OPEN-AI SENTIMENT ANALYSIS

From Reddit

POPOS

TABLE OF CONTENTS

APPLICATION SCENARIO

Uncovering Public Opinion

We gain **valuable insights into public opinion** on various topics, because of Reddit's **diverse user base**

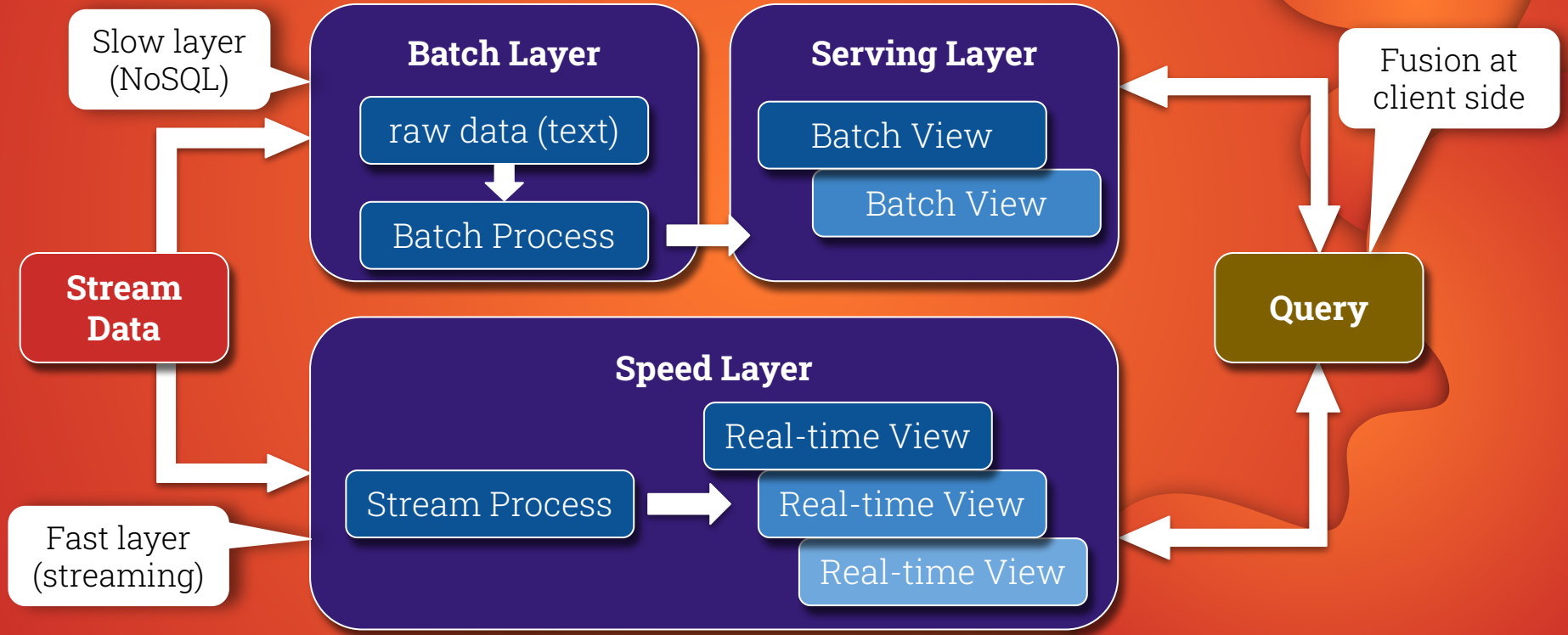
Real-time Analysis

Allow users to **stay updated on the evolving sentiments** surrounding their interests or target subjects

Customization and Scalability

Designed to be **adaptable to various topics** and **scalable** to handle **large volumes of data**

DATA ARCHITECTURE: LAMBDA



ANALYTICAL GOALS

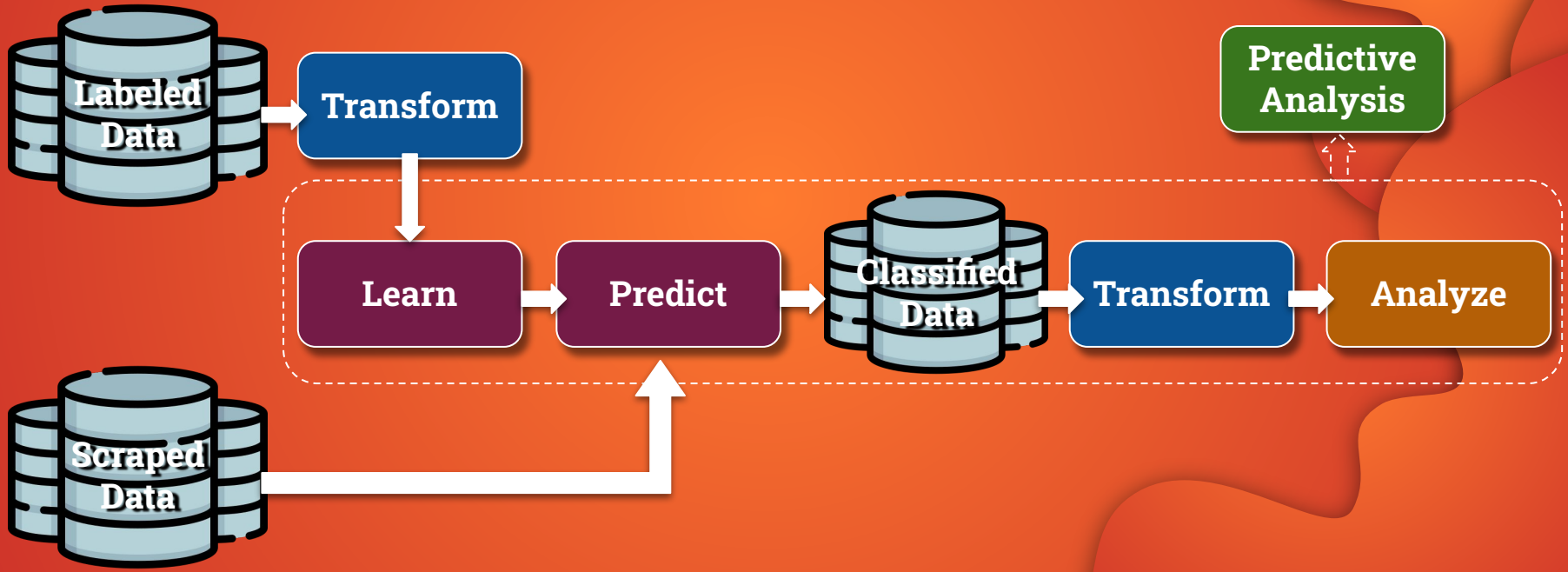
Gain **insights** into the **sentiment** expressed about **OpenAI**

Track **sentiment changes over time** to identify trends, shifts, or events affecting the OpenAI community mood

Identify **key topics** associated with different sentiments within the subreddit

Evaluate **responses** to OpenAI **announcements**, provide insights for **content moderation**, and support **data-driven decision-making** for community management

IMPLEMENTED DATA VALUE CHAIN



MODEL IMPLEMENTATION

Datasets

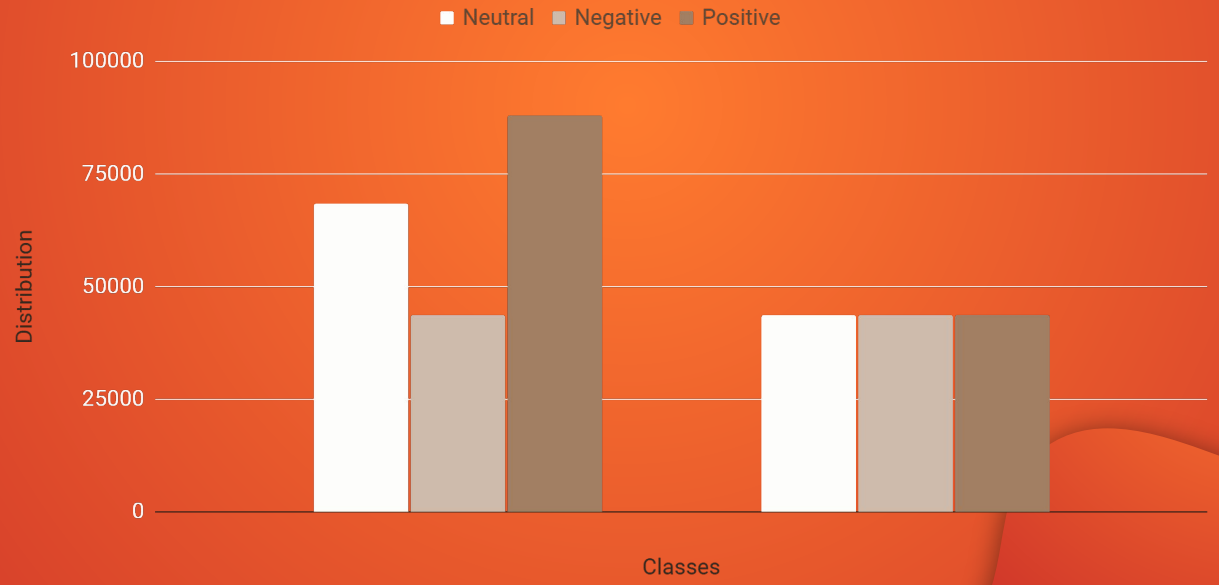
- **Twitter and Reddit comments labeled: Model Training**
  - 131358 samples
  - Positive, Neutral and Negative

- **Scraped Data from Reddit: Analysis**
  - 764 samples
  - Subreddit head
  - Subreddit body
  - Subreddit timestamp
  - Comments body
  - Comments timestamp

MODEL IMPLEMENTATION

Data exploration

Classes distribution in Training Dataset:

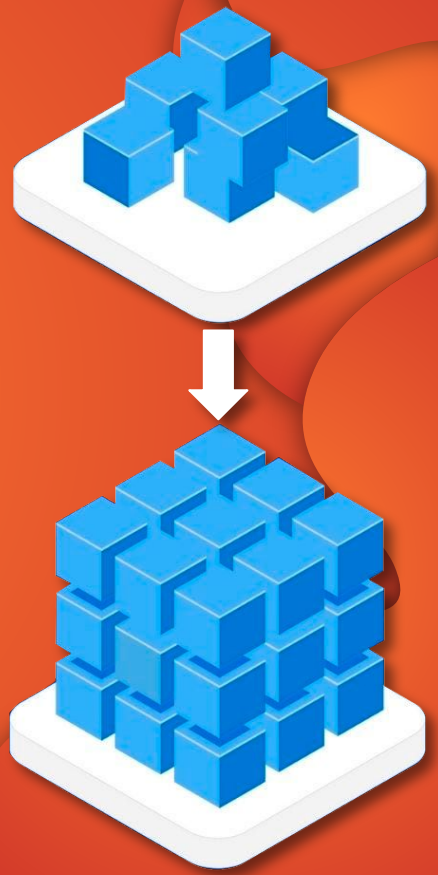


MODEL IMPLEMENTATION

Data transformations

Training Dataset

- **Balance** classes distribution
- **Refactor** labels:
  - (0: Neutral, 1: Positive, -1: Negative) to (0: Neutral, 1: Positive, 2: Negative)
- **Split** Dataset:
  - Train(80%), Validation (10%), Test (10%)
- Transform to Pytorch **Tensor** Dataset
- **Tokenize** Dataset



MODEL IMPLEMENTATION

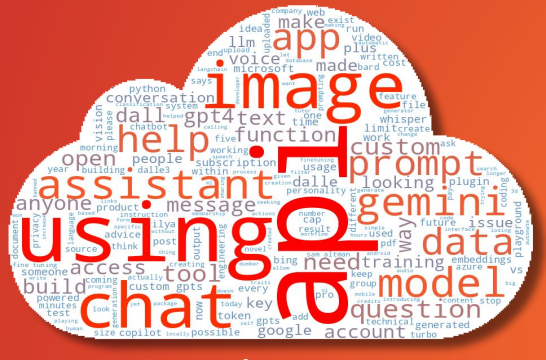
Text Classification Model

- Model Name: **XLM-RoBERTa-base**
- Batch Size: **16**
- Epochs: **2**
- Learning Rate: **2e-5**
- Optimizer: **Adam**
- Dataset: **Twitter and Reddit comments**

RESULTS EVALUATION

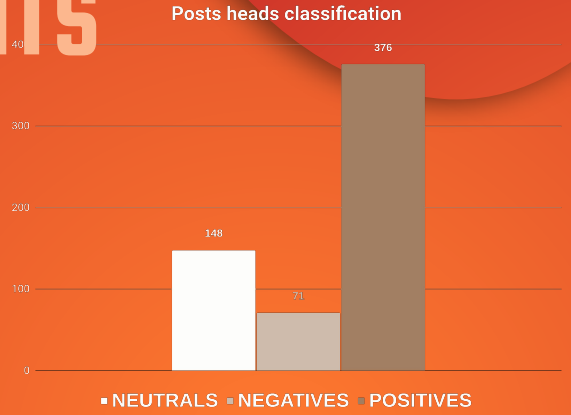
	ACCURACY	LOSS
TRAINING	0.98	0.13
VALIDATION	0.96	0.15
TEST	0.97	0.14

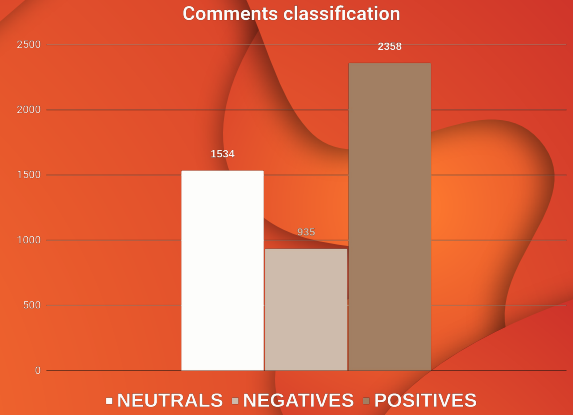
RESULTS INSIGHTS

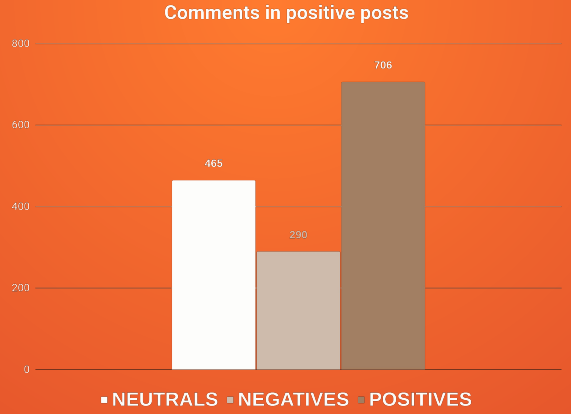


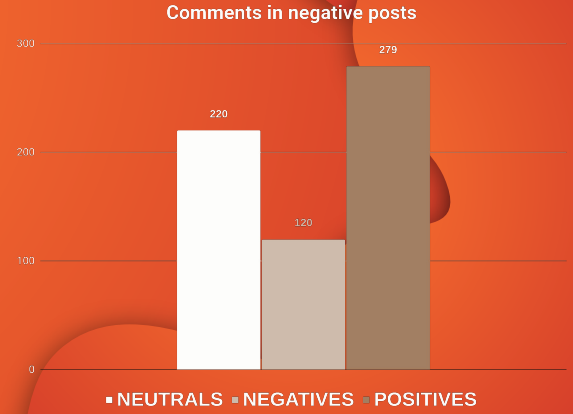
Negative posts

Positive posts









⚠ No correlation between the classification of the post and the comments

USE OF LLM

CONCLUSIONS FUTURE INSIGHTS

TOPICS OF INTEREST
Topics of interest of the community

RESPONSE TO EVENTS
Response of users to specific events

USER ENGAGEMENT
User engagement for an specific brand

REAL -TIME PROCESSING
Check the sentiment in real-time for topics of interest and user engagement

UNIVERSITAT JAUME I

valgrAI

THANK YOU!

Any question?