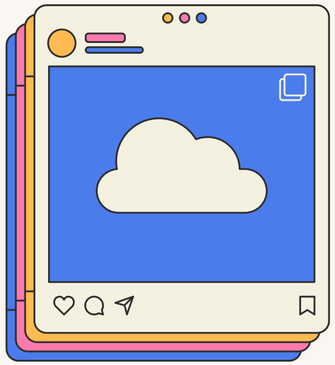
End of Year Project

Personality detection of Instagram users

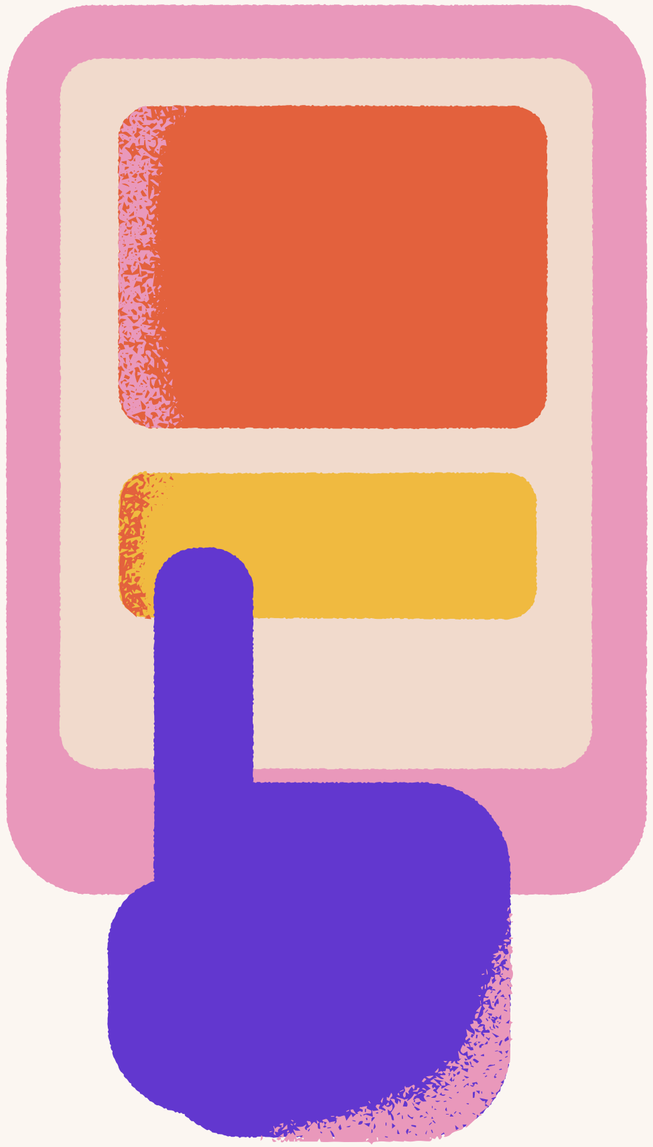
Made by :
Ben Omrane Mohamed Salim
Turki Mohamed Seddik
Hammami Omar
Abdelkader Iheb

Supervisor : Mrs. Sana Hamdi

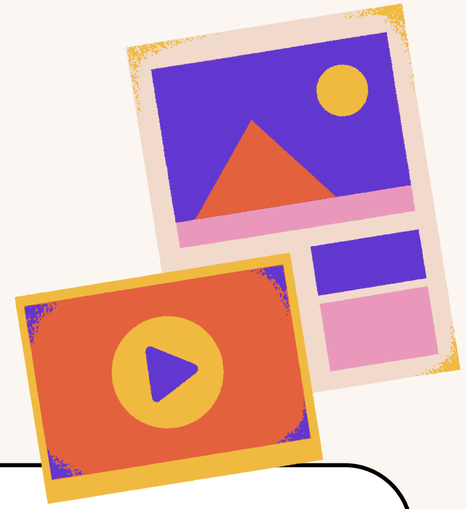
Reviewer : Mrs. Wided MILED SOUID



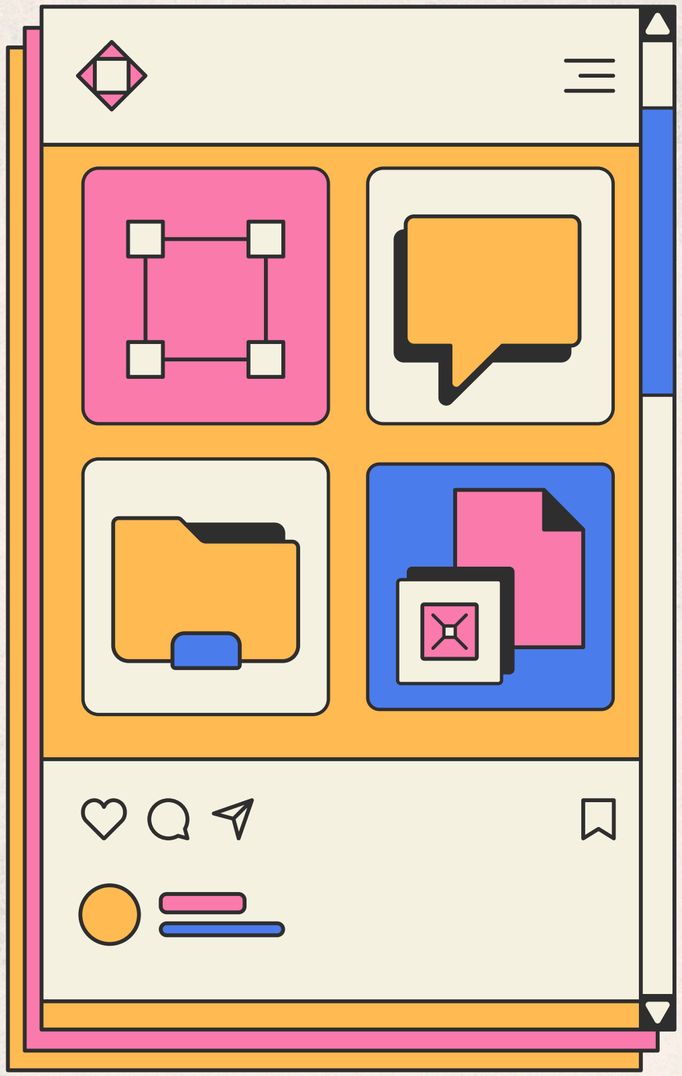
Study year: 2023/2024



Agenda



- Project Overview
- Methodology and Project lifeCycle
- Data Gathering and Understanding
- Modelling and Evaluation
- Final Results and Conclusion

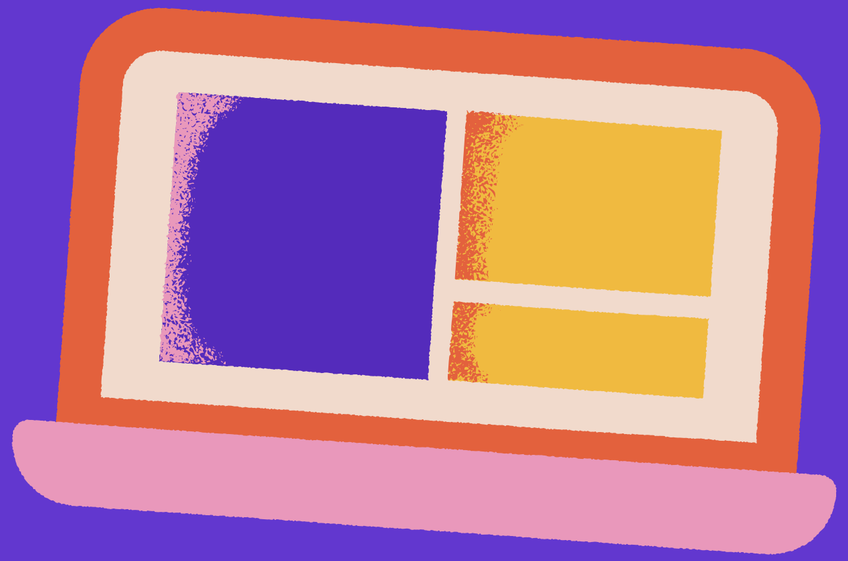


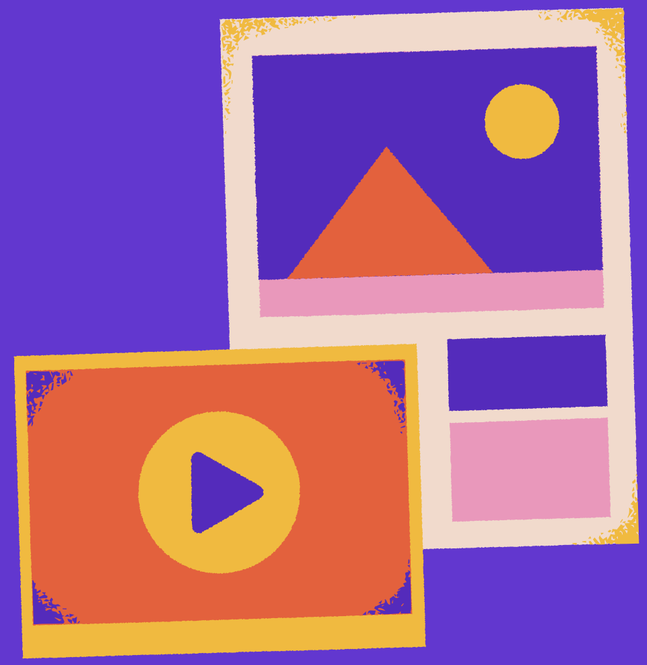
Project Overview

In the age of social media, Instagram profiles offer a window into the personalities of users through their shared images and text. This project leverages artificial intelligence to analyze Instagram content, aiming to predict personality traits accurately.

By understanding these personality traits, our AI technology can unlock new potentials across various domains, enhancing business operations, refining user engagement strategies, and streamlining service personalization.

This initiative not only showcases the integration of AI with psychological insights but also promises substantial advancements in how we interact with digital content.

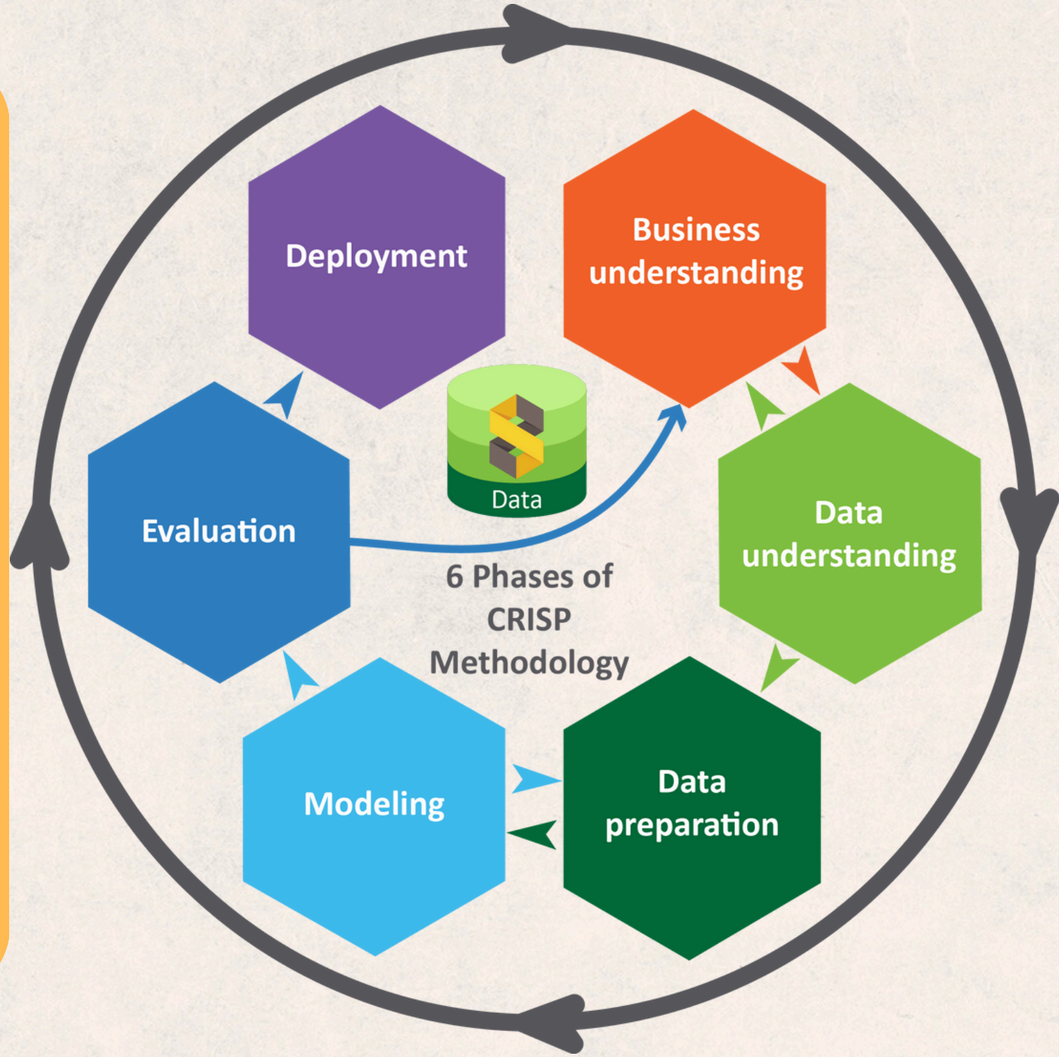




Methodology and Project LifeCycle

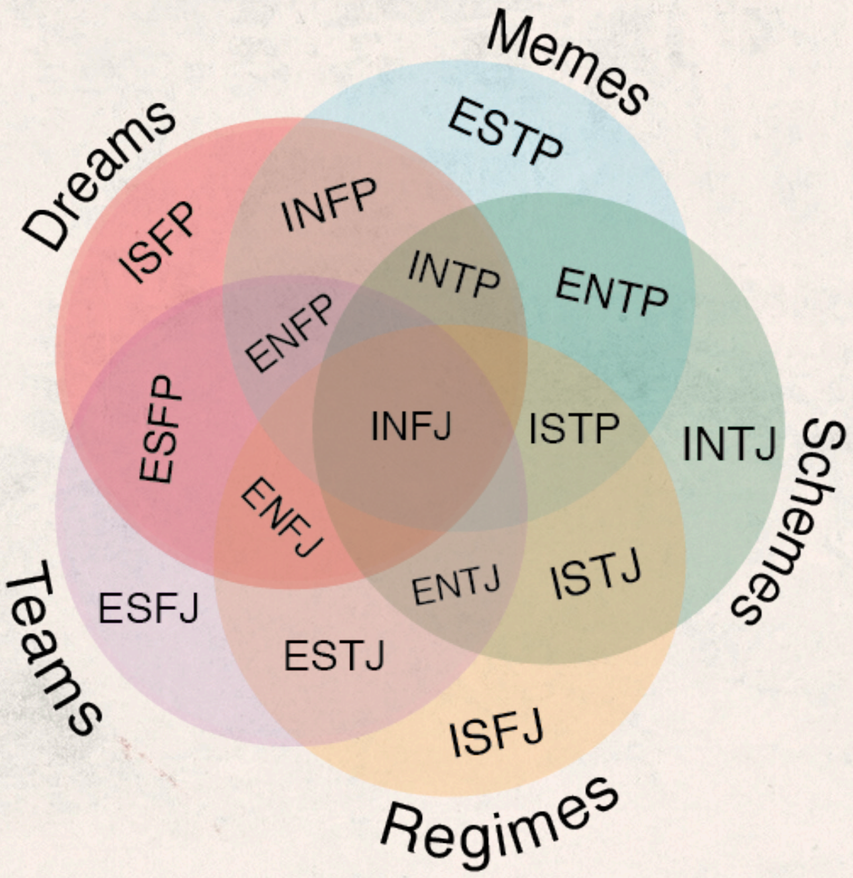
Methodology and Project LifeCycle

- WE ADOPTED THE CRISP METHODOLOGY
- ROBUST AND WELL PROVEN



Bussiness Understanding

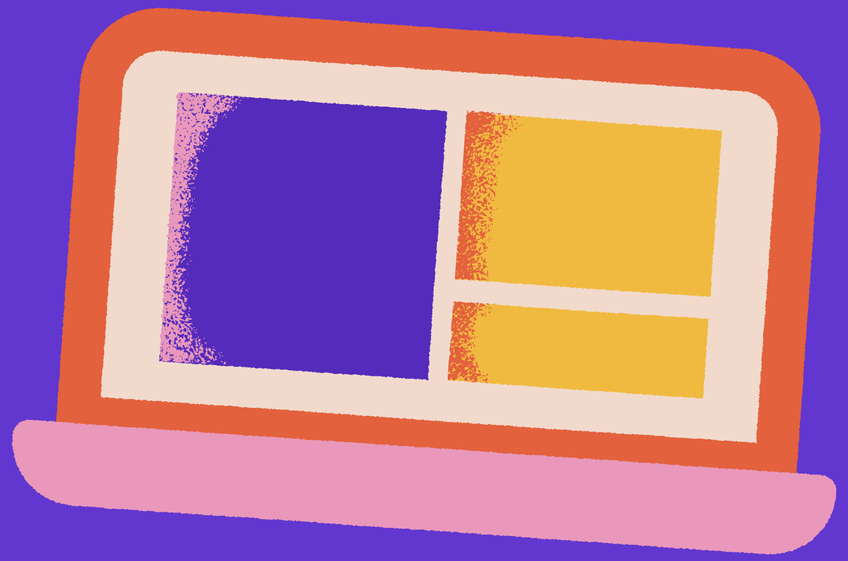
Myers-Briggs Type Indicator (MBTI)

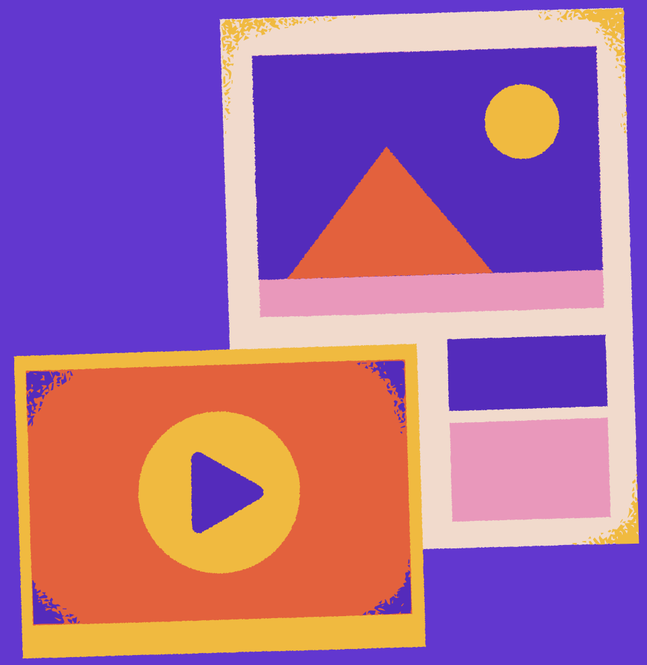


Big Five Personality Traits

WHY MBTI ?

- Detailed classification into 16 personality types provides a nuanced framework that is particularly effective for analyzing user behavior.
- Aids in understanding and predicting content preferences, making it valuable for targeted marketing and personalization strategies.
- Widespread recognition and ease of interpretation make it accessible to a broad audience, aligning well with Instagram's diverse user base.





Data Gathering and Data Understanding

DATA GATHERING

TEXTUAL

VISUAL

SOLUTIONS

PROS

- Handles dynamic content.
- It can interact with web elements like a real user.

CONS

- Slower and resource-intensive due to browser automation.

PROS

- User-friendly Interface
- Easily handles large-scale scraping operations with efficient resource management.

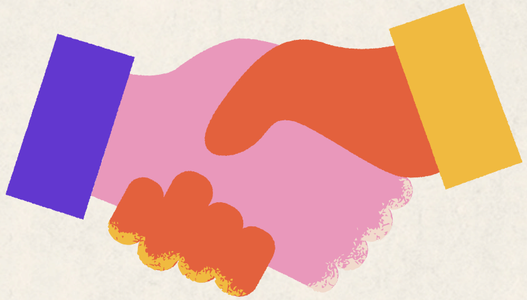
CONS

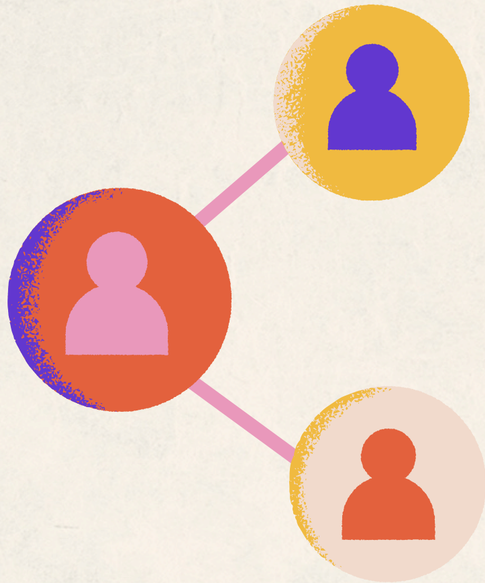
- Can be expensive, especially for extensive or frequent scraping tasks.

APIFY

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.

DATASET





SIZE

800+ Scraped users.

POSTS

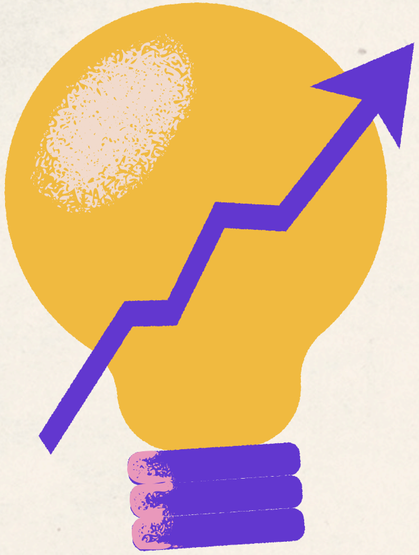
More than 100000 scraped posts.

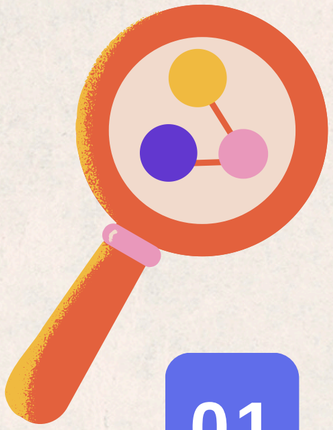
LANGUAGES

More than 40 languages.

TYPES

Athletes, Influencers, Singers, Actors, Writers, Streamers...etc





Data Labeling

01

We used the BOO website to identify the personality type of the target

02

We also used the personality database website to identify the personality type of the target

03

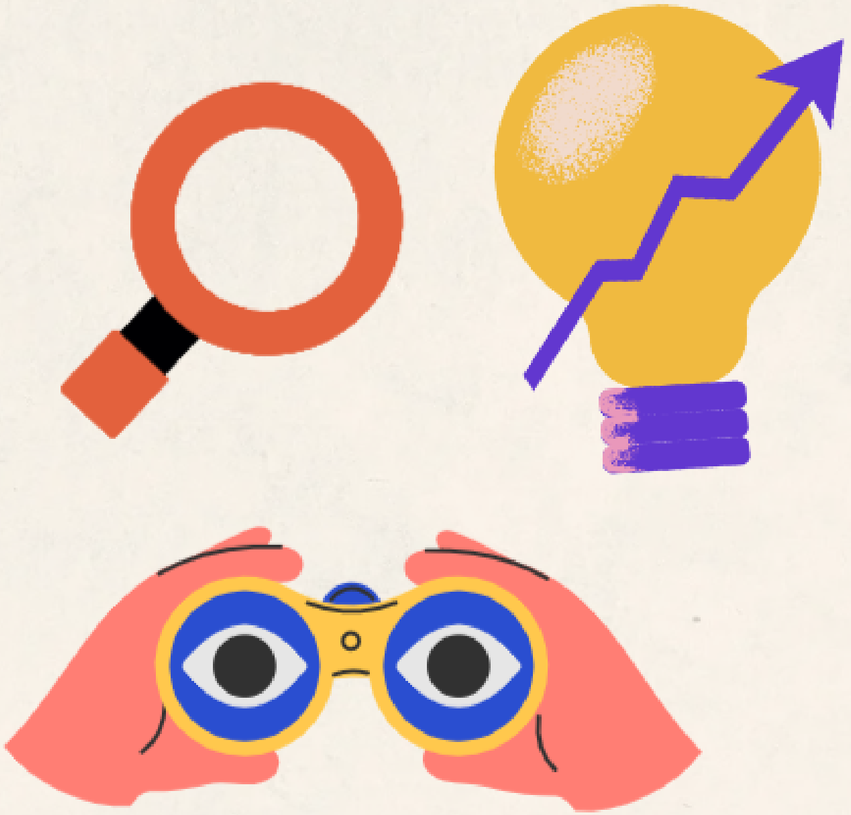
We finally used Crystal, a personality data platform to verify the personality type.

04

In case we didn't find the target's personality neither on these websites nor on the internet we can proceed to the manual identification method

Data Understanding

- TO GAIN INSIGHTS FROM OUR DATA, WE PERFORMED VARIOUS VISUALIZATIONS



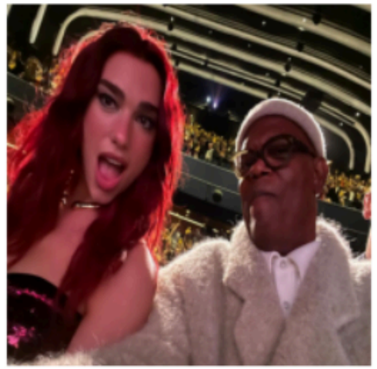
SAMPLE FROM THE DATASET

Shakira

Cohete #LMYNL

Samuel L Jackson

ARGYLLE PREMIERE



Neha Kakkar

#EnnaAkhiyan ♥

Alok

Meryl Streep

❤ #sagawards

Jason Statham

#StoneIslandSS24 @davidsimsofficial

JACLYN TORREY

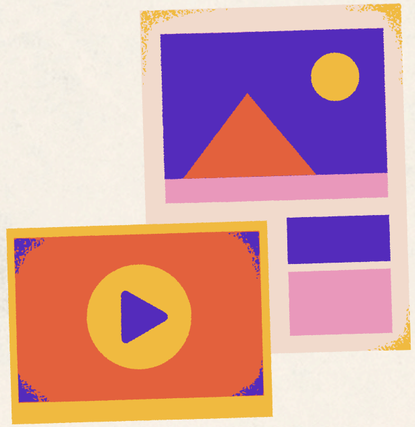
Okay, Tom...

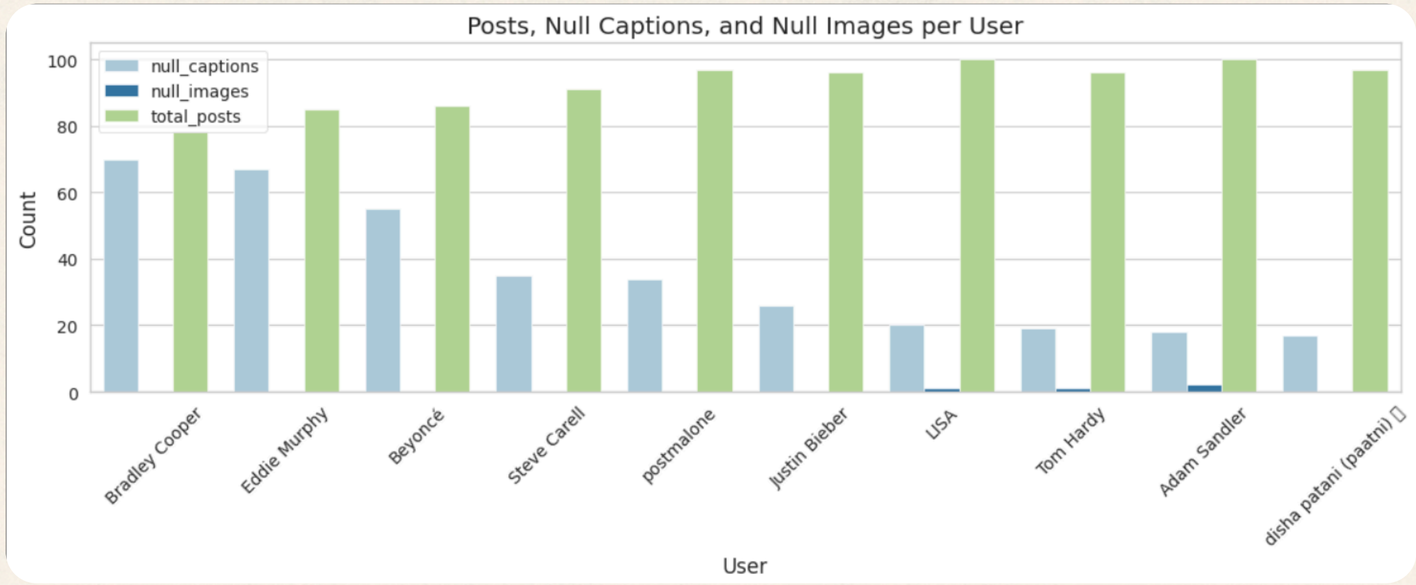
Niall Horan

Night Twenty-Two Düsseldorf

INSTANCES WITH NULL CAPTIONS AND NULL IMAGES

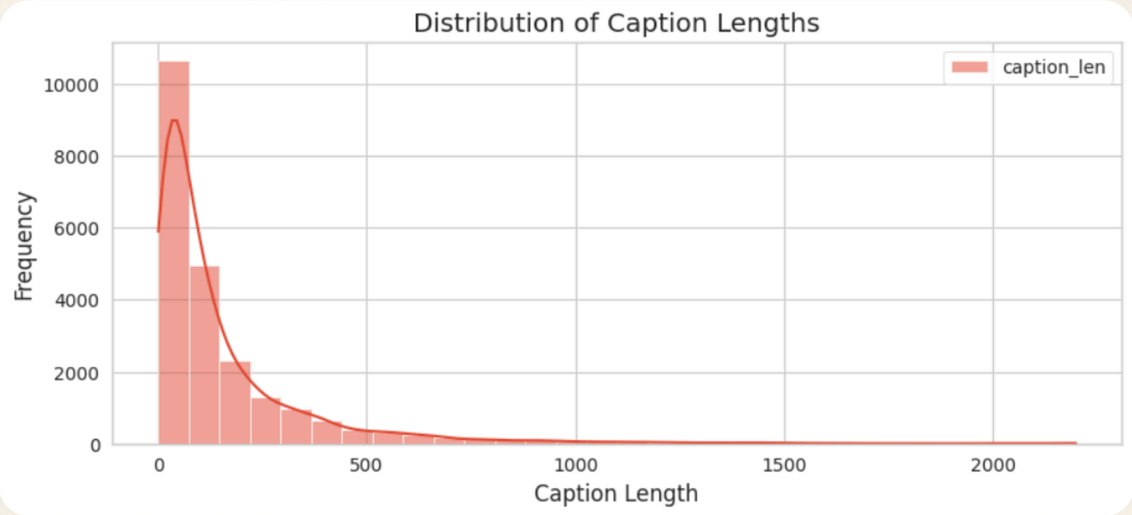
	fullName	null_captions	null_images	total_posts
0	21savage	5	0	100
1	50 Cent	0	0	100
2	6ix9ine	0	1	100
3	ANADEARMAS	3	2	100
4	AKON	13	0	100





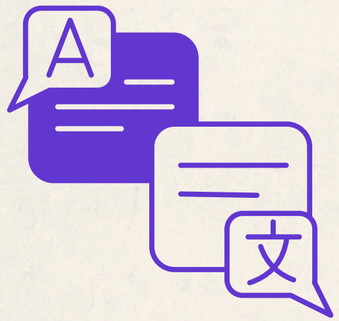
Visualizations related to captions

DISTRIBUTION OF CAPTION LENGTHS



MEAN OF CAPTION LENGTHS PER USER

	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



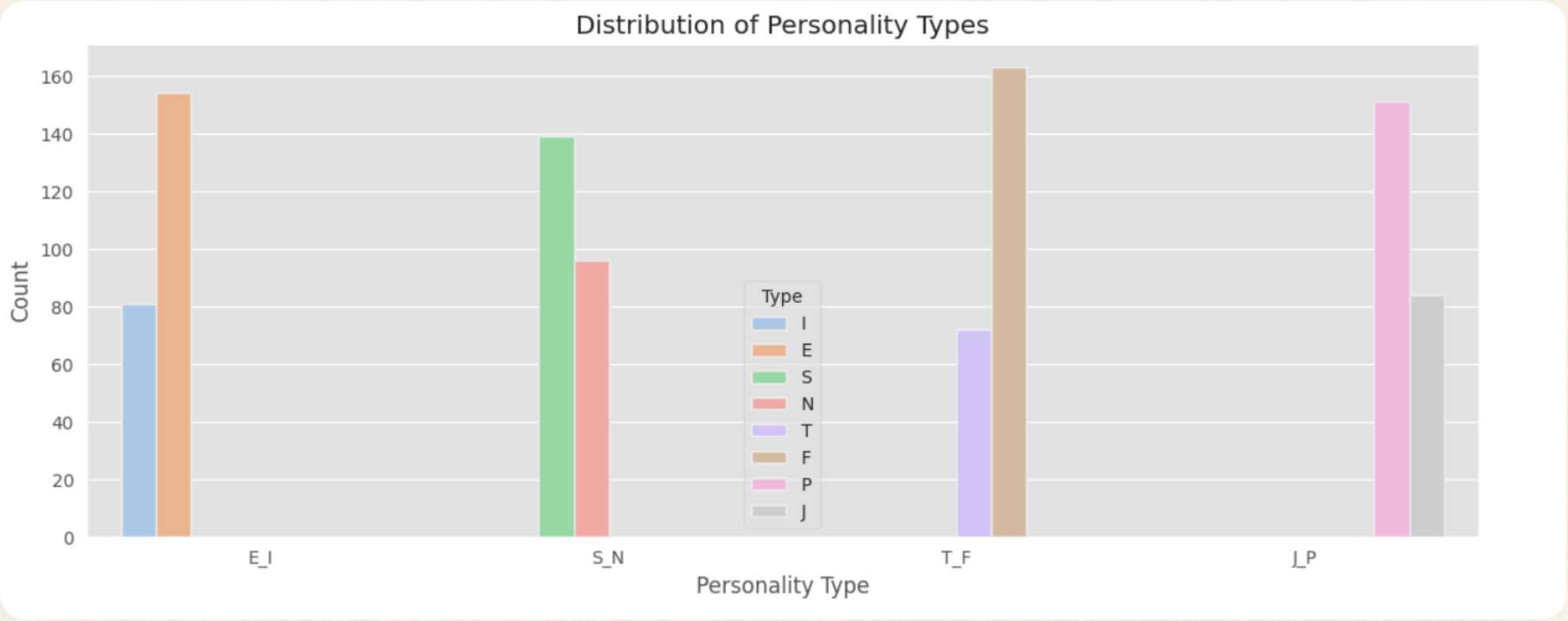
LANGUAGES USED IN CAPTIONS

	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

LANGUAGES USED IN CAPTIONS PER USER

	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	A N A D E A R M A S	{so, sl, it, fr, en, tr, fi, pl, ro}
4	AKON	{cs, pt, hr, it, ca, sw, et, fr, de, en, af, l...

Personality Distribution



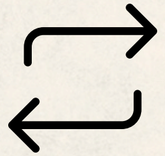
Data cleaning

- DIFFERENT TYPES OF DATA (NUMERICAL, IMAGES, AND TEXT) REQUIRE DIFFERENT CLEANING TECHNIQUES. HERE ARE SOME TECHNIQUES FOR EACH TYPE

General Data Cleaning

- REMOVE DUPLICATE ROWS

- CONVERT COLUMNS TO APPROPRIATE DATA TYPES



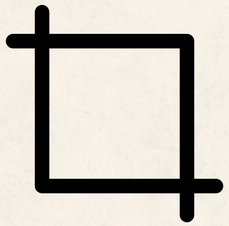
- REMOVE ROWS WITH NULL VALUES

Numerical Data

- REMOVE MISSING VALUES
- SCALING AND NORMALIZATION

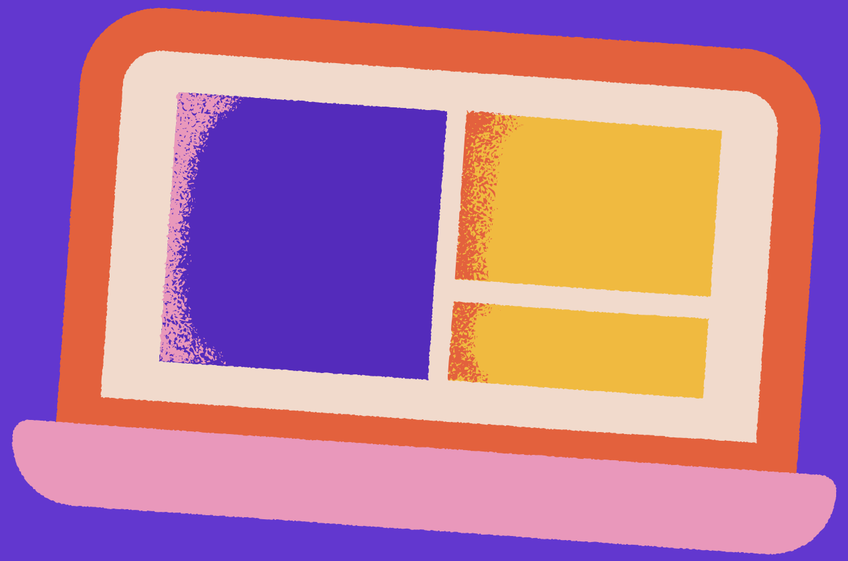
Image Data

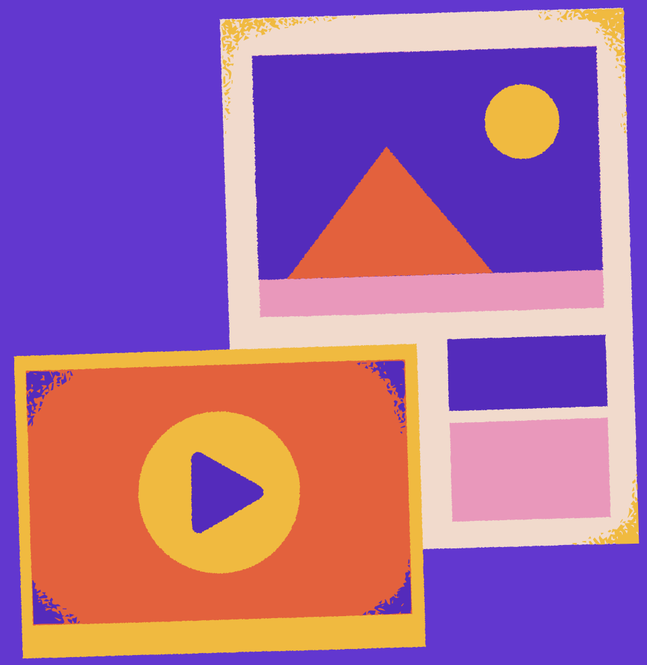
- RESIZING IMAGES



Text Data

- REMOVE NULL VALUES



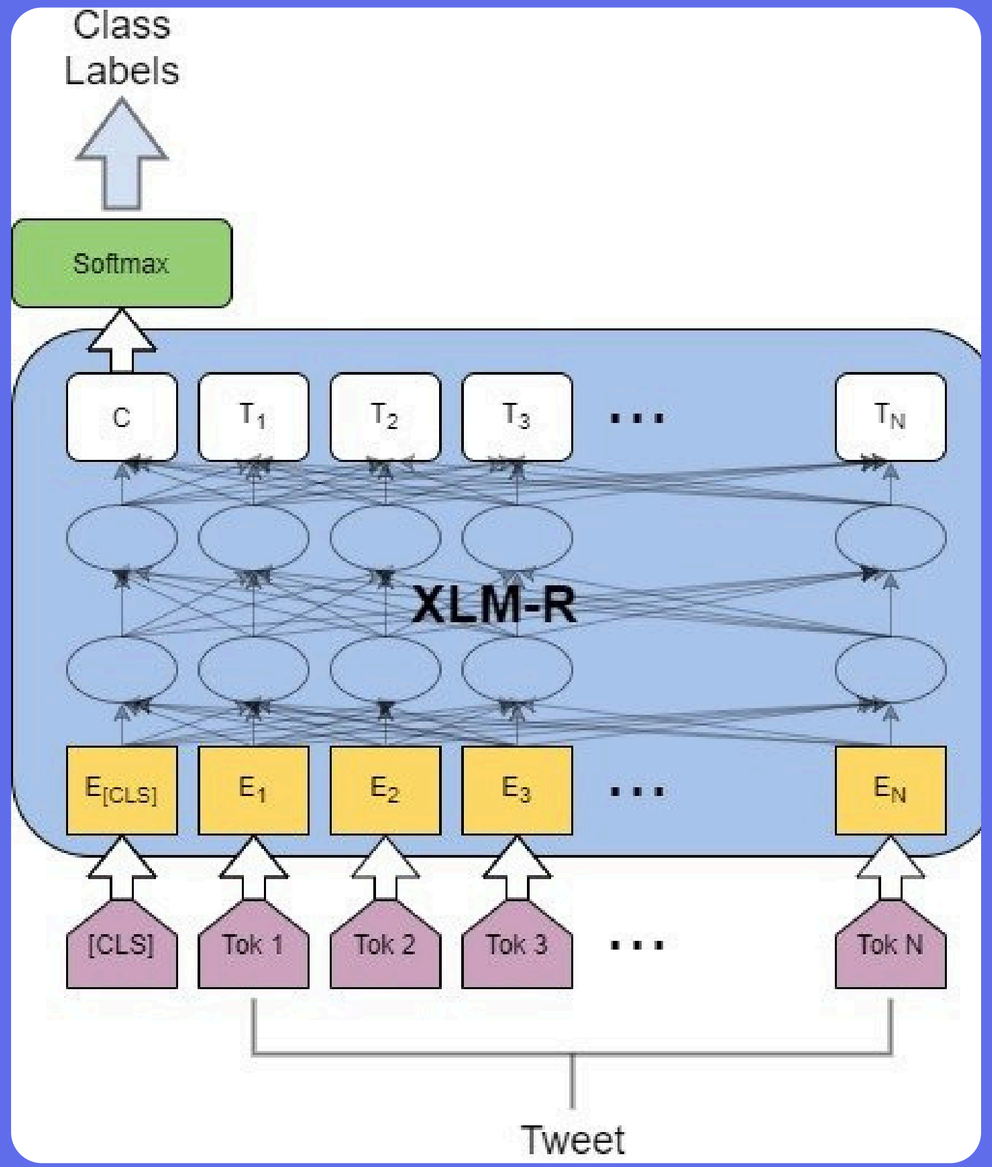


Modelling and Evaluation

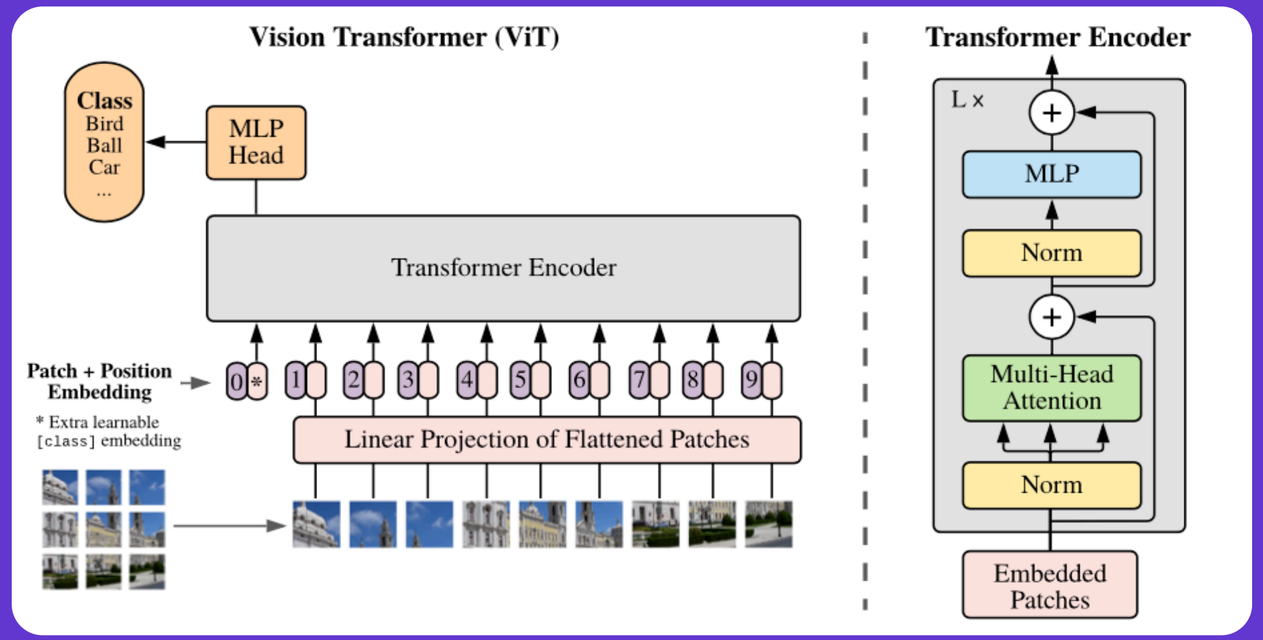
Model Selection

XLM-RoBERTa

we chose XLM-RoBERTa model as our text feature extraction due to its capacity to understand both multilingual text, emojis and hashtags.



Model Selection



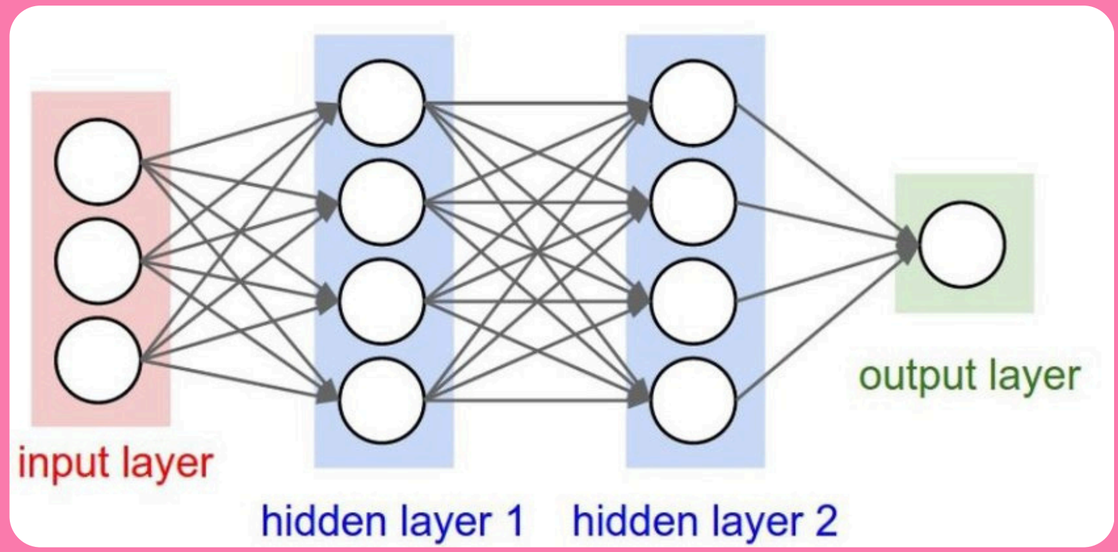
VIT

To extract image features we opted for ViT,which excels in capturing the relationships between the different parts of the image

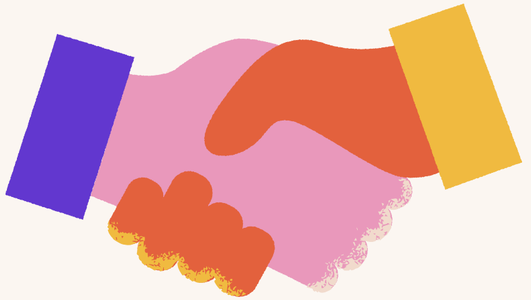
Model Selection

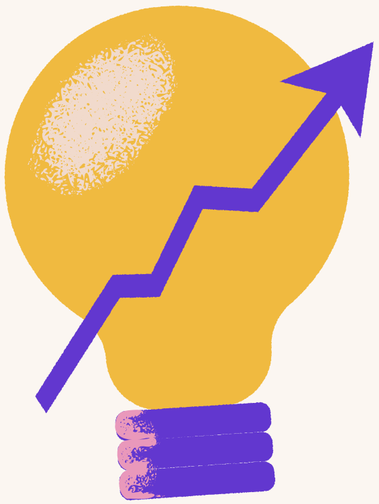
DNN

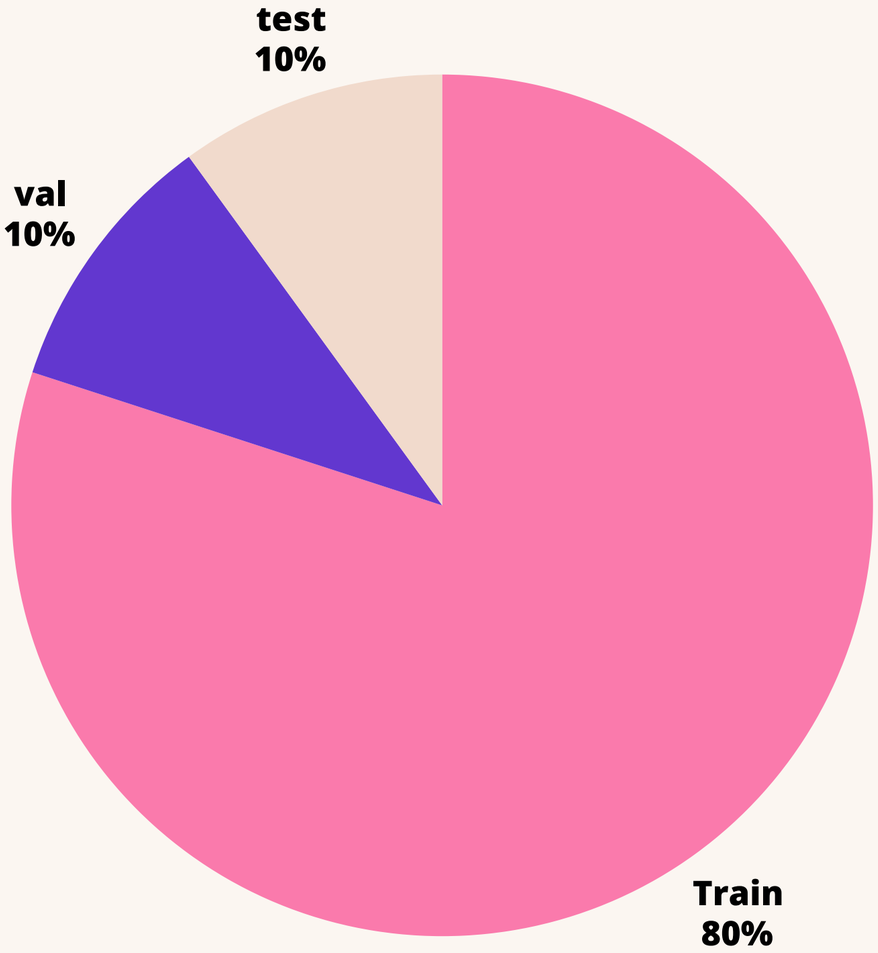
For numerical features we chose to use a simple DNN due to the low dimensionality of the numerical features

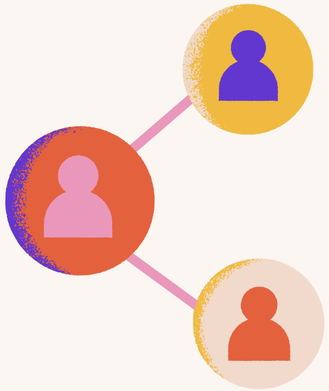


Evaluation Setup





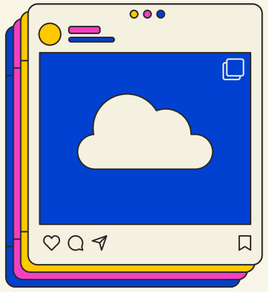




Model Architectures

1 XLM-RoBERTa

2 ViT



3 XLM-RoBERTa + ViT

4 Bio + Numerical

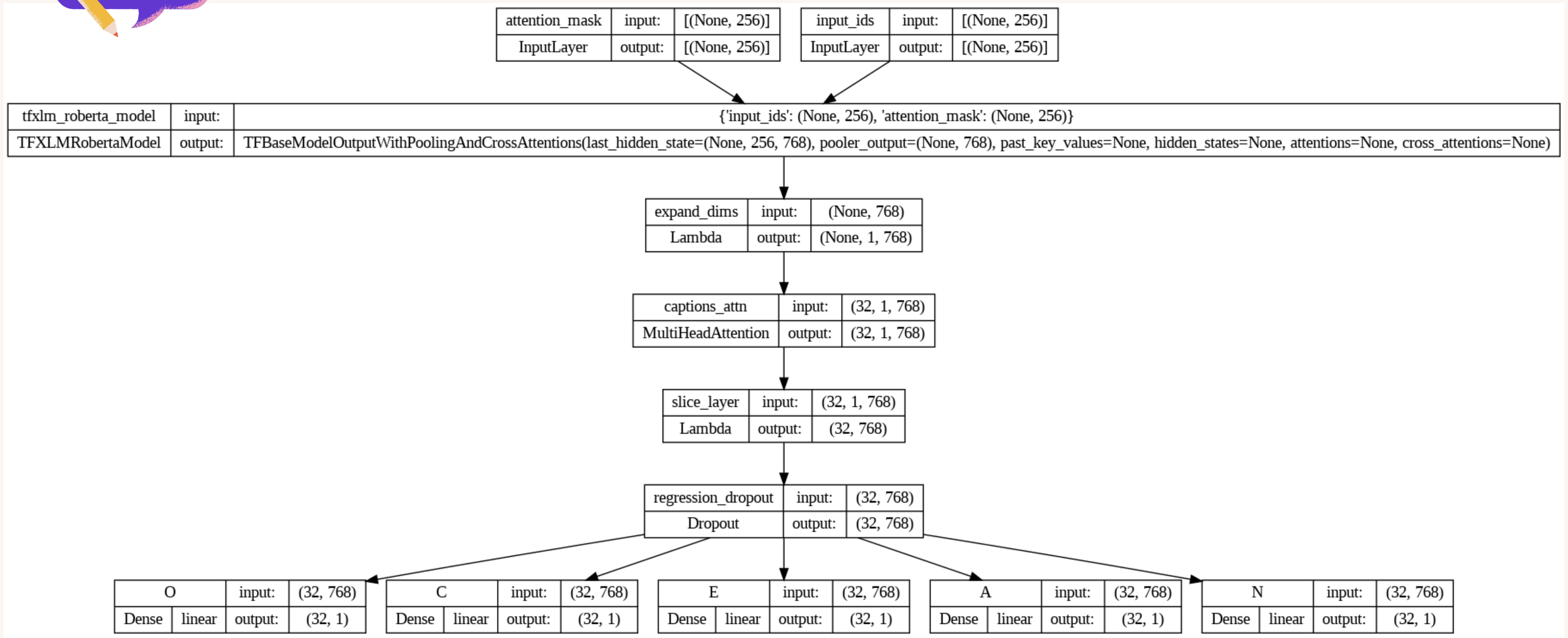
5 Numerical

1

XLM serves as a backbone for our model to extract text features, these features passes throught an Attention layer to select the best captions that best describes the user’s personality

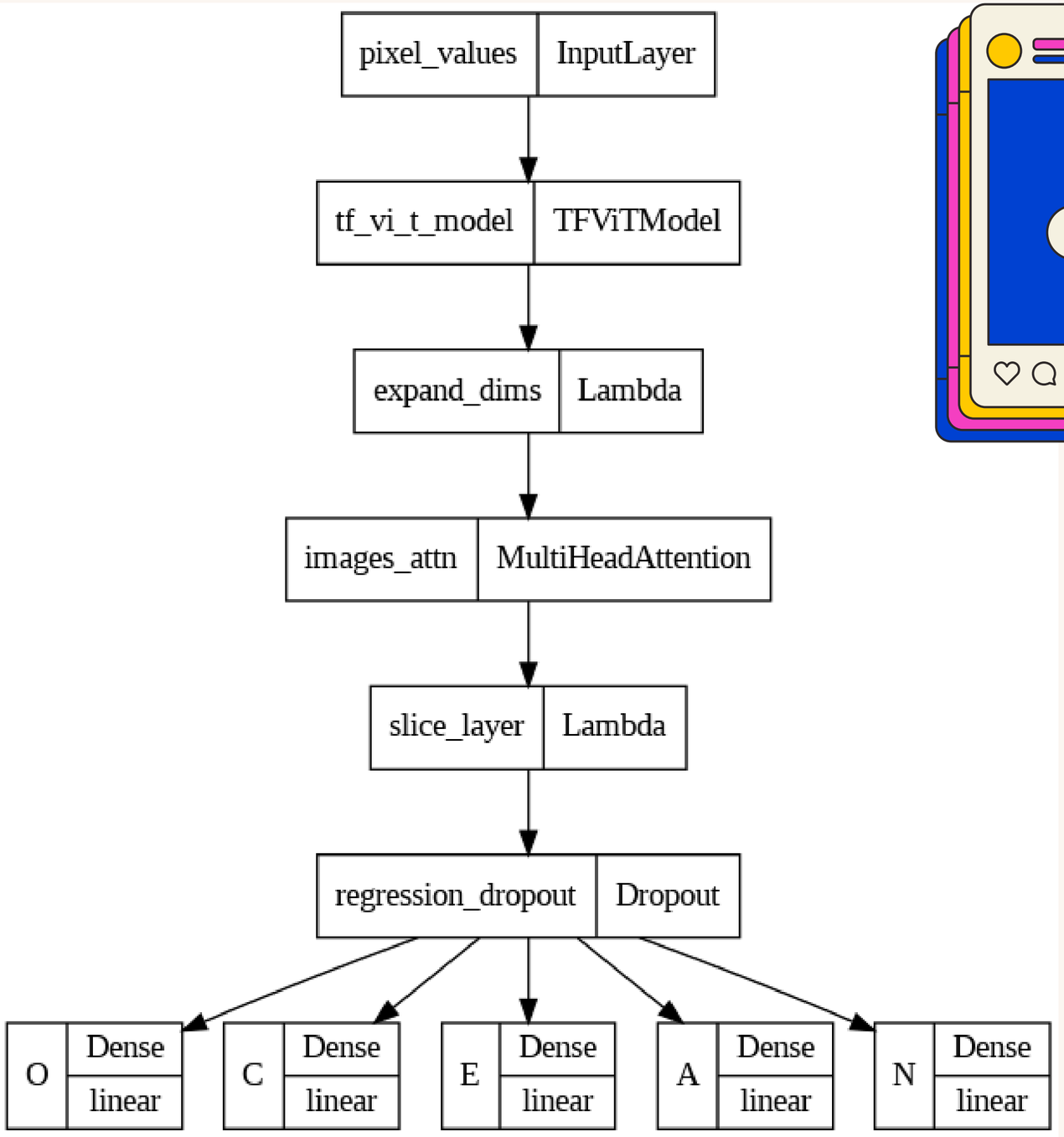
XLM-R for Text

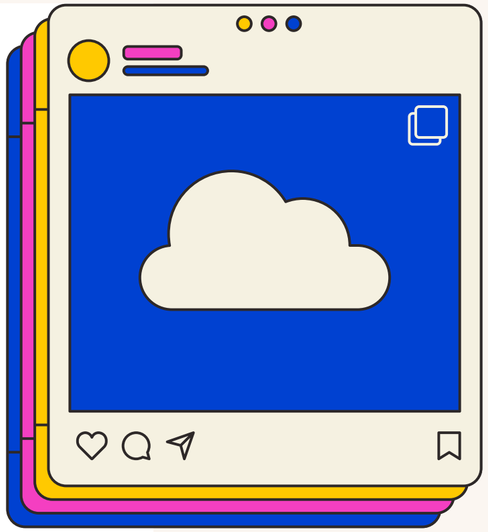
Model Architecture



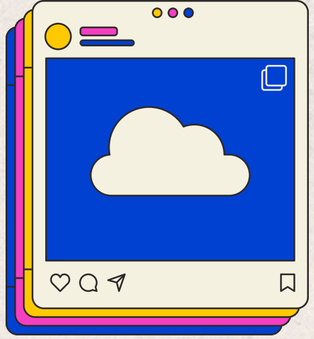
2

ViT serves as a backbone for our model to extract Image features and goes throught an Attention layer to select the best images related to the user’s personality





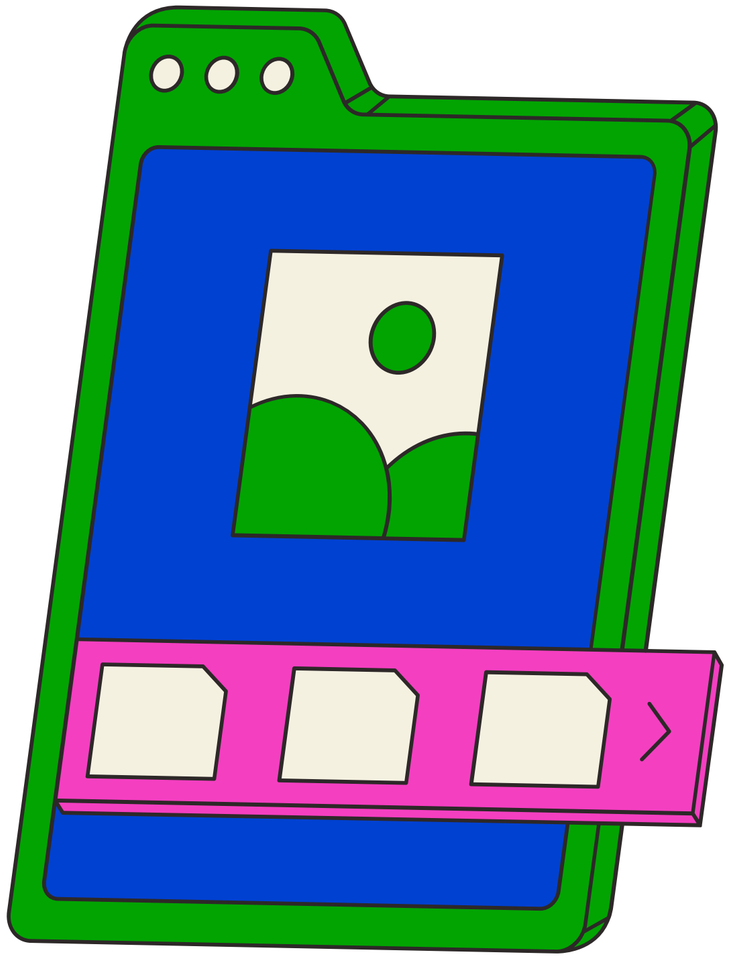
Model Architecture



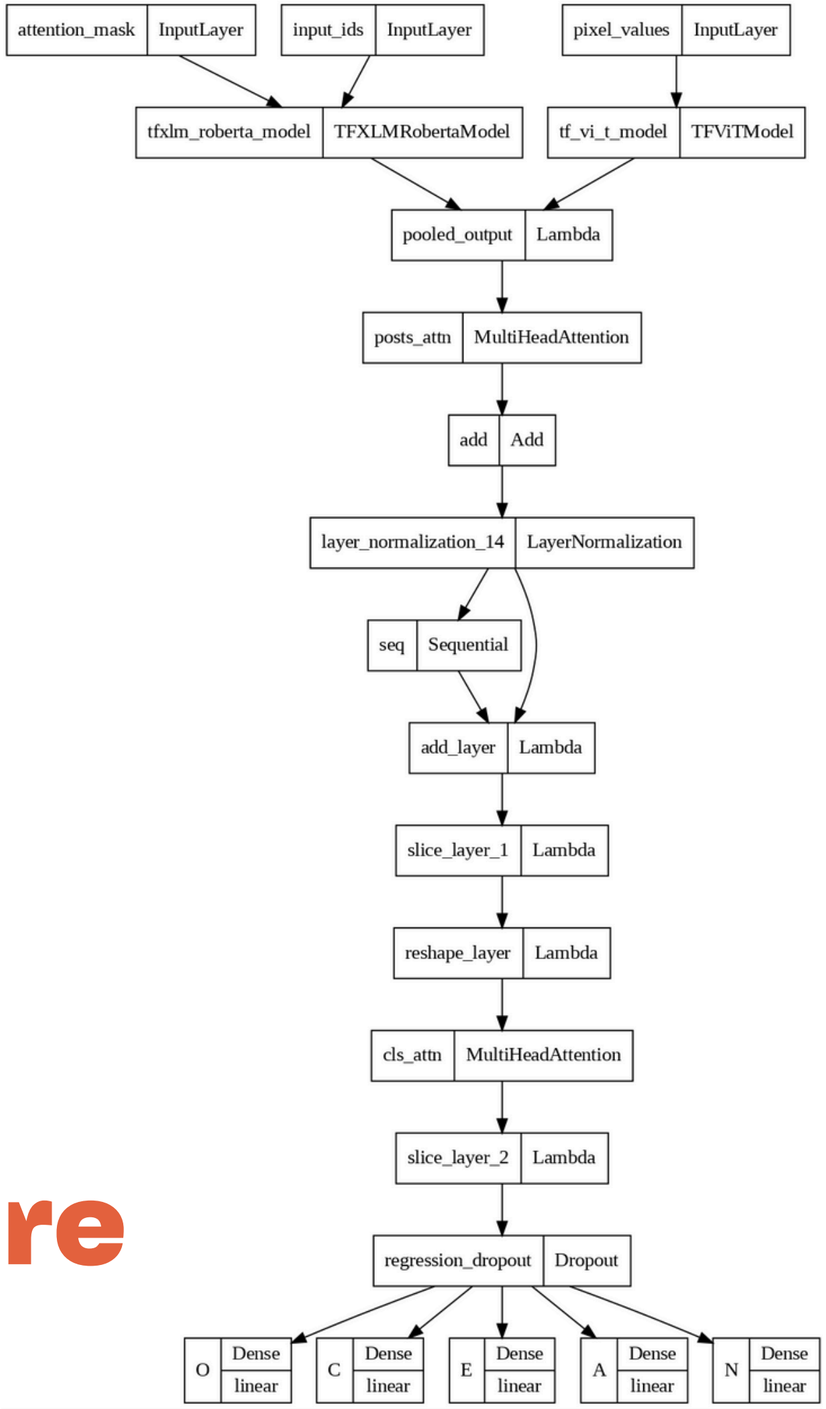
XLM-R + ViT

3

XLM-R extracts text features and ViT images’, these two features passes throught an attention layer to combine the informations from the same post to get a single post feature, these new posts features passes throught another attention layer to select the best posts for predictions

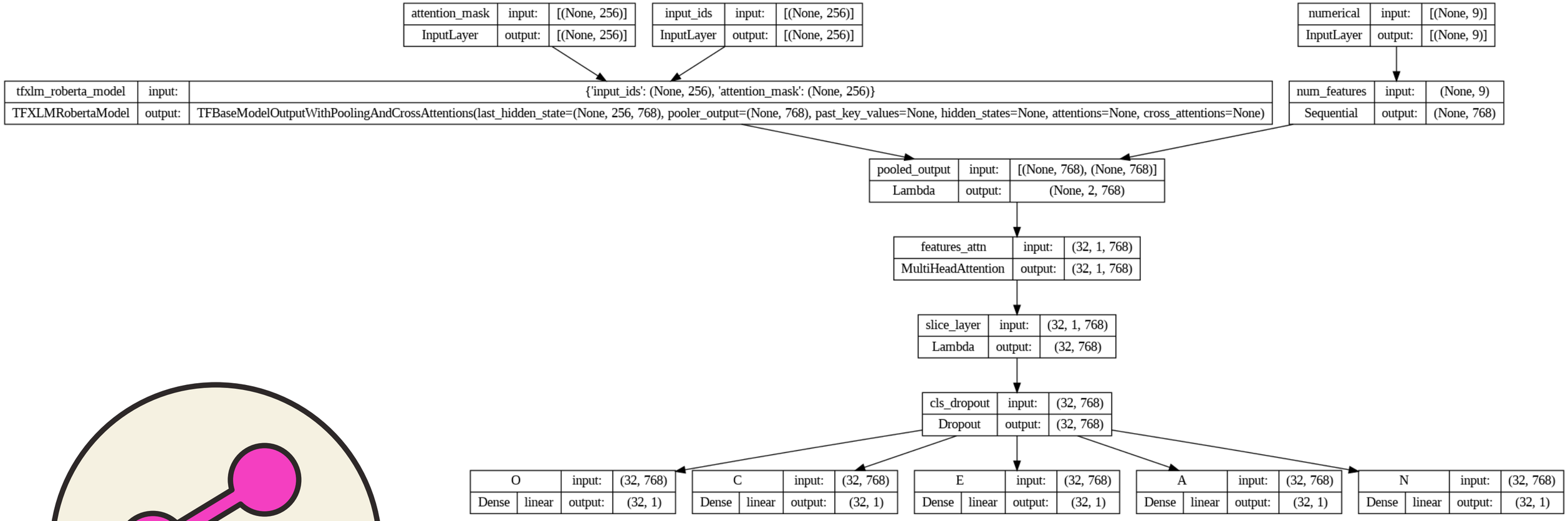


Model Architecture



4

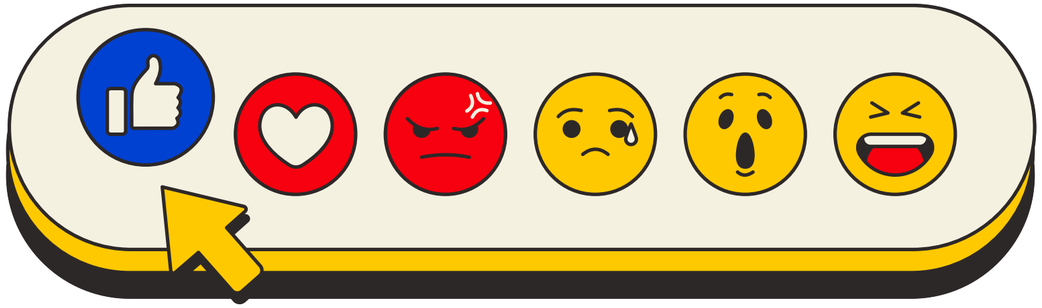
The model now works with Biography and Numerical data. XLM-R extracts the bio features and the simple DNN extracts the numerical features, then these two are combined with attention mechanism for prediction



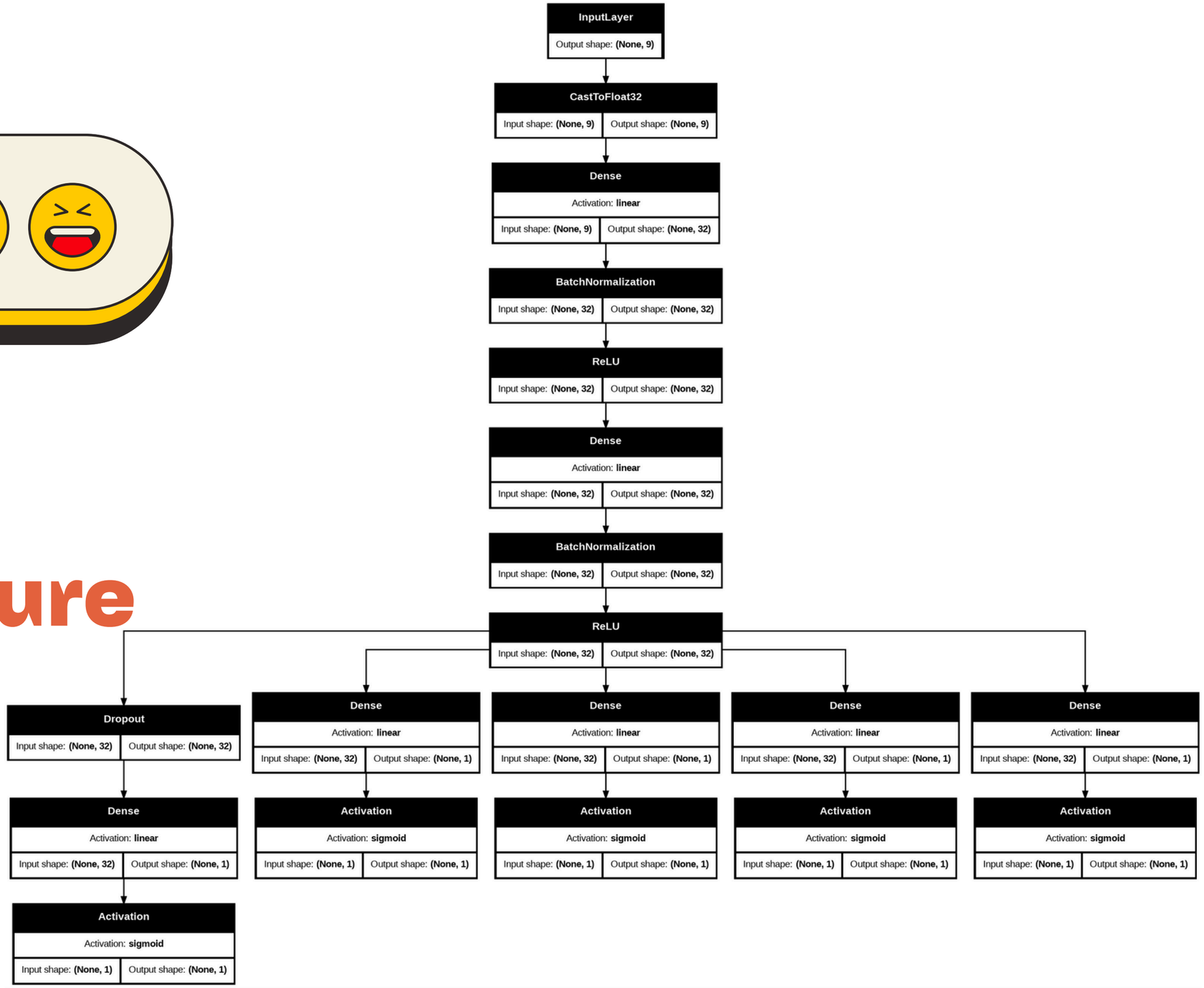
Model Architecture

5

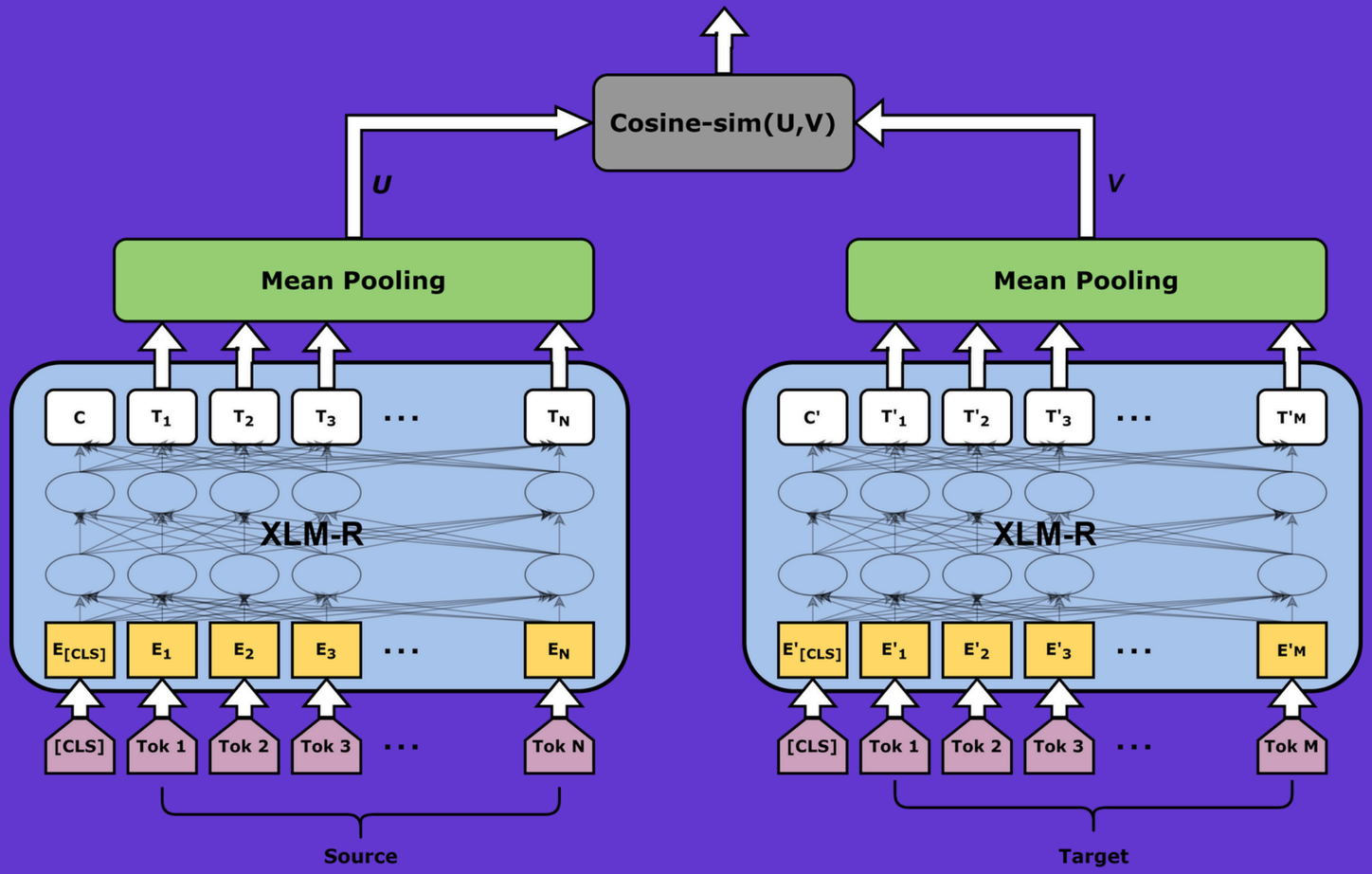
This model on numerical data exrtracted from the user account data and captions features, the model was then built with an automation tool to build the best model and find the best hyperparemeters



Model Architecture



Siamese XLM-R



PANDORA DATASET

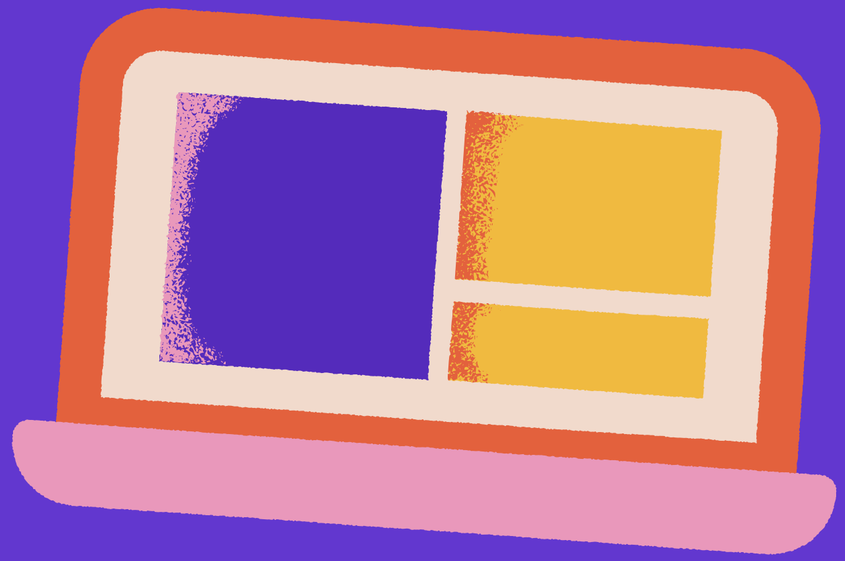
text string	agreeableness float64	openness float64	conscientiousness float64	extraversion float64	neuroticism float64
his name was kim kimble originally wow thats some messed up parents	9	61	13	4	72
theyre better than the normal posts on ryugioh id rather have them then the same topic posted multiple 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

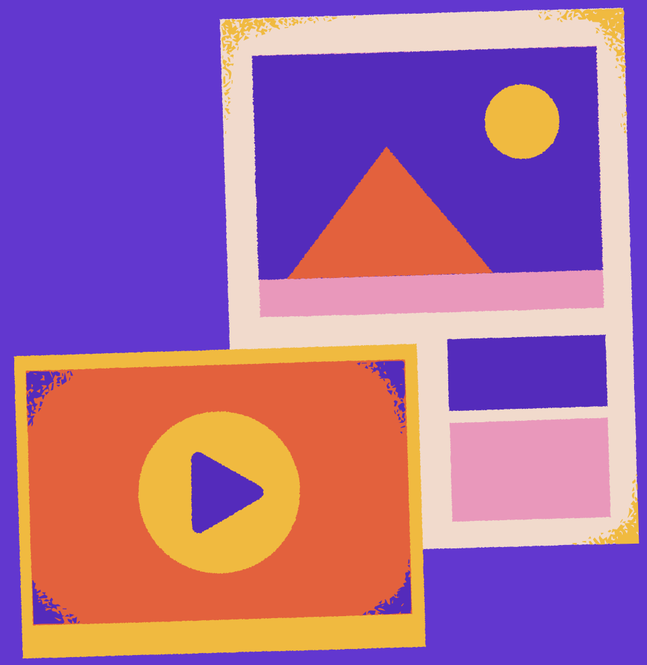
Implementation

We stumbled upon a dataset called PANDORA for OCEAN personlity collected and processed by experts so we decided to use it. We harnessed the power of siamese XLM-R to predict the similarities between the users' captions and the comments found in the PANDORA dataset and based on these similarities we calculate the weighted sum of the personalities to get our final predictions

Models Accuracy

XLM-R Text	**0.582**
ViT Image	**0.582**
XLM-R + ViT	**0.582**
Bio + Numerical Data	**0.72**
Numerical Data	**O:0.7, C:0.73, E:0.67, A:0.78 , N:0.82**
Siamese XLM-R	**O:0.69, C:0.53, E:0.63, A:0.72 , N:0.77**





Deployment

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

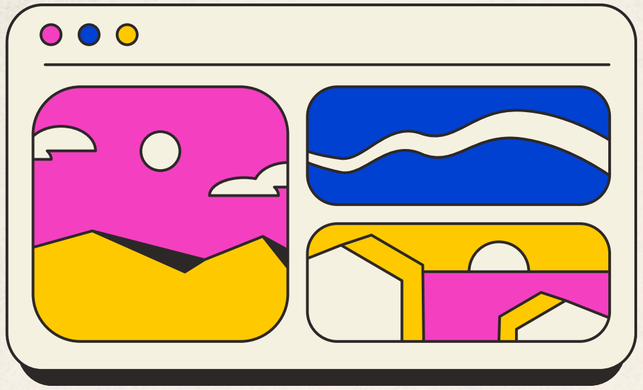
From pre-K to MD—I'm so proud

Little known fact: the corner

STREAMLIT

- THE STREAMLIT LIBRARY WAS THE MAIN TOOL USED FOR DEVELOPING THE WEB APPLICATION.

STEPS



STEP 1

The user enters the instagram profile URL

STEP 2

The application scrapes information about the searched profile

STEP 3

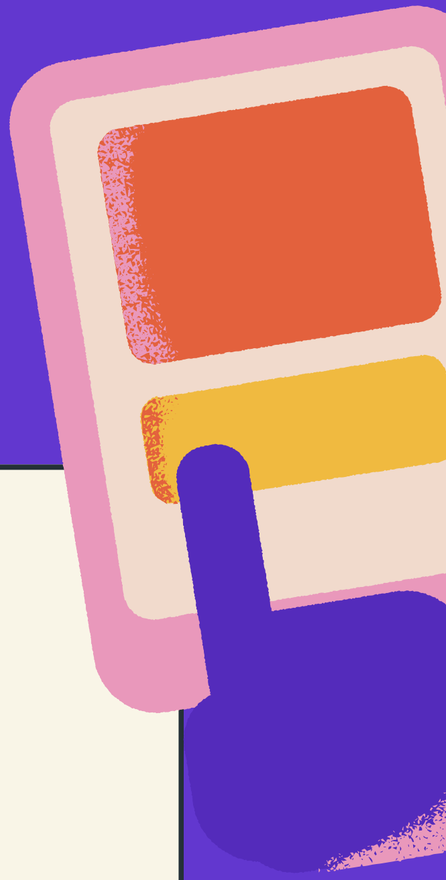
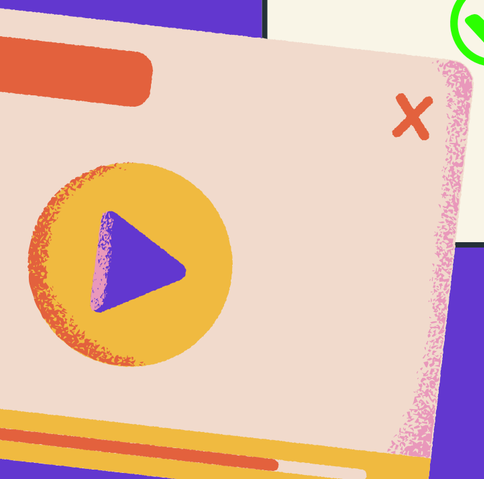
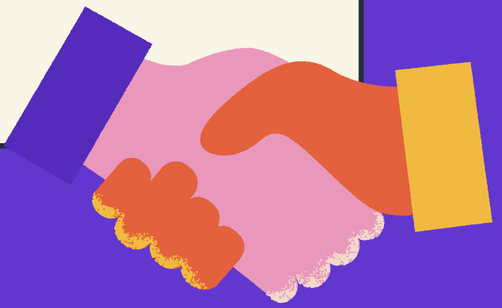
The collected data to the model and predict the right personality of the desired person

STEP 4

The personality of the desired user is then rendered as well as some general information.

Perspectives

- Opportunities for further improvement and exploration
- Deployment and validation in real-world settings
- Collaboration with domain experts for refining the model

Thank You For Your attention!