First Impressions Matter: Predicting Visual Content Popularity through Multimodal Model

Dhawal Shah

Shounak Powar

Rohail Alam

Overall idea

- We are building a system that predicts how much engagement a social-media post will receive (Very Low → Very High).
- Our model looks at both the image and its text description to understand what makes a post engaging.
- We use **OpenAI CLIP-ViT-Base-Patch32**, a multimodal AI model, to extract deep features from each post.

Dataset Utilized - PixelRec

Dataset Overview
Large-scale multimedia dataset of social media posts combining images, metadata, and engagement metrics.

Each sample includes:

- Image
- Text metadata (title, tag, description)
- Raw engagement signals (likes, comments, shares, favorites, views

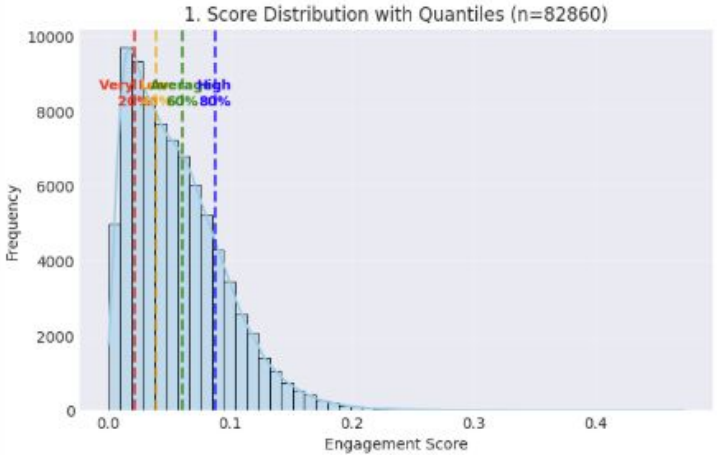
Binning the Dataset and Aggregating Features

Data Processing

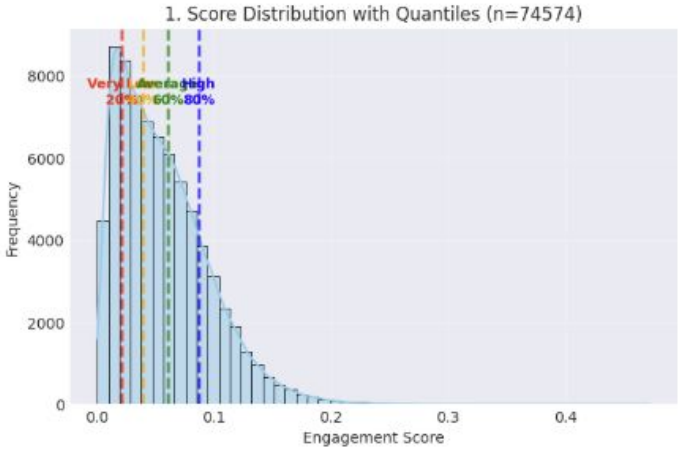
- Computed engagement score for each item.
- Created 5 engagement classes.
- Validated image files and removed missing items.
- Exported final multimodal dataset (image + metadata + score + label).

How Engagement Score Was Calculated

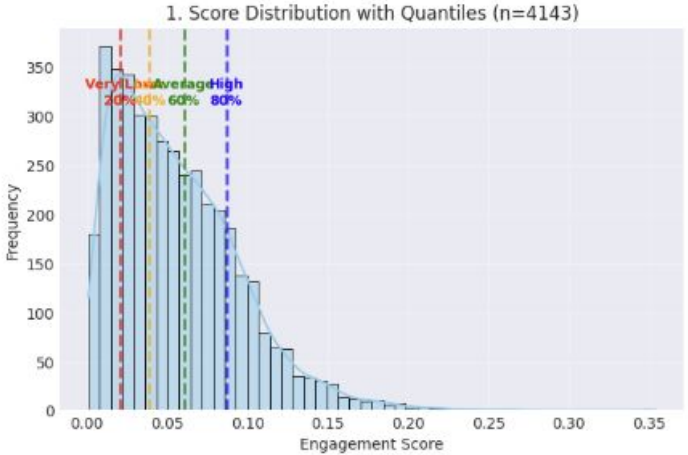
- Used a weighted formula capturing meaningful interactions:
  Engagement Score = (1×likes + 2×comments + 3×shares) / views
- Prevents division by zero using max(views, 1).
- Scores reflect interaction intensity relative to how many people viewed the post.
- Global percentiles (20/40/60/80) used to map items into:
  Very Low, Low, Average, High, Very High.



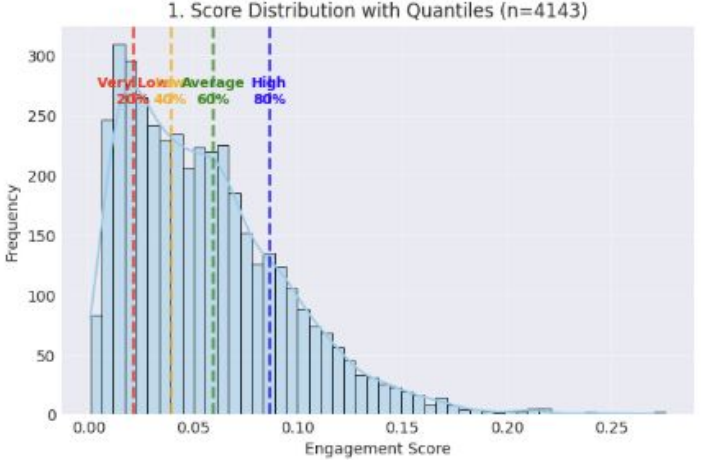
All Dataset Combined



Train Dataset

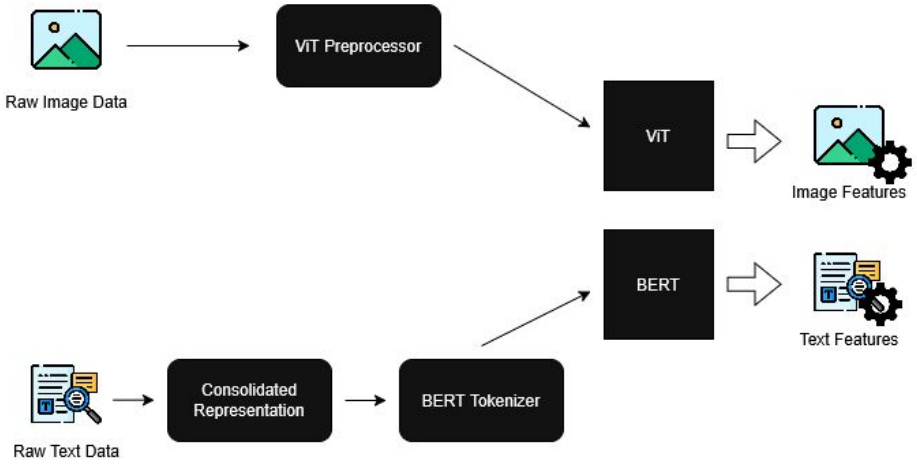


Validation Dataset



Test Dataset

Implementation - Finetuning with ViT + BERT



Models Utilized

ViT : google/vit-base-patch16-224

BERT : bert-base-uncased

Feature Fusion Strategies:

Concatenation, a good baseline
Cross-Attention, inter-modal interactions
Gated, scaled contribution based on importance

Implementation - CatBoost

Image and Text features are extracted using the **CLIP Pytorch Library**, which comes built-in with the required methods, which can be viewed HERE

Model used for feature extraction: **OpenAI CLIP-ViT-Base-Patch32**

Additional Features:

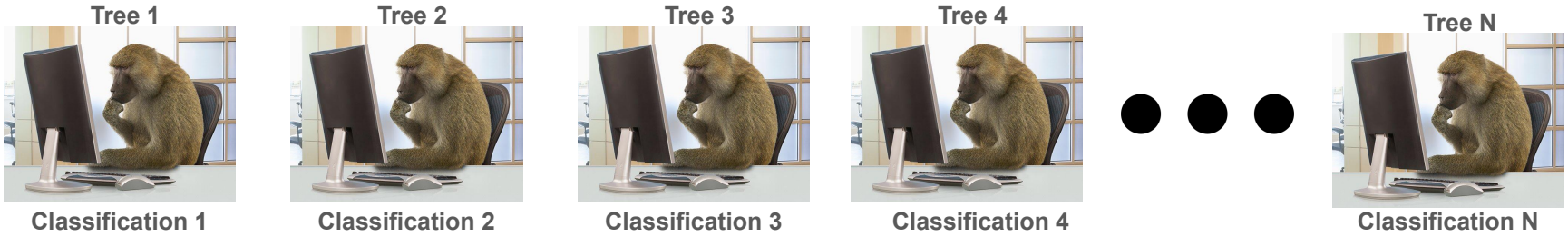
- Cross-modal similarity, the cosine similarity between image and text embeddings.
- “Engagement Rate”, makes engagement prediction more robust

$$engagement\ rate = \frac{likes + 2 \times comments + 3 \times shares}{views}$$

Implementation - CatBoost

Training is done on the CatBoost Classifier - ensemble of Decision Trees

An accumulation of decisions made by thousands of trees, each considering unique combinations of the multimodal features.



A smart ensembled classification

Implementation - Regressive Prediction

As evident in our results for the classification approach, the model is unable to learn to classify correctly.

So an alternative approach is considered - by formulating a new “Popularity Score” metric:

$$\texttt{popularity\_score} = (1 - \text{Cor}(V, L)) \cdot l + (1 - \text{Cor}(V, S)) \cdot s + (1 - \text{Cor}(V, C)) \cdot c + (1 - \text{Cor}(V, F)) \cdot f$$

The multimodal model is trained to predict $log(\texttt{popularity\_score})$

Experimental Setup

- We are using Nova clusters to run our training.
- Iterations per Fold: 50/100/250/500
- LR: 0.01
- Depth of Decision tree: 5/10/14
- Bootstrap_type: Bernoulli
- Loss Function: Multiclass Cross Entropy

Classification Results

Metric	Value
Validation Accuracy	0.3207
Test Accuracy	0.3128
Cross-Validation Accuracy	0.3116 ± 0.0019

Ablation Study

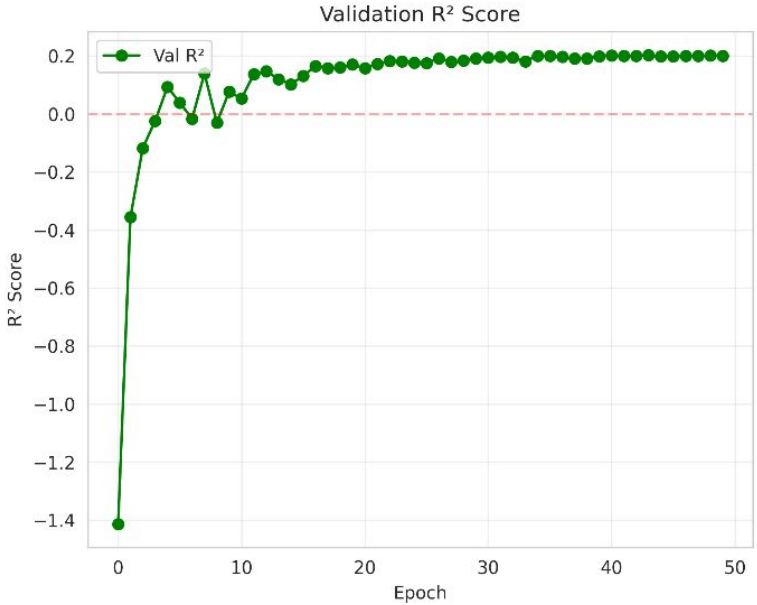
Visual Features play an **important role** in the classification process.

Training F1 score without visual features dips significantly - **~0.28**

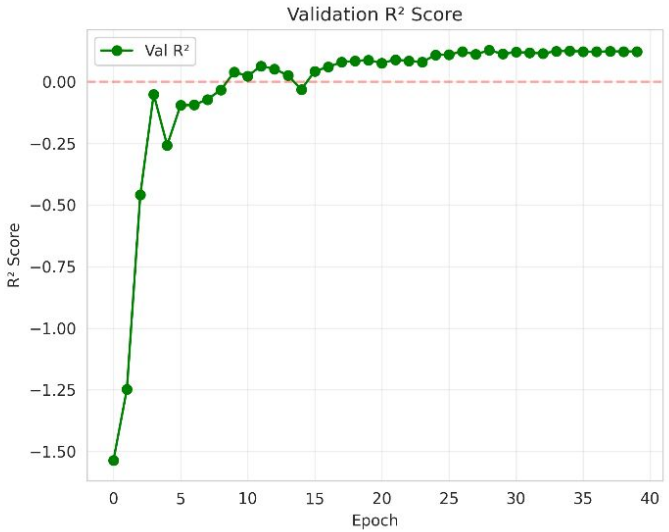
Textual Features are also a contributing factor, although not as significant as Visual

Training F1 score without text features - ~0.30

Regression Results



SigLIP + DeBERTa :
Best R² **0.2033**



CLIP + DeBERTa :
Best R² **0.1284**

```
{
  "image": "../cover/i89258.jpg",
  "engagement": {
    "likes": 2607,
    "comments": 291,
    "shares": 35,
    "views": 177538,
    "favorites": 817
  },
  "title": "Frame-by-frame plagiarism, shoddy production of the worst national comics: Porky Pig, surprisingly copied from Cat and Mouse!",
  "tag": "Miscellaneous",
  "description": "Tom and Jerry.",
  "engagement_score": 0.01855377440322635,
  "engagement_label": "Very Low"
},
```



```
Predicted Label: Very Low
Confidence: 0.2091

Class Probabilities:
  Very Low: 0.2091
  Low: 0.2021
  Average: 0.1987
  Very High: 0.1956
  High: 0.1944
```