National Institute Of Applied Science
And Technology
CARTHAGE UNIVERSITY

End of Year Project

Branch
Networks and telecommunications

Personality detection of Instagram users

Made by :
Ben Omrane Mohamed Salim
Turki Mohamed Seddik
Hammami Omar
Abdelkader Iheb

Supervisor : Mrs. Sana Hamdi

Reviewer : Mrs. Wided MILED SOUID

Study year: 2023/2024

Contents

List of Figures

List of Tables

Introduction

This project harnesses the power of artificial intelligence, specifically through the application of deep learning models, to analyze and predict personality traits based on Instagram profiles. By examining the visual and textual content shared by users, our AI-driven approach aims to decode the underlying personality dimensions that influence these posts. This innovative method not only enhances our understanding of digital personas but also bridges the gap between psychological theories and their practical implications in digital interactions.

The potential applications of this technology extend far beyond mere academic interest. By integrating personality detection into systems, we can revolutionize recommendation algorithms, enhance targeted marketing strategies, and improve social media interfaces to align more closely with individual user preferences. In recruitment and professional settings, this technology can aid in understanding the cultural fit of potential candidates through their digital footprints. Furthermore, it opens up avenues in mental health services by providing insights into personal well-being indirectly expressed through social media behavior.

Through this project, we aim to set a benchmark in personality detection using social media data, pushing the boundaries of AI in psychology and offering tools that are not only innovative but also immensely practical in various domains

1

Project Settings

1 Project Context and objectives

In this section, we will provide an overview of our project context and our primary focus, as well as the goals we are currently trying to achieve.

2 Project Context

Social media platforms like Instagram have become a mirror reflecting the personality of its users by allowing individuals to create and share content that highlights their interests, values, and lifestyle.
Through the images, captions, and stories they post, users can express their unique identities and connect with others who share similar passions, effectively creating a digital persona that mirrors their real-world selves.

2.1 Problematic

Our challenge centers on accurately predicting user personalities based on their Instagram posts. While textual data offers valuable insights, integrating visual content, such as images, adds another layer of complexity. This project aims to tackle this challenge by investigating the impact of combining image analysis using Vision Transformer (ViT) with text analysis using XLM-RoBERTa. We aim to explore whether this fusion of text and image analysis enhances predictive performance and provides added value in accurately predicting user personalities.

2.2 Objectives

- **- Data Collection** The initial step is to look and search for adequate public instagram profiles with enough content and posts that are reflective of their personalities. After that scraping and storing the data will be the priority.
- **- Data Preparation** The next step involves a comprehensive exploration and understanding of the collected dataset, which includes user profiles and posts on Instagram. This process will entail data cleaning, preprocessing, and analyzing user behavior, demographics, and posting patterns as well as studying their followees and followers.

3 Project Methodology

3.1 CRISP-DM

To ensure the smooth progression of our work, it is essential to establish a clear and well-structured methodology that aligns with the data science nature of our project. We have chosen to adopt the CRISP-DM methodology, which is the most common methodology for data mining, analytics, and data science projects.. The CRISP-DM methodology consists of six primary stages (1.1) that guide us throughout the project, from understanding the business problem to deployment.

- **- Business Understanding** The Business Understanding phase focuses on understanding the objectives and requirements of the project.

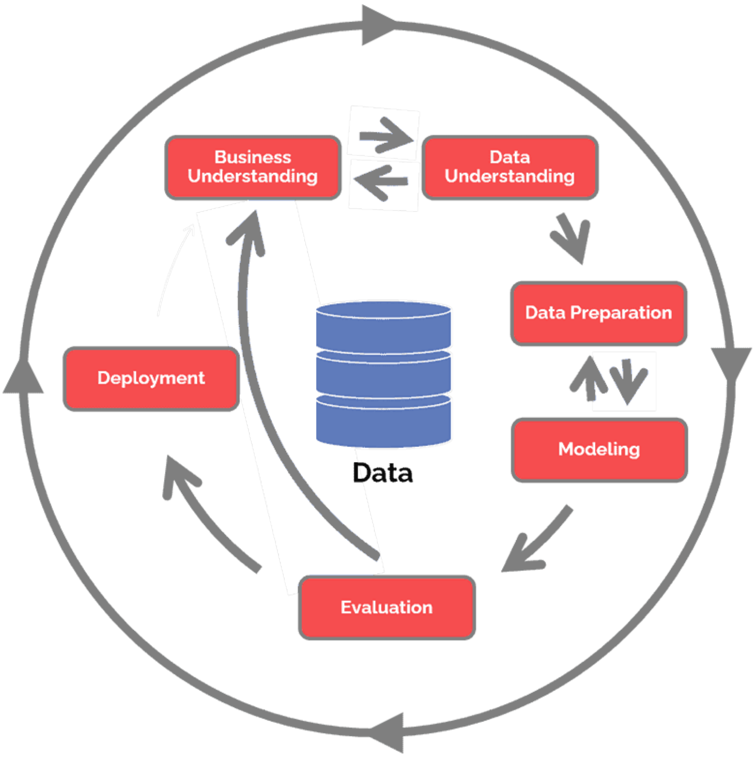


Figure 1.1: CRIPS-DM PROCESS [3]

- **- Data Understanding** Adding to the foundation of Business Understanding, this step drives the focus to identify, collect, and analyze the data sets that can help us accomplish the project goals.
- **- Data Preparation** Also known as data construction or the creation of the DataHub, this phase involves activities aimed at transforming raw data sources into a well-defined dataset ready for analysis. This phase takes the majority of the project.
- **- Modeling** The modeling phase is the core of our project. It involves selecting, configuring, and creating different algorithms that best suit our project's objectives.
- **- Evaluation** Whereas the Modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next.
- **- Deployment** The final stage of the process involves putting the acquired knowledge, obtained through the created models, into practical use for end users.

2

State of the Art

Introduction

This section provides an overview of the foundational concepts and prevalent models used in personality psychology, which underpin our analysis of Instagram user data. Understanding these models is crucial for developing effective tools for personality prediction and understanding user behavior on social media platforms.

1 Basic Concepts

Two of the most influential models in the study of personality are the Big Five Personality Traits and the Myers-Briggs Type Indicator (MBTI). These models offer frameworks for analyzing and categorizing human behavior and traits, which are integral to our predictive modeling approach.

1.1 Big Five Personality Traits:

- **Openness:** Appreciation for art, emotion, adventure, unusual ideas, imaginative and curious.
- **Conscientiousness:** A tendency to show self-discipline, act dutifully, and aim for achievement, planned rather than spontaneous behavior.
- **Extraversion:** Energy, positive emotions, surgency, and the tendency to seek stimulation in the company of others.
- **Agreeableness:** A tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others.
- **Neuroticism:** The tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, or vulnerability.

1.2 16 Personality Types (MBTI):

The MBTI is a personality framework that classifies individuals into one of 16 distinct types based on four dichotomies: Introversion/Extraversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving. Each type is typically represented by a four-letter code:

- **ISTJ (Inspector):** Practical and fact-minded individuals, whose reliability cannot be doubted.
- **ISFJ (Defender):** Very dedicated and warm protectors, always ready to defend their loved ones.
- **INFJ (Advocate):** Quiet and mystical, yet very inspiring and tireless idealists.
- **INTJ (Architect):** Imaginative and strategic thinkers, with a plan for everything.
- **ISTP (Virtuoso):** Bold and practical experimenters, masters of all kinds of tools.
- **ISFP (Adventurer):** Flexible and charming artists, always ready to explore and experience something new.
- **INFP (Mediator):** Poetic, kind, and altruistic people, always eager to help a good cause.
- **INTP (Logician):** Innovative inventors with an unquenchable thirst for knowledge.

- **ESTP (Entrepreneur):** Smart, energetic, and very perceptive people, who truly enjoy living on the edge.
- **ESFP (Entertainer):** Spontaneous, energetic, and enthusiastic entertainers – life is never boring around them.
- **ENFP (Campaigner):** Enthusiastic, creative, and sociable free spirits, who can always find a reason to smile.
- **ENTP (Debater):** Smart and curious thinkers who cannot resist an intellectual challenge.
- **ESTJ (Executive):** Excellent administrators, unsurpassed at managing things – or people.
- **ESFJ (Consul):** Extraordinarily caring, social, and popular people, always eager to help.
- **ENFJ (Protagonist):** Charismatic and inspiring leaders, able to mesmerize their listeners.
- **ENTJ (Commander):** Bold, imaginative, and strong-willed leaders, always finding a way or making one.

By understanding these personality frameworks, we can better analyze Instagram user data, enabling a more nuanced approach to predicting user behavior based on their interactions with content.

1.3 Model Selection: Why MBTI?

We have chosen the Myers-Briggs Type Indicator (MBTI) over the Big Five Personality Traits for our study due to several advantages. MBTI's detailed classification into 16 personality types provides a nuanced framework that is particularly effective for analyzing user behavior and communication styles on Instagram. This specificity aids in understanding and predicting content preferences, making it valuable for targeted marketing and personalization strategies. Additionally, MBTI's widespread recognition and ease of interpretation make it accessible to a broad audience, aligning well with Instagram's diverse user base. By leveraging MBTI, we aim to enhance user engagement through personalized content recommendations, utilizing its clear insights into how individuals perceive the world and make decisions.

2 Machine learning concepts

2.1 Supervised Learning

Supervised learning is a type of machine learning where the model is trained on a labeled dataset. This means the data used for training includes both the input data and the correct output. The goal of supervised learning is to learn a function that, given a sample of data and desired outputs, approximates the relationship between input and output variables. This approach is commonly used for classification and regression tasks.

2.2 Unsupervised Learning

Unsupervised learning involves training a model on data that does not have labeled responses. The model attempts to identify the underlying patterns or distributions in the data. Common applications include clustering and dimensionality reduction, where the goal is to group similar data points together or reduce the number of random variables under consideration, respectively.

2.3 Machine Learning (ML)

Machine learning is a field of artificial intelligence that enables systems to learn from data, identify patterns, and make decisions with minimal human intervention. ML algorithms are designed to improve their performance at a task progressively with experience.

2.4 Deep Learning (DL)

Deep learning is a subset of machine learning that uses algorithms inspired by the structure and function of the brain called artificial neural networks. Deep learning models are capable of learning from data in an unstructured form such as images, text, and sound. These models have proven to be very effective at tasks that involve a large amount of data and complex patterns that are difficult for a human to extract.

3 Deep Learning Models

3.1 Introduction to Transformers

Our choice to employ transformer models over other machine learning models is driven by several key factors that align with the unique demands of our project. Transformers are renowned for their superior ability to handle sequential data, making them particularly suited for processing and generating natural language and analyzing images in sequences. This capability stems from their core mechanism, the attention mechanism, which allows the model to focus on different parts of the input data, providing a dynamic understanding of the context.

Additionally, transformers are highly parallelizable, which significantly reduces training times and allows them to manage large datasets efficiently. This is crucial for handling the vast amount of data typically involved in analyzing Instagram posts, which includes both textual and visual content. Their ability to integrate and learn from multiple data modalities simultaneously makes transformers especially powerful for multimodal tasks that are central to our project, such as sentiment analysis and personalized content recommendation.

3.2 Transformers

Transformers are a type of deep learning model that relies on mechanisms called attention mechanisms, which weigh the influence of different parts of the input data differently. Transformers are particularly effective for processing sequences of data, such as natural language for tasks like translation and text summarization. They are known for their parallelizability and efficiency in handling long inputs.

3.3 XLM-Roberta

XLM-Roberta (Cross-lingual Language Model - Roberta) is a deep learning model based on the RoBERTa architecture which itself is an optimization of BERT (Bidirectional Encoder Representations from Transformers). XLM-Roberta is pre-trained on a large corpus of text in multiple languages, making it effective for tasks that involve understanding and generating text across different languages. It is often used in tasks like sentiment analysis, language translation, and content classification.

3.4 ViT (Vision Transformer)

The Vision Transformer (ViT) model is an adaptation of the transformer architecture for image recognition tasks. Instead of processing the image as a whole, ViT divides the image into patches and processes these patches sequentially as if they were elements in a sequence. This method allows ViT to apply the powerful self-attention mechanisms of transformers to image analysis, leading to state-of-the-art performance on many benchmarks.

3.5 Siamese XLM-RoBERTa

Siamese XLM-RoBERTa is a model architecture that leverages the XLM-RoBERTa transformer model in a Siamese network configuration. This setup involves using two parallel instances of XLM-RoBERTa to process pairs of text inputs, typically for tasks like sentence similarity, text matching, or paraphrase identification. The embeddings produced by each instance are compared or combined to produce a final similarity score or classification. This architecture benefits from XLM-RoBERTa's multilingual capabilities and deep contextual understanding while being optimized for tasks involving text pairs.

3

Data collection and Data Understanding

1 Data Gathering

1.1 Selenium : [1]

1.1.1 Definition :

Selenium is an open-source framework used primarily for automating web browsers. It's particularly effective for repetitive web tasks like web scraping, allowing for automated interactions with web pages.

1.1.2 Scraping Process Using Selenium :

Our project utilizes Selenium to automate data collection from Instagram profiles to analyze user behaviors and personality traits. The process includes:

1. **Logging In:** Selenium navigates to Instagram and logs in using predefined credentials.
2. **Profile Access:** It searches for and navigates to specific user profiles based on usernames.
3. **Data Extraction:** Selenium scrapes essential data such as:
    - **Profile User Name**
    - **Number of Followers**
    - **Number of Follows**
    - **Biography**
    - **Captions and Images from Posts**
4. **Dynamic Data Handling:** The script simulates scrolling to load and capture data from dynamic content like post feeds.
5. **Data Storage:** Extracted data is organized and saved for further analysis.

1.2 Apify: [2]

1.2.1 Definition :

Apify is a platform where developers can build, deploy, and publish web scraping, data extraction, and web automation tools. It works seamlessly with both Python and JavaScript and supports libraries like Playwright, Puppeteer, Selenium, and Scrapy. Apify allows developers to turn their code into Apify Actors, which are serverless microapps that are easy to develop, run, share, and integrate. The platform provides infrastructure, proxies, and storage solutions ready for immediate use.

1.2.2 Scraping Process Using Apify :

Our project utilizes Apify to automate data collection from Instagram profiles to analyze user behaviors and personality traits. The process includes:

1. **Accessing Apify Actors:**We utilize two existing Apify actors (Instagram Profile Scraper and Instagram Posts Scraper). These actors are capable of extracting detailed profile information and post data, respectively.
2. **Parallel Data Collection** Apify allows us to scrape multiple Instagram profiles simultaneously by providing the API with a list of usernames, significantly speeding up and streamlining the data collection process.
3. **Configuration and Execution:**
   - **Using API:** We access Apify's API to execute these actors. The API is configured with our account token to ensure secure access.
   - **Web Interface:**Alternatively, Apify provides a graphical interface on their website to manage and run the scraping tasks.
4. **Dynamic Data Handling:** Apify's infrastructure supports dynamic data handling, ensuring that all relevant content, including dynamically loaded posts, is captured effectively.
5. **Data Storage:** The scraped data is organized and saved using Apify's storage solutions for further analysis.

1.3 Comparison between Selenium and Apify :

Table 3.1: Comparison between Selenium and Apify

Comparison	Pros	Cons
Selenium	• **Flexible:**Highly adaptable and can handle a wide range of web automation tasks. • **Specific Task Handling:** Capable of performing very specific and complex tasks that other tools might not be able to manage.	• **Verbose Code:**Requires writing a significant amount of code, which can be time-consuming and complex. • **Proxy Issues:** Can encounter problems with proxies, often leading to being blocked on platforms like Instagram.
Apify	• **Faster:**Optimized for quick and efficient data collection. • **Parallel Data Collection:** Supports scraping multiple profiles simultaneously, enhancing speed and efficiency. • **User-Friendly:**Easy to use with a straightforward setup process. • **Proxy Management:**Includes robust proxy management to avoid blocks, ensuring smooth operation.	• **Limited by Pre-existing Actors:**May be limited for some tasks if the necessary actors are not available in the Apify library. • **Less Customizable:** While easy to use, it may not offer the same level of detailed customization for specific tasks as Selenium. • **Not Free:** Apify is not free and provides only a 5$ credit for initial use, which may limit extensive scraping activities.

⇒ We chose Apify because it is faster and supports parallel data collection. Additionally, the existing actors on Apify can efficiently handle our specific scraping tasks, making it the ideal tool for our project.

2 Data cleaning :

Cleaning data is a critical step in preparing your dataset for analysis or machine learning. Different types of data (numerical, images, and text) require different cleaning techniques.

Here are some techniques for each type:

1. **Numerical Data:**
   - **Remove Missing Values:** Remove rows or columns with missing values if they are not essential.
   - **Scaling and Normalization:** Normalize or standardize numerical features to bring them to a similar scale.
2. **Image Data:**
   - **Resizing Images:** Resize images to a uniform size.
3. **Text Data:**
   - **Lowercasing:** Convert all characters to lowercase to maintain consistency.
4. **General Data Cleaning Steps:**
   - **Duplicated Data:** Remove duplicate rows.
   - **Correct Data Types:** Convert columns to appropriate data types.
   - **Remove Rows with Null Values:** Remove rows with null values in specific columns, such as caption and images.

3 Data Labeling :

We utilize two methods for data labeling:

3.1 Websites :

These platforms offer insights into user personalities through user and expert voting systems, allowing us to determine the personality traits of the profiles we search for.

- **personality-database [4]**
- **crystalknows [5]**
- **boo.world [6]**

3.2 Manual :

In cases where profiles aren't available on websites, we employ manual identification, relying on the following method:

- **Extraversion (E) vs. Introversion (I):** Describes whether you focus on the outer world of people and things (extraversion) or your inner world of ideas and impressions (introversion).
- **Sensing (S) vs. Intuition (N):** Reflects how you prefer to gather information. Sensing types focus on concrete, tangible details and experiences, while intuitive types focus on patterns, possibilities, and meanings.
- **Thinking (T) vs. Feeling (F):** Indicates how you make decisions. Thinking types prioritize logic, consistency, and objective criteria, while feeling types prioritize empathy, harmony, and the impact on others.
- **Judging (J) vs. Perceiving (P):** Describes how you approach the outside world. Judging types prefer structure, organization, and closure, while perceiving types prefer flexibility, spontaneity, and adaptability.

4 Data Understanding (EDA):

To gain insights from our data, we performed various visualizations:

- **Data Samples:** We plotted samples of user names along with one of their post images and captions to get a better sense of the content.

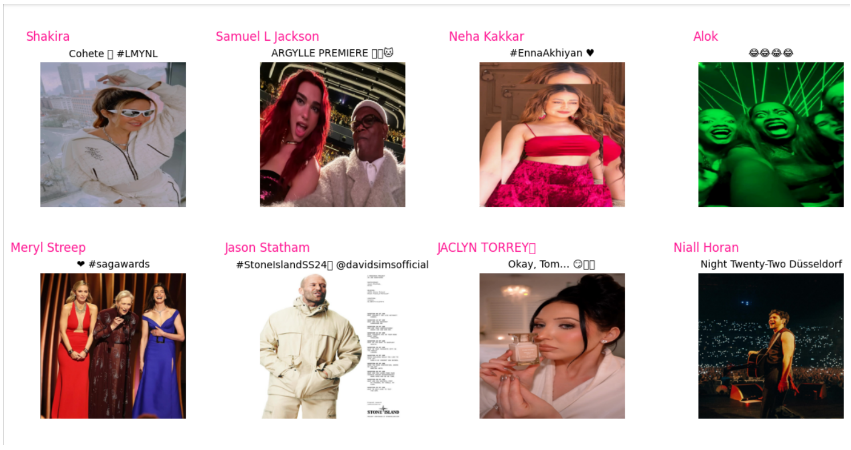
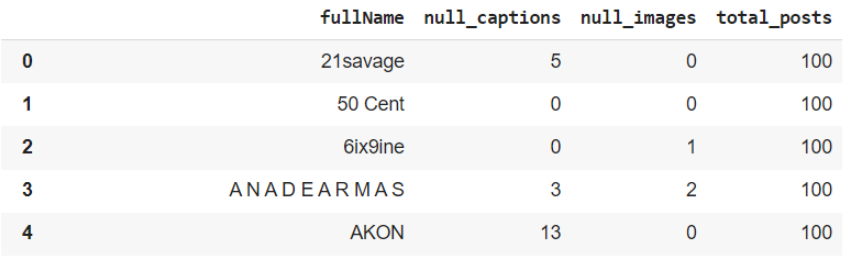


Figure 3.1: Sample dataset

- **Head of Dataset:** We examined the first five rows of our dataset, focusing on instances with null captions and null images. Additionally, we plotted the count of posts, null captions, and null images per user.



	fullName	null_captions	null_images	total_posts
0	21savage	5	0	100
1	50 Cent	0	0	100
2	6ix9ine	0	1	100
3	ANADEARMAS	3	2	100
4	AKON	13	0	100

Figure 3.2: instances with null captions and null images

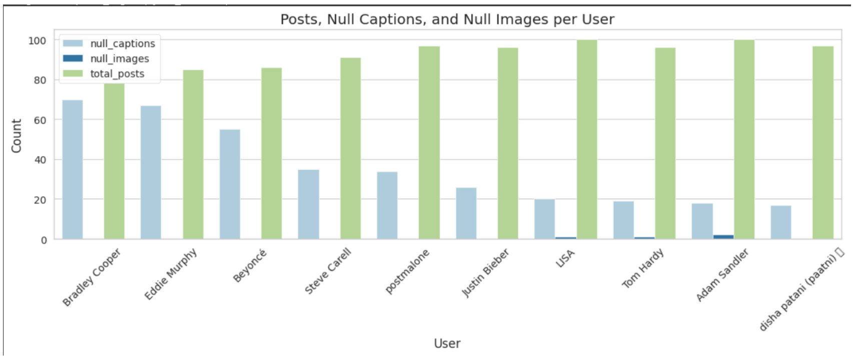


Figure 3.3: The count of posts, null captions, and null images per user

We created several visualizations related to captions:

- **Distribution of Caption Lengths:** We plotted the distribution of caption lengths across the dataset to understand the variety in content length.

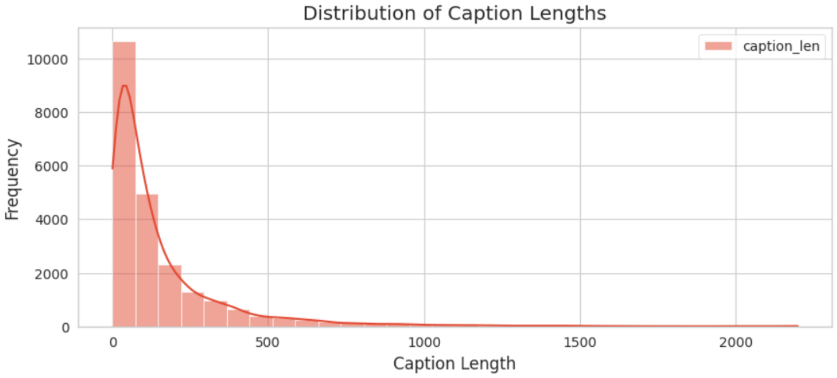
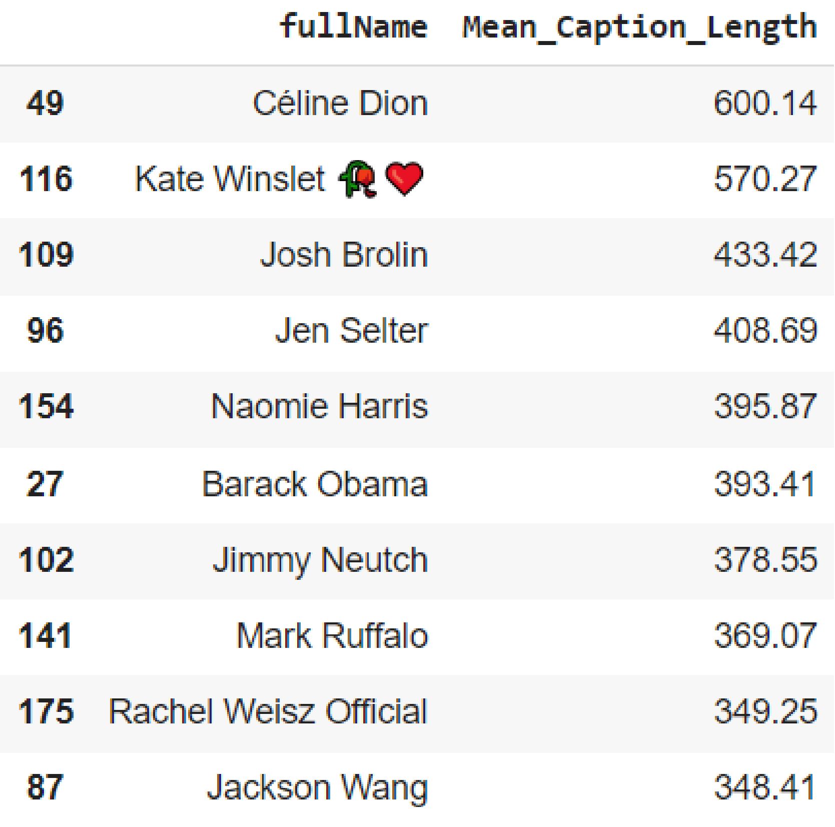


Figure 3.4: Distribution of Caption Lengths

- **Mean of Caption Lengths per User:** We calculated and plotted the mean caption length for each user to see how verbose different users are in their posts.

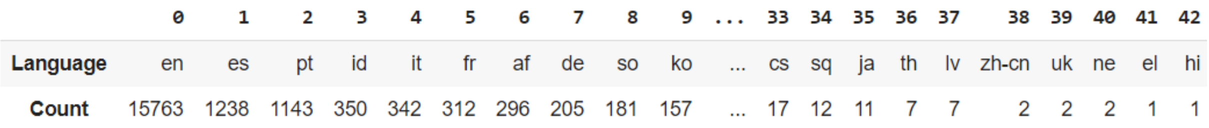


	fullName	Mean_Caption_Length
49	Céline Dion	600.14
116	Kate Winslet 🌹❤️	570.27
109	Josh Brolin	433.42
96	Jen Selter	408.69
154	Naomie Harris	395.87
27	Barack Obama	393.41
102	Jimmy Neutch	378.55
141	Mark Ruffalo	369.07
175	Rachel Weisz Official	349.25
87	Jackson Wang	348.41

Figure 3.5: Mean of Caption Lengths per User

Languages Used in Captions: We analyzed the languages used in captions:

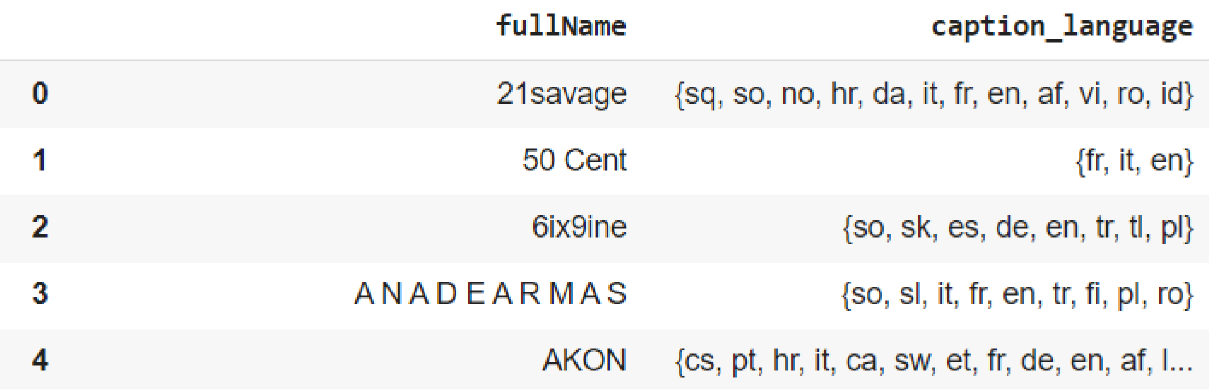
- **Table of Language Counts:** We compiled a table showing the count of each language used in the entire dataset.



	0	1	2	3	4	5	6	7	8	9	...	33	34	35	36	37	38	39	40	41	42
Language	en	es	pt	id	it	fr	af	de	so	ko	...	cs	sq	ja	th	lv	zh-cn	uk	ne	el	hi
Count	15763	1238	1143	350	342	312	296	205	181	157	...	17	12	11	7	7	2	2	2	1	1

Figure 3.6: Language Counts

- **Languages per User:** We examined how many different languages each user used in their captions and visualized this data.



	fullName	caption_language
0	21savage	{sq, so, no, hr, da, it, fr, en, af, vi, ro, id}
1	50 Cent	{fr, it, en}
2	6ix9ine	{so, sk, es, de, en, tr, tl, pl}
3	ANADEARMAS	{so, sl, it, fr, en, tr, fi, pl, ro}
4	AKON	{cs, pt, hr, it, ca, sw, et, fr, de, en, af, l...

Figure 3.7: Languages per User

- **Personality Distribution:** we plotted the distribution of personality traits across the dataset to understand the variety and prevalence of different personality types among the users.

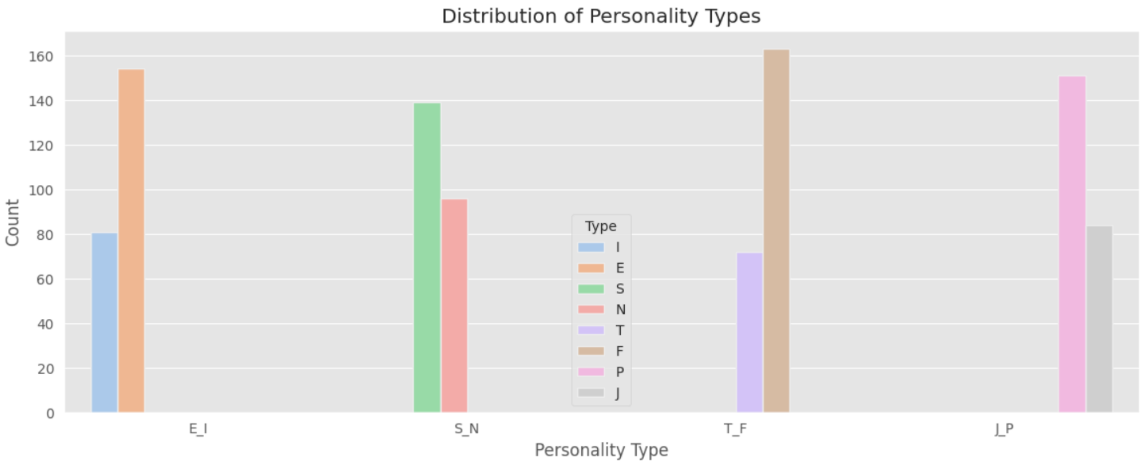


Figure 3.8: Personality Distribution:

4

Modeling and Evaluation

1 Introduction

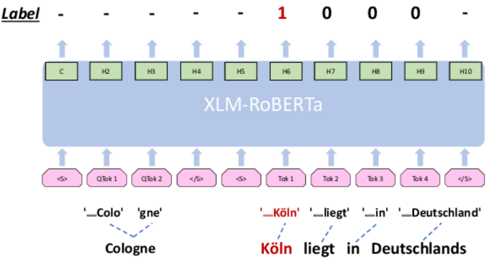
In this section we will present the model building process, in which we have trained different models, each model trained on a subset in order to select the most suitable model.

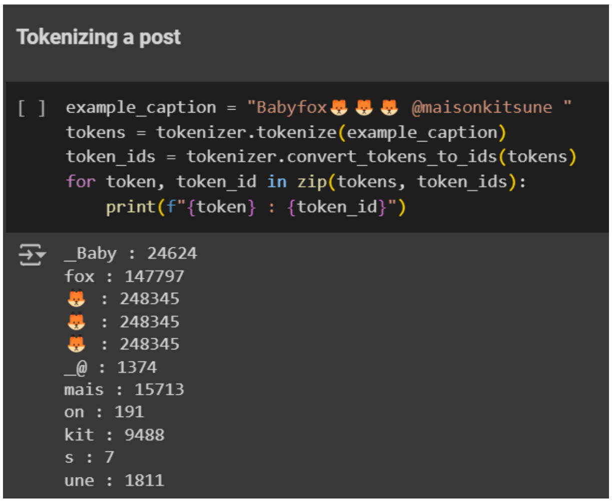
2 Model Selection Intuition

In this Part, we aim to select the models that best serves our interests. Since our data is comprised of texts, images and numerical data, we chose to select the best models commonly used for each modality. In the next sections we will dive deeper into each modality and each model.

2.1 XLM-RoBERTA for Text

Since our data is scraped from a social media website, we tend to find a lot of users posting in different languages, some posts contain multiple languages at the same time. Also, text from Instagram poasts with a lot of emojis and hashtags which contains critical features for our models decisions. Therefore we opted to choose XLM-RoBERTa model as our text feature extraction due to its capacity to understand both multilingual text, emojis and hashtags.


Label
- - - - - 1 0 0 0 -
C H2 H3 H4 H5 H6 H7 H8 H9 H10
XLM-RoBERTa
<S> QTok 1 QTok 2 </S> <S> Tok 1 Tok 2 Tok 3 Tok 4 </S>
'_Colo' 'gne'
'_Köln' '_liegt' '_in' '_Deutschland'
Cologne
Köln liegt in Deutschlands



Tokenizing a post

```
[ ] example_caption = "Babyfox🦊🦊🦊 @maisonkitsune "
    tokens = tokenizer.tokenize(example_caption)
    token_ids = tokenizer.convert_tokens_to_ids(tokens)
    for token, token_id in zip(tokens, token_ids):
        print(f"{token} : {token_id}")
```

```
_Baby : 24624
fox : 147797
🦊 : 248345
🦊 : 248345
🦊 : 248345
_@ : 1374
mais : 15713
on : 191
kit : 9488
s : 7
une : 1811
```

Figure 4.1: XLM-RoBERTa Tokenizing emojis

2.2 Visual Transformers (ViT) for Image

To extract meaningful features from the images, we opted for the Vision Transformer (ViT) model. While CNNs have been popular for image recognition tasks, they can be limited in their ability to learn the relation between different parts of an image. ViT, on the other hand, excels in this area, therfore has the capacity to capture information of visual features from Instagram images.

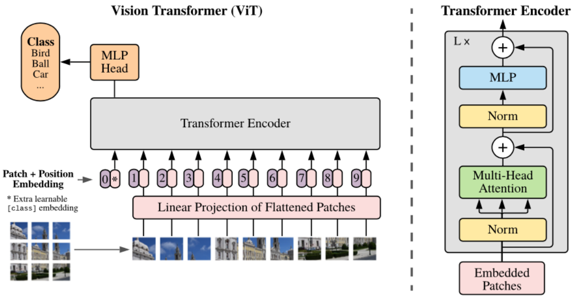


Figure 4.2: ViT Architecture

2.3 Dense Neural Network for Numerical data

In addition to text and image analysis, we also extracted numerical features from Instagram posts. Due to the relatively low dimensionality of the numerical data, we opted for a simple Deep Neural Network (DNN) for feature extraction.

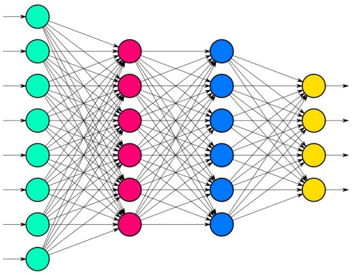
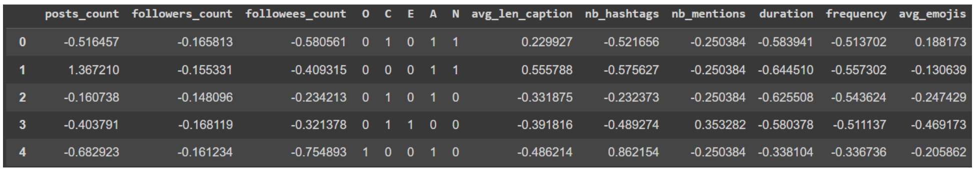


Figure 4.3: DNN Architecture

Some Numerical data are extracted from users' profiles such as N° posts, N° followers, and some other features we hand crafted using data engineering like frequency of posting, N° of emojis in the captions ... Here's the complete list of the numercial features:



	posts_count	followers_count	followees_count	O	C	E	A	N	avg_len_caption	nb_hashtags	nb_mentions	duration	frequency	avg_emojis
0	-0.516457	-0.165813	-0.580561	0	1	0	1	1	0.229927	-0.521656	-0.250384	-0.583941	-0.513702	0.188173
1	1.367210	-0.155331	-0.409315	0	0	0	1	1	0.555788	-0.575627	-0.250384	-0.644510	-0.557302	-0.130639
2	-0.160738	-0.148096	-0.234213	0	1	0	1	0	-0.331875	-0.232373	-0.250384	-0.625508	-0.543624	-0.247429
3	-0.403791	-0.168119	-0.321378	0	1	1	0	0	-0.391816	-0.489274	0.353282	-0.580378	-0.511137	-0.469173
4	-0.682923	-0.161234	-0.754893	1	0	0	1	0	-0.486214	0.862154	-0.250384	-0.338104	-0.336736	-0.205862

Figure 4.4: Num Features

3 Evaluation Setup

To ensure the generalizability and robustness of our AI system, we employed a common practice in machine learning: splitting the data into separate training, validation, and testing sets. So for this purpose we splitted the users in our data, we opted for an 80/10/10 split, 80 percent of users in the training set, 10 percent in the validation set for hyperparameter tuning, and 10 percent in the test set for final model evaluation.

4 Model Architectures

To comprehensively understand the role of different data modalities in personality detection, we trained and evaluated five unique models. The first model utilized XLM-RoBERTa to extract features solely from the text. The second model employed ViT to analyze the visual features within images. To explore the potential of combining these modalities, a third model was created by combining text features from XLM-RoBERTa and image features from ViT. Additionally, a fourth model was trained on a combination of biography and numerical data. Finally, a fifth model was trained exclusively on the numerical data to assess its conntribution to personality prediction. In the next sections we will dive deeper into each trained model.

4.1 XLM-RoBERTa Model

This model is solely trained on the text only where XLM-RoBERTa model, serves as the backbone for our text-based feature extraction, the text-pooled outputs for each user are passed throught a MultiHead Attention layer to select the best and most coherent captions that best describes the user's personality. The output of the MHA layer is then fed into the classification heads, each head is specific for a personality trait from the OCEAN personalities.

During model evaluation, we assessed the loss of the model. We noticed that the loss function did not exhibit a decrease across training epochs. This suggests potential difficulties in model learning and its ability to capture the intricate relationships between text data and the target personality traits.

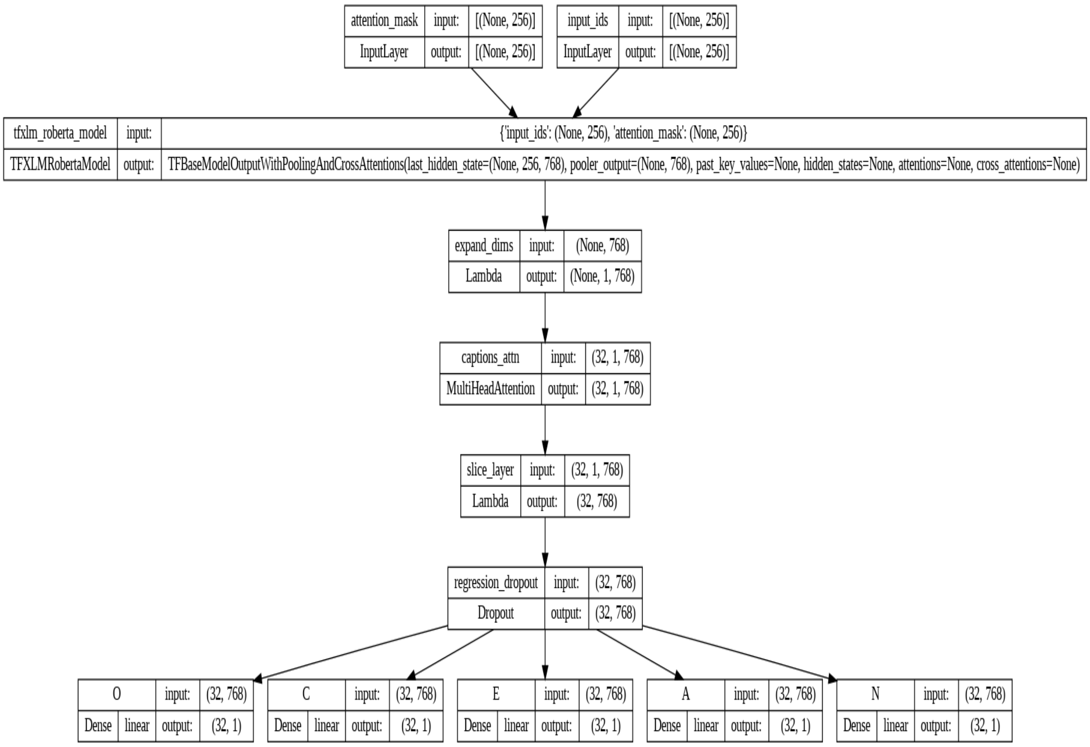

attention_mask | input: | [(None, 256)]
InputLayer | output: | [(None, 256)]
input_ids | input: | [(None, 256)]
InputLayer | output: | [(None, 256)]
tfxlm_roberta_model | input: | {'input_ids': (None, 256), 'attention_mask': (None, 256)}
TFXLMRobertaModel | output: | TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 256, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)
expand_dims | input: | (None, 768)
Lambda | output: | (None, 1, 768)
captions_attn | input: | (32, 1, 768)
MultiHeadAttention | output: | (32, 1, 768)
slice_layer | input: | (32, 1, 768)
Lambda | output: | (32, 768)
regression_dropout | input: | (32, 768)
Dropout | output: | (32, 768)
O | input: | (32, 768)
Dense | linear | output: | (32, 1)
C | input: | (32, 768)
Dense | linear | output: | (32, 1)
E | input: | (32, 768)
Dense | linear | output: | (32, 1)
A | input: | (32, 768)
Dense | linear | output: | (32, 1)
N | input: | (32, 768)
Dense | linear | output: | (32, 1)

Figure 4.5: Text Model Architecture

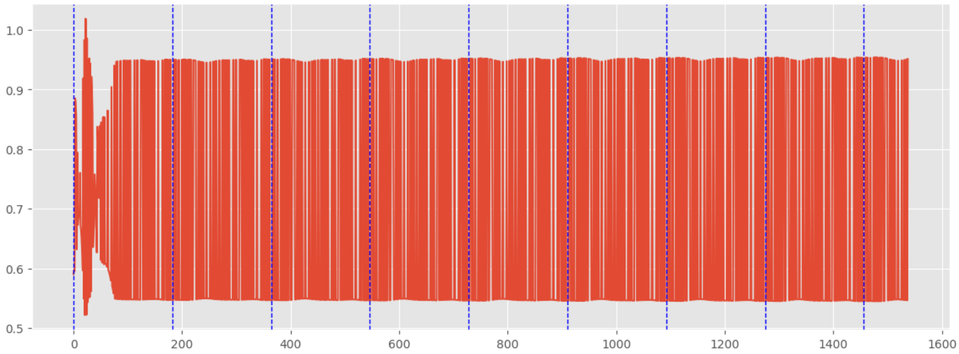


Figure 4.6: Text Model Loss over the epochs

4.2 ViT Model

Due to the limitations of the text model and its poor performance, we decided to look into the image features and train a model using images only. The same strategy and model architecture from the text model has been adopted, the only change is that ViT model serves now as the backbone for our image-based feature extraction.

During model evaluation, we observed the loss, accuracy and logits of the model. We noticed that the loss function did not decrease across training epochs. The accuracy did not get any higher and the logits were the same for different classes (the model predicts the same class regardless of the input), so the model still struggles to indentify

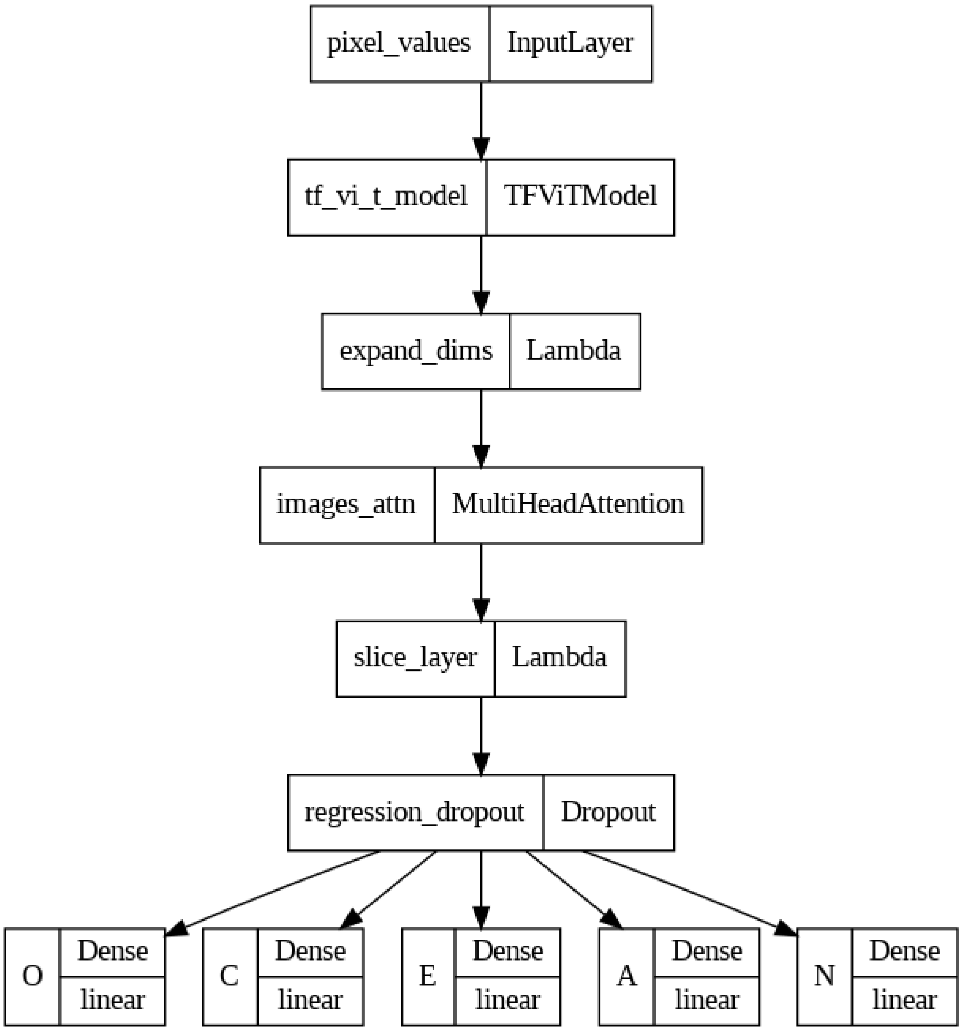
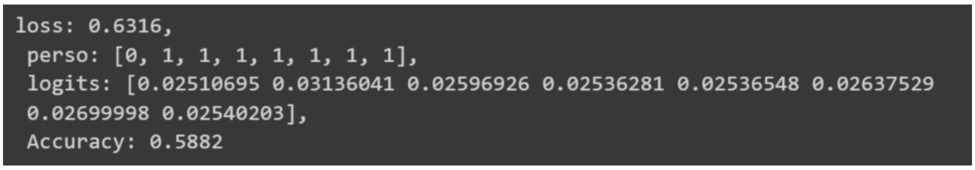


Figure 4.7: Image Model Architecture

the relationships between the data and the target personality traits.



```
loss: 0.6316,
 perso: [0, 1, 1, 1, 1, 1, 1, 1],
 logits: [0.02510695 0.03136041 0.02596926 0.02536281 0.02536548 0.02637529
 0.02699998 0.02540203],
 Accuracy: 0.5882
```

Figure 4.8: Image Model Evaluation for C Trait

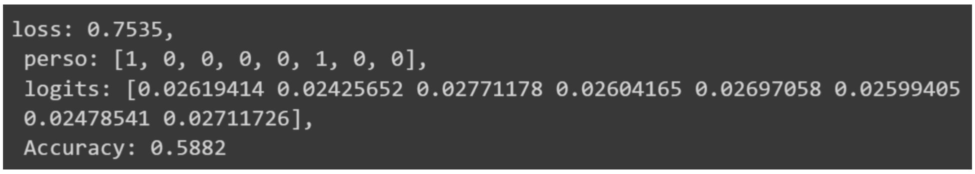
4.3 XLM-RoBERTa + ViT Model

We observed poor performances from the text model and the image model so we decided two combine the modalties.

The ViT and xlm roberta models serves as the backbone for this model here's an overview: the text is passed throught roberta for feature extraction as well as the image throught

ViT.
these two features from text and image are passed throught a MultiHead Attention layer, so that the text and image informations from the same post are combined to form a single post feature. The new posts features are then passed throught another multihead attention layer to select the best posts that best describes the user personality and then feed them into the classification heads as mentionned before in previous sections.

For evaluation, We observed the model's loss, accuracy and logits. The model still struggles to identify the users' personalities and predicts a single class for each input, so we concluded that the model could not capture the dataset from images and text.



```
loss: 0.7535,
 perso: [1, 0, 0, 0, 0, 1, 0, 0],
 logits: [0.02619414 0.02425652 0.02771178 0.02604165 0.02697058 0.02599405
 0.02478541 0.02711726],
 Accuracy: 0.5882
```

Figure 4.9: Text Image Model Evaluation for C trait

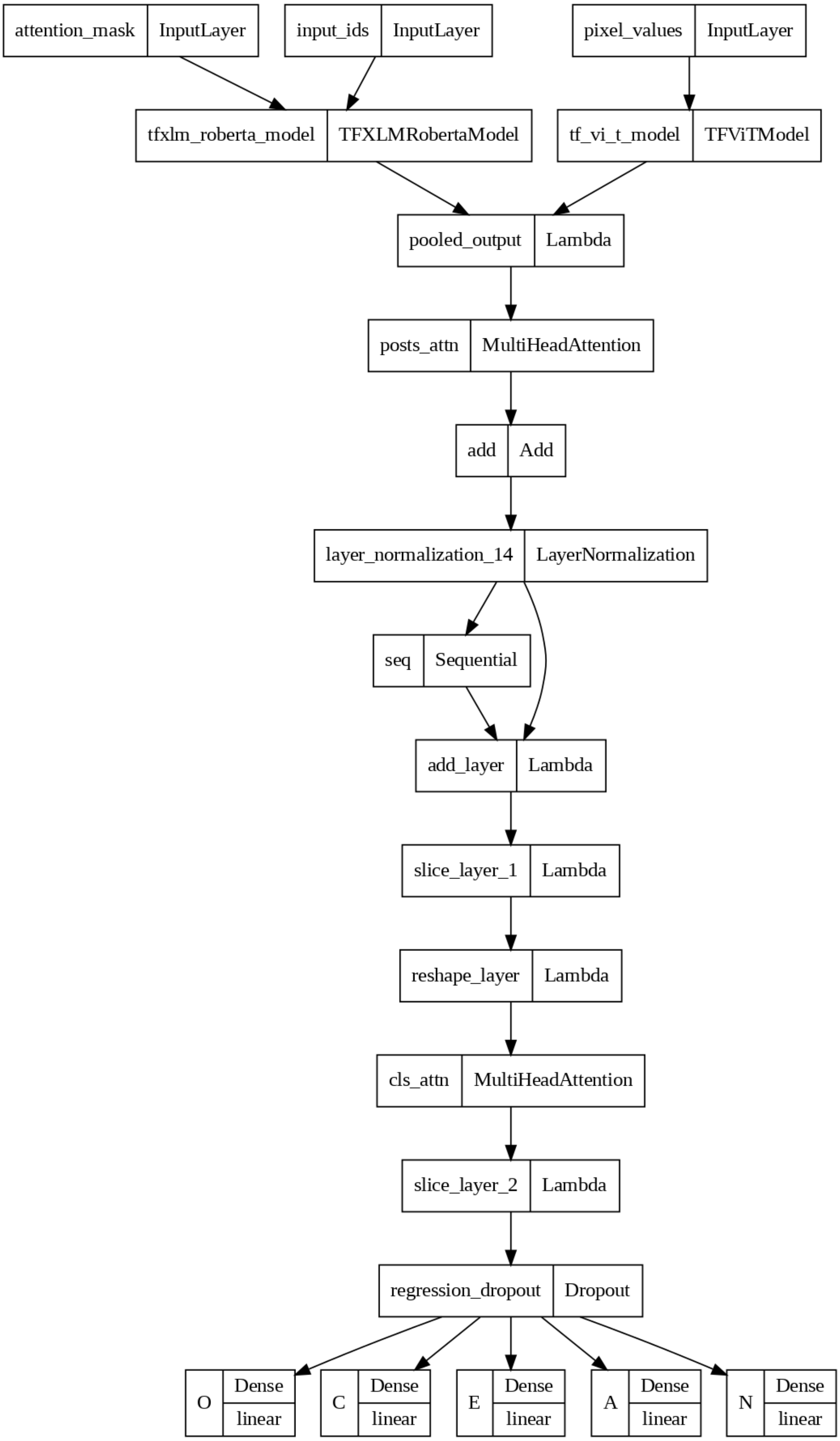

attention_mask | InputLayer
input_ids | InputLayer
pixel_values | InputLayer
tfxlm_roberta_model | TFXLMRobertaModel
tf_vi_t_model | TFViTModel
pooled_output | Lambda
posts_attn | MultiHeadAttention
add | Add
layer_normalization_14 | LayerNormalization
seq | Sequential
add_layer | Lambda
slice_layer_1 | Lambda
reshape_layer | Lambda
cls_attn | MultiHeadAttention
slice_layer_2 | Lambda
regression_dropout | Dropout
O | Dense | linear
C | Dense | linear
E | Dense | linear
A | Dense | linear
N | Dense | linear

Figure 4.10: Text Image Model Architecture

We will try with other modalities like bio and numerical features.

4.4 Biography and Numerical data Model

Considering the poor performances of the captions and image models, we decided to work with other modalities like Bio and numerical. This model extract the features from the bio using xlm roberta and DNN to extract numerical features, then these two features are combined with attention layer and fed into the classification heads.

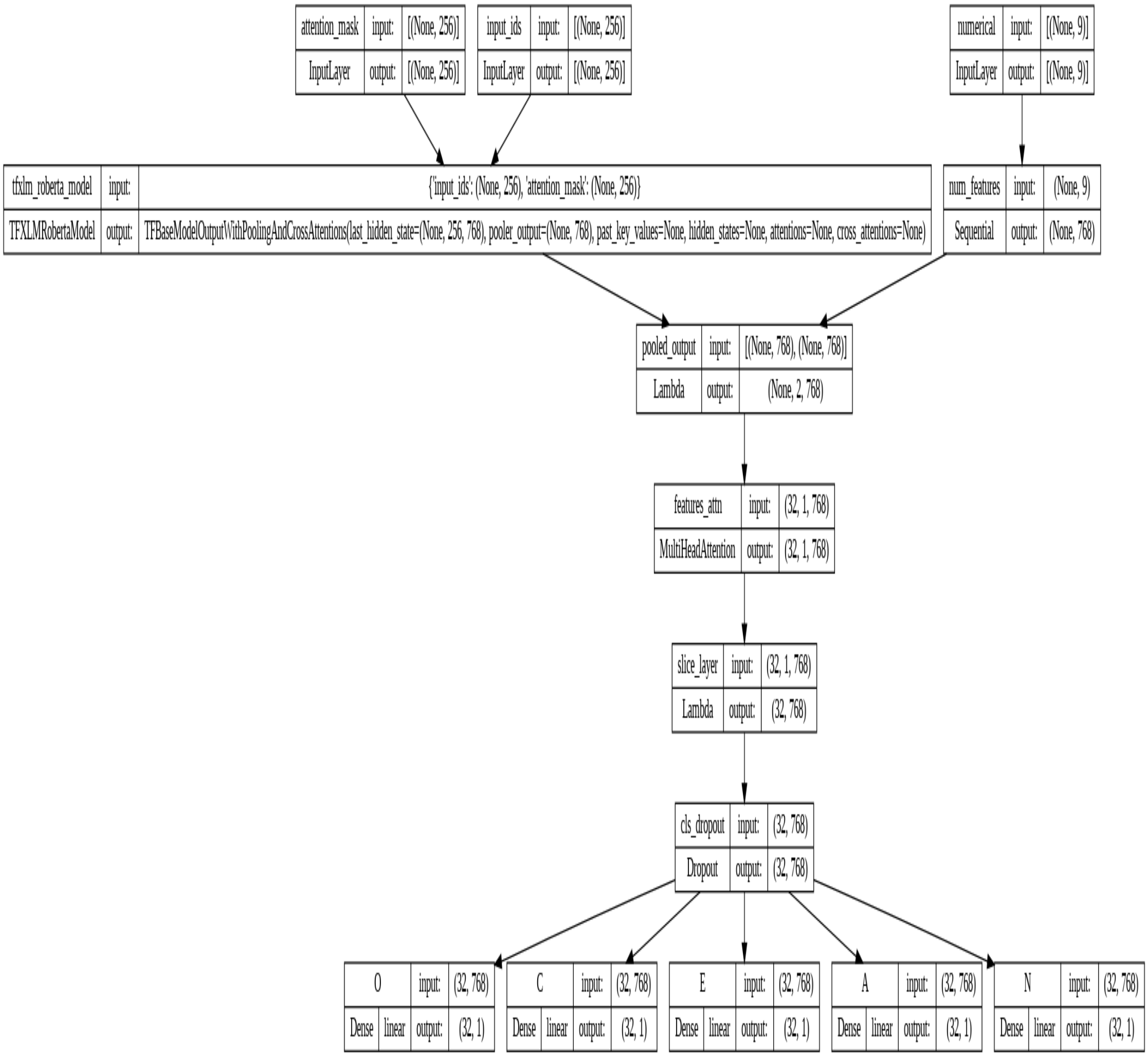
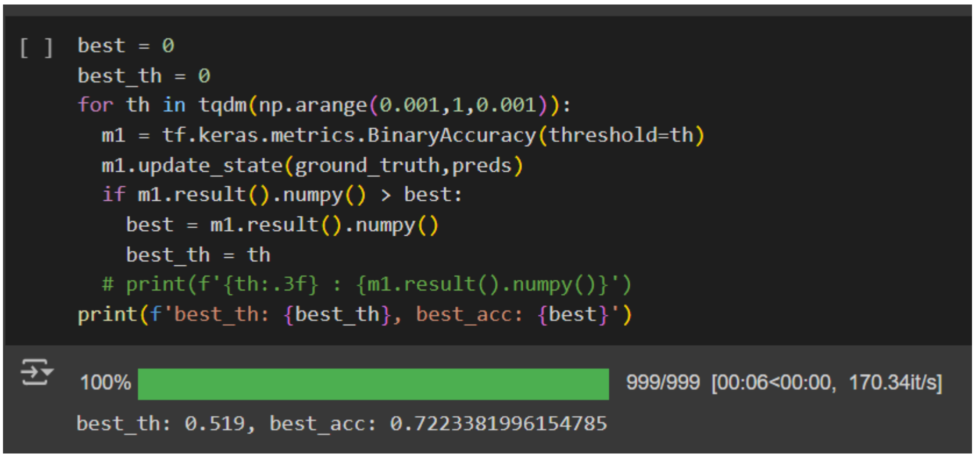

attention_mask | input: | [(None, 256)]
InputLayer | output: | [(None, 256)]
input_ids | input: | [(None, 256)]
InputLayer | output: | [(None, 256)]
numerical | input: | [(None, 9)]
InputLayer | output: | [(None, 9)]
tfxlm_roberta_model | input: | {'input_ids': (None, 256), 'attention_mask': (None, 256)}
TFXLMRobertaModel | output: | TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 256, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)
num_features | input: | (None, 9)
Sequential | output: | (None, 768)
pooled_output | input: | [(None, 768), (None, 768)]
Lambda | output: | (None, 2, 768)
features_attn | input: | (32, 1, 768)
MultiHeadAttention | output: | (32, 1, 768)
slice_layer | input: | (32, 1, 768)
Lambda | output: | (32, 768)
cls_dropout | input: | (32, 768)
Dropout | output: | (32, 768)
O | input: | (32, 768)
Dense | linear | output: | (32, 1)
C | input: | (32, 768)
Dense | linear | output: | (32, 1)
E | input: | (32, 768)
Dense | linear | output: | (32, 1)
A | input: | (32, 768)
Dense | linear | output: | (32, 1)
N | input: | (32, 768)
Dense | linear | output: | (32, 1)

Figure 4.11: Bio Numerical Model Architecture

We evaluated this model and observed the loss, the logits and the validation accuracy and we found out that this model's performances is great in comparison to the last models and did not struggle to identify the classes.

While this archtecture has significantly improved the model's performance, when we investigated the model's attention scores we found that the model focus solely on the



```
best = 0
best_th = 0
for th in tqdm(np.arange(0.001,1,0.001)):
  m1 = tf.keras.metrics.BinaryAccuracy(threshold=th)
  m1.update_state(ground_truth,preds)
  if m1.result().numpy() > best:
    best = m1.result().numpy()
    best_th = th
  # print(f'{th:.3f} : {m1.result().numpy()}')
print(f'best_th: {best_th}, best_acc: {best}')
```

```
100% 999/999 [00:06<00:00, 170.34it/s]
best_th: 0.519, best_acc: 0.7223381996154785
```

Figure 4.12: Bio Numerical Model Evaluation for C trait

numerical features in his predictions so this led us into the idea of using a model that works solely on numerical features

4.5 Numerical data Model

This model works solely on numerical features. The numerical features consists of postsCount, followersCount, foloweesCount, duration and frequency of the posting which can be synthesises from the user's profile data. The numerical data also contains some information relevant to the captions, which include the average of captions lengths, the number of hashtags and emojis and the number of mentions. The numerical features served as data for our model which we have trained using Autokeras for optimal architecture and hyperparameter tuning, the best architecture is presented below:

The model performance was great, we tracked the accuracy for each personality trait and find the best thresholds for each class as mentionned in the figure below:

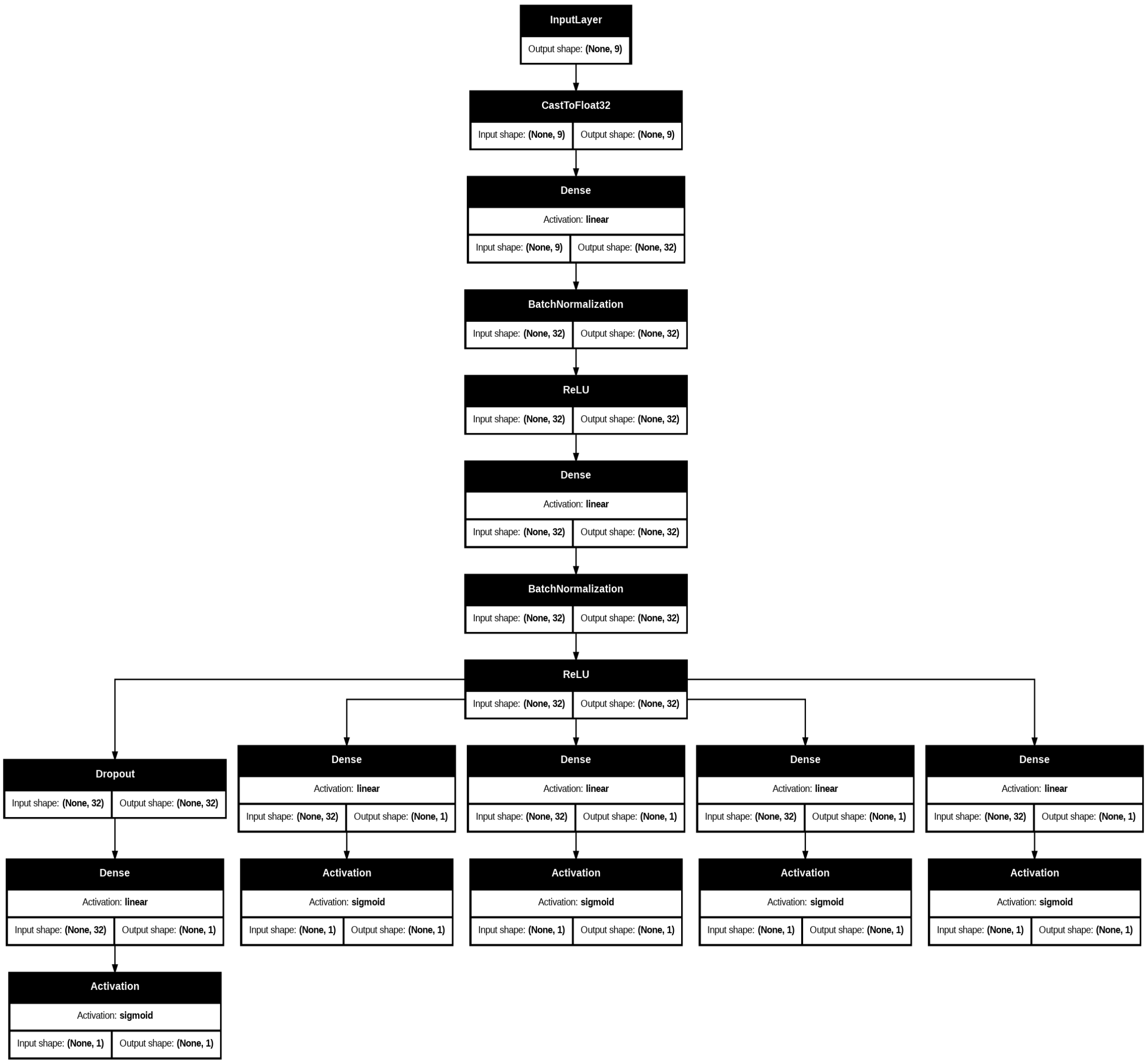
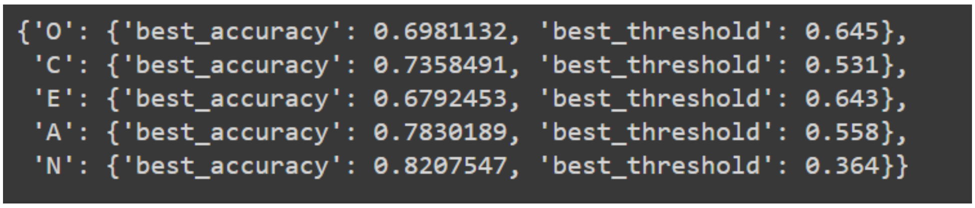


Figure 4.13: Numerical Model Architecture

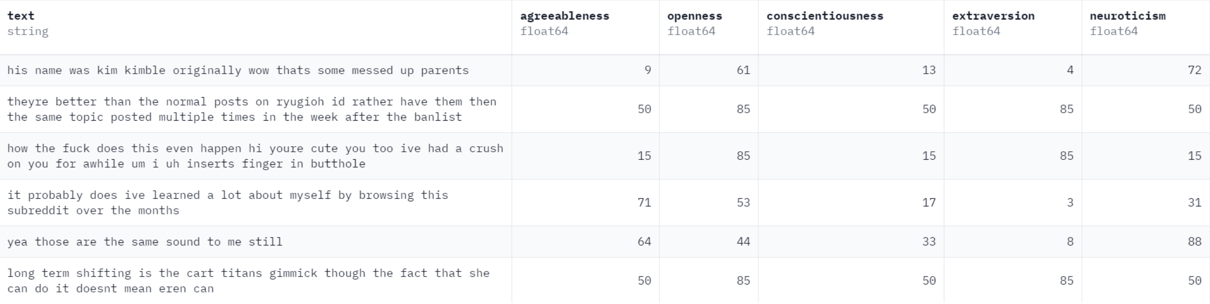


```
{'O': {'best_accuracy': 0.6981132, 'best_threshold': 0.645},
 'C': {'best_accuracy': 0.7358491, 'best_threshold': 0.531},
 'E': {'best_accuracy': 0.6792453, 'best_threshold': 0.643},
 'A': {'best_accuracy': 0.7830189, 'best_threshold': 0.558},
 'N': {'best_accuracy': 0.8207547, 'best_threshold': 0.364}}
```

Figure 4.14: Numerical Model Evaluation for C trait

5 Siamese XLM-RoBERTa Model

While searching the datasets hubs, we stumbled upon a dataset called PANDORA [7] for OCEAN personlity collected and processed by experts So we decided to use this dataset in building our AI system.



text	agreeableness	openness	conscientiousness	extraversion	neuroticism
string	float64	float64	float64	float64	float64
his name was kim kimble originally wow thats some messed up parents	9	61	13	4	72
theyre better than the normal posts on ryugioh id rather have them then the same topic posted multiple times in the week after the banlist	50	85	50	85	50
how the fuck does this even happen hi youre cute you too ive had a crush on you for awhile um i uh inserts finger in butthole	15	85	15	85	15
it probably does ive learned a lot about myself by browsing this subreddit over the months	71	53	17	3	31
yea those are the same sound to me still	64	44	33	8	88
long term shifting is the cart titans gimmick though the fact that she can do it doesnt mean eren can	50	85	50	85	50

Figure 4.15: 5 personality traits Dataset

While investigating the dataset, we come to a solution on how to use it with our project, we used siamese xlm roberta to predict the similarities between the users' captions and texts found in the PANDORA dataset and based on these similarites we calculate the weighted sum of the personalities to get our final predictions

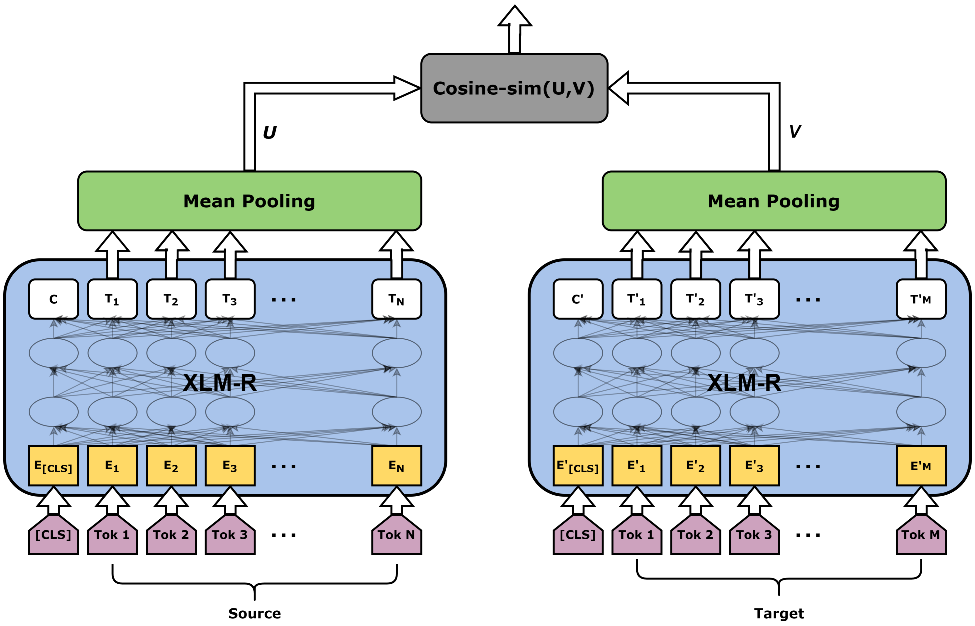


Figure 4.16: Siamese XLM-RoBERTa Model Architecture

Here's a simple output of the model for a certain Instagram user

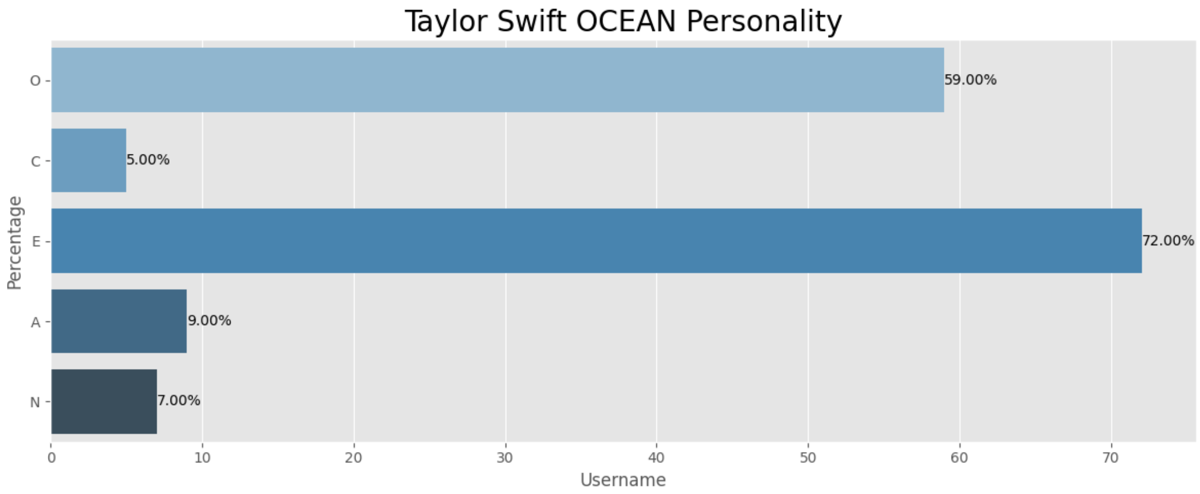


Figure 4.17: example output

6 Comparison between models

Table 4.1: Comparison between models

model	**Accuracy**
XLM-RoBERTa	0.5882
ViT Model	0.5882
XLM-RoBERTa + ViT Model	0.5882
Biography and Numerical data Model	0.72
Numerical data Model	• **Openness :0.7** • **Conscientiousness :0.73** • **Extraversion :0.67** • **Agreeableness :0.78** • **Neuroticism :0.82**
Siamese XLM-RoBERTa Model	• **Openness :0.7** • **Conscientiousness :0.55** • **Extraversion :0.7** • **Agreeableness :0.6** • **Neuroticism :0.7**

5

Deployment

1 Introduction

In this section, we will showcase the Deployment phase where we successfully integrated the trained model into a live environment where it can be used to make personality predictions in real-time. This phase included creating a visually attractive and interactive web interface for the goal of Visualization

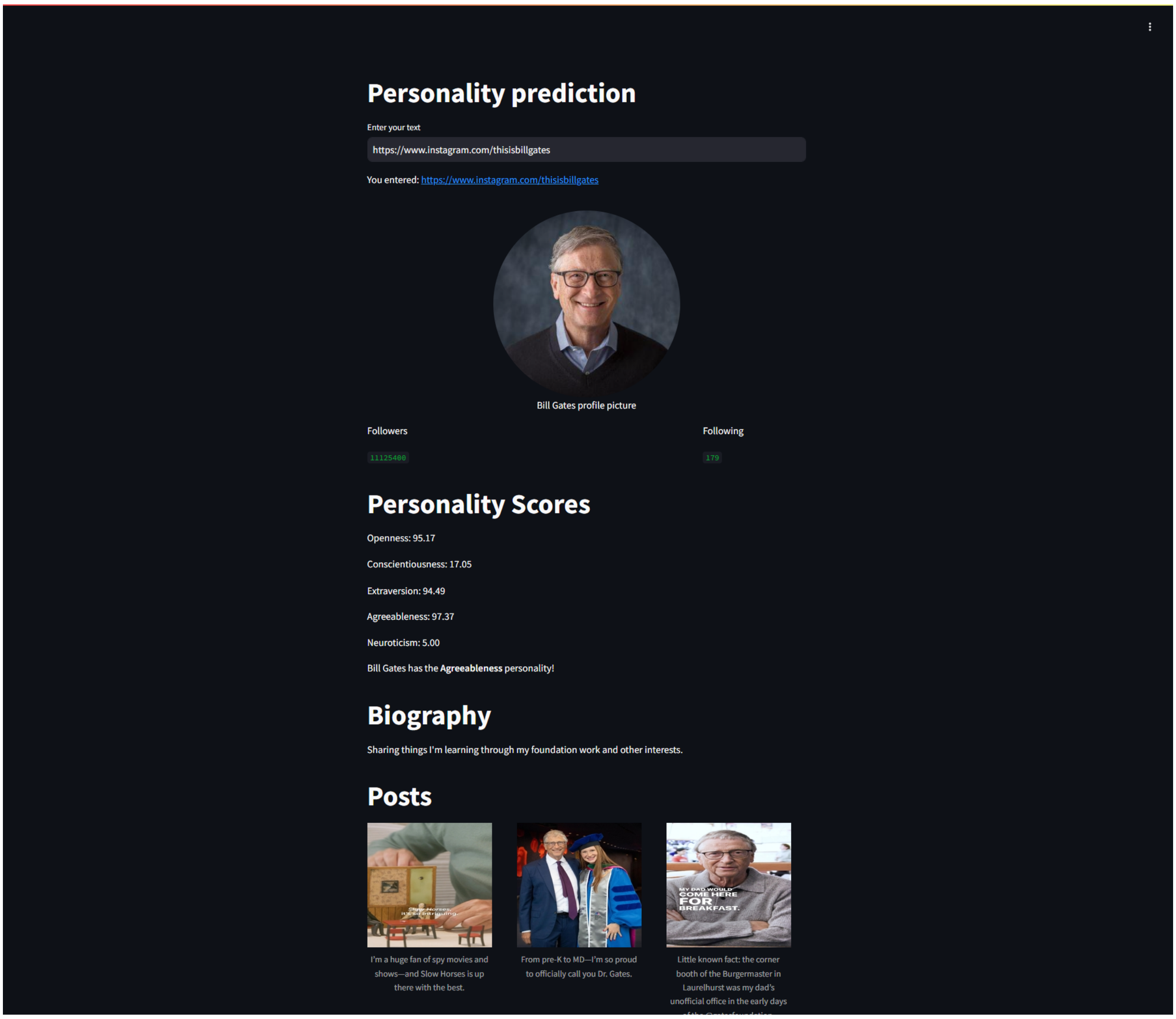
2 Tools

The Streamlit library(Figure 5.2) was the main tool used for developing the web application.

Streamlit is an open-source Python library designed for creating interactive, web-based applications for data science and machine learning projects. It allows users to quickly build and deploy custom applications using simple and intuitive syntax.

Figure 5.1: Streamlit

3 Process overview



Personality prediction

Enter your text

https://www.instagram.com/thisisbillgates

You entered: https://www.instagram.com/thisisbillgates

Bill Gates profile picture

Followers

11125400

Following

179

Personality Scores

Openness: 95.17

Conscientiousness: 17.05

Extraversion: 94.49

Agreeableness: 97.37

Neuroticism: 5.00

Bill Gates has the **Agreeableness** personality!

Biography

Sharing things I'm learning through my foundation work and other interests.

Posts

I'm a huge fan of spy movies and shows—and Slow Horses is up there with the best.

From pre-K to MD—I'm so proud to officially call you Dr. Gates.

Little known fact: the corner booth of the Burgermaster in Laurelhurst was my dad's unofficial office in the early days

Figure 5.2: final result

Predicting personalities through our web application consists of these steps:

1. The user enters the instagram profile URL of the person that he wants to predict his personality.
2. The application scrapes information about the searched profile such as his name, his biography, his followers and followees count and his posts. The informations are then rendered in the page.
3. The user can then press the "Predict Personality" button to send the collected data to the model and predict the right personality of the desired person.

4 Conclusion

In this last chapter, we showcased the degree to which the final result meets the business objectives. We then presented the main interfaces and the technical choices we used to develop our web application and deploy our model.

Our application can also be improved.As a first step we may try to provide an additional prediction based on the 16 personalities and let the user choose between which prediction type he wants to use(Big 5 personalities vs the 16 personalities).

Other possible improvement is to develop a support for other social networks like Facebook or Tiktok

Conclusion and Perspectives

We focused on developing a pipeline for detecting Instagram users' personalities based on various features and data sources. By scraping Instagram profiles data, and using various personality and psychology sources for labeling the dataset, we have made significant progress in improving the accuracy and prediction capabilities of the model.

During the project, we faced and conquered various challenges. Grasping the complexities of Instagram user data and the subtleties of detecting users' personalities demanded significant effort and time. Moreover, integrating diverse data sources, including textual information and image data, and effectively merging them to improve the model's performance were crucial steps in achieving satisfactory outcomes.

Looking ahead, deploying the model in a real-world environment and conducting comprehensive testing and validation will be essential. Collaborating with domain experts and utilizing their insights can significantly enhance the model's refinement and ensure its practical applicability for user detection tasks.

In conclusion, this project has made substantial strides in detecting Instagram users personality by leveraging diverse data sources and employing effective modeling techniques. By addressing the areas identified for improvement and considering the outlined future perspectives, the model can be further refined, resulting in enhanced accuracy and broader applicability for user detection tasks on Instagram and potentially other social media platforms like Facebook and TikTok.

Bibliography

[1] Website seilinium. https://www.selenium.dev.

[2] Website apify. https://apify.com.

[3] What is crisp dm? datascience.pm. https://www.datascience-pm.com/crisp-dm-2/.

[4] Website pdb. https://www.personality-database.com/profile/1008/miley-cyrus-pop-contemporary-mbti-personality-type.

[5] Website crystalknows. https://www.crystalknows.com/famous-people.

[6] Website boo world. https://boo.world/fr/database/infps.

[7] here's the link of the dataset. https://huggingface.co/datasets/Fatima0923/Automated-Personality-Prediction.