

Social Media Prediction — A Comparison Between Gradient-Boosted Trees and Towered Neural Networks Architectures

Jack Vuturo

jackvuturo@u.boisestate.edu

Abstract

This paper investigates popularity prediction on the SMPD-Image dataset from the Social Media Prediction Challenge, comparing a towered multimodal neural network against a CatBoost gradient-boosted tree baseline. Both models consume CLIP image embeddings, CLIP text embeddings, and tabular post metadata, evaluated under 5-fold GroupKFold cross-validation grouped by user ID to prevent leakage. CatBoost achieves a test MAE of 1.6391 ± 0.0446 , outperforming the towered network (1.7148 ± 0.0425) and both mean and median baselines (~ 1.96). The result suggests that despite the inclusion of dense CLIP embeddings, SMPD-Image is fundamentally a tabular prediction problem better suited to tree-based models than to neural architectures. We further argue that the dataset’s exclusion of user-level features, particularly follower counts, imposes a ceiling on what any content-only model can achieve, and we point toward graph-based approaches as a more promising direction for this task.

1 Introduction

The introduction of the internet, and more specifically of social platforms, has changed the world of advertising. Marketing campaigns have always been measurable through conversion metrics: how many customers can I get to perform a desired action? Social media platforms add a new set of metrics relating to reach — likes, comments, and views.

Both large companies and small businesses benefit from building campaigns targeted toward going viral, since increasing reach often translates into more opportunities for conversion. Optimizers for reach existed even before the internet became popular: companies like “AAA” named themselves specifically to land at the top of the phone book. Search Engine Optimization (SEO) was developed to place websites higher in search engine results. Paid advertising can lead to higher rankings, but it

comes at a cost. Social Media Optimization (SMO) instead aims to optimize a post to receive increased organic traffic.

Today, no general social media optimization algorithm exists. With the introduction of recommendation algorithms into ranking, predicting how a post will perform has become increasingly challenging. Phone books were simply alphabetical, and search engines use a combination of semantic analysis and heuristics, but social media algorithms are largely peer-based. They use techniques like cosine similarity to surface content liked by users perceived to be similar to the current user. These algorithms are difficult to reproduce because they depend so heavily on user behavior *after* a post has been published.

2 Background

This paper sits at the intersection of several active research areas: social media popularity prediction, multimodal representation learning, and gradient-boosted trees on tabular data.

2.1 Social Media Prediction

Popularity prediction on social media has been studied for a while, originally focused on text and news content. Szabo and Huberman (2010) were among the first to formalize the problem, showing that early view counts on YouTube and Digg posts could predict long-term popularity reasonably well. Bandari et al. (2012) extended this to news articles on Twitter, using content features and source reputation to classify popularity ranges with around 84% accuracy. These early works established that popularity is at least partially predictable from content and context, even though the actual propagation of a post depends on user behavior that’s hard to model directly.

Image popularity prediction specifically took off with Khosla et al. (2014), who built a regression model on Flickr data using a mix of low-level vi-

sual features (color, gradient, texture, GIST), deep learning features from a CNN, and social cues like follower counts and number of photos uploaded. Their main finding was that image content alone has weak predictive power — most of the signal comes from social context. This result has held up across follow-up work and is part of the motivation for our argument that the SMPD-Image dataset’s exclusion of follower information is a significant limitation. Gelli et al. (2015) built on this by adding visual sentiment features, and ? proposed a multi-modal pipeline combining visual, textual, and social features on the SMP-T1 dataset (a predecessor to SMPD).

More recent work has moved toward end-to-end multimodal deep learning. Abousaleh et al. (2021) proposed VSCNN, a unified network that fuses CNN-based visual features with social features in a single architecture — conceptually similar to the towered NN we evaluate in this paper. The SMP Challenge itself has run annually since 2017, with overview papers from Wu et al. (2019) and ? summarizing the task formulation, dataset construction, and trends in submitted approaches. Submissions have generally moved from hand-crafted feature pipelines toward deep learning and ensemble methods, though gradient-boosted trees remain competitive — a pattern consistent with the broader finding that tree-based models still outperform deep learning on tabular data (Grinsztajn et al., 2022).

Finally, graph-based approaches have started to appear in this space. Cao et al. (2020) proposed coupled graph neural networks for popularity prediction on social platforms, modeling both user-user and user-content relationships. Approaches like this point toward what’s missing from the SMPD-Image setup: the underlying network structure that determines how far a post can actually reach.

2.2 Social Media Prediction Challenge

This paper leans heavily on the work developed in relation to the Social Media Prediction Challenge. The challenge is broken down into two distinct subtasks based on two provided datasets: SMPD-Image and SMPD-Video. SMPD-Image consists of data from Flickr; SMPD-Video consists of data from TikTok.

	SMPD-Image	SMPD-Video
# Posts	486k	6k
# Users	70k	4.5k
# Categories	756	120
Duration (M)	16	24
# Tags	250k	40k

Table 1: Summary statistics for the two SMPD datasets. This paper focuses on SMPD-Image.

2.2.1 Popularity Score Calculation

The models described in this paper predict the popularity score as formulated by the SMP Challenge. The formulation as described below:

$$s = \log_2 \left(\frac{r}{d} \right) + 1 \quad (1)$$

where s is the normalized popularity score, r is the view count of the post, and d is the number of days since the post was published.

This metric naturally differentiates a post with a million likes in a single day versus a post with a million likes gained over five years.

2.3 CLIP for Multimodal Embeddings

CLIP (Contrastive Language-Image Pre-training) is a multimodal model from Radford et al. (2021) trained on 400 million image-text pairs scraped from the internet. It consists of two encoders, a vision transformer for images and a text transformer for captions, trained jointly. This results in image and text embeddings that can be compared directly with tools like cosine similarity.

3 Model and Approach

3.1 Data Preparation

CLIP is utilized for two reasons. First, it gives us pretrained image and text encoders that don’t need fine-tuning to produce meaningful features. We can pass a Flickr photo through the vision encoder and its title/tags through the text encoder and get dense embeddings that already capture semantic content. These are precomputed before training, where we have hundreds of thousands of posts and training image/text encoders from scratch would be expensive. Second, because both modalities share an embedding space, the image and text features are roughly comparable in scale and structure, which makes them easier to fuse in the towered architecture without a lot of per-modality preprocessing. This makes CLIP a natural fit for popularity prediction, where the visual content of a post and its

caption/tags are both contributing signal and need to be combined cleanly.

Precomputed embeddings using the CLIP encoders are cached to disk as NumPy (.npz) files. With roughly 486k posts in SMPD-Image, re-encoding on every epoch or every cross-validation fold would have been prohibitively expensive, so the embeddings live on disk and training runs only pay the I/O cost of loading them.

3.2 Evaluation

We evaluate both models using mean absolute error (MAE) on log-transformed popularity scores, as defined by the SMP Challenge. MAE is reported as the primary metric; we also track MSE and RMSE during development. We use two splits: a simple random 80/10/10 partition into training, validation, and test, and a 5-fold GroupKFold cross-validation (~305k posts per training fold), grouped by user ID so that no user appears in both training and validation within a fold. The grouping matters because individual users have distinct posting patterns and follower bases. Without it, the model can memorize per-user popularity rather than learning content-level signal, which would inflate validation scores in a way that doesn’t generalize well.

3.3 CatBoost Model

CatBoost was trained for 3000 iterations per fold with a learning rate of 0.05. The CLIP image embeddings, CLIP text embeddings, and tabular metadata features were concatenated into a single flat feature vector. Since gradient-boosted trees already handle this kind of mixed dense-and-categorical input natively, so no scaling, normalization, or per-modality preprocessing was needed.

3.4 Towered NN

The towered network was trained with AdamW for 20 epochs per fold, using a learning rate of $3e-4$ and weight decay of $1e-4$. Each modality, image embeddings, text embeddings, and tabular features, is processed by its own tower (a small MLP) before the outputs are concatenated and passed through a shared fusion head that produces the popularity prediction.

Per-fold training curves and validation MAE for the towered network are shown in Figure 2, and the corresponding curves for CatBoost are shown in Figure 1.

4 Results

Model	MAE
Tower-NN	1.7148 ± 0.0425
CatBoost	1.6391 ± 0.0446
Mean (Baseline)	1.9669 ± 0.0538
Median (Baseline)	1.9622 ± 0.0558

Table 2: Test MAE on SMPD-Image, averaged across 5 GroupKFold splits ($\pm$ standard deviation). Lower is better.

Table 2 reports MAE on the held-out test set, averaged across the five GroupKFold splits with standard deviations. Both models clear the mean and median baselines by a comfortable margin (~0.25 MAE), confirming that the CLIP embeddings and tabular metadata carry real signal beyond per-user popularity averages. CatBoost outperforms the towered neural network by 0.0757 MAE (1.6391 vs. 1.7148).

5 Implications

The performance of CatBoost suggests that the SMPD-Image dataset is tabular at its core, even with the CLIP image and text embeddings mixed in. Gradient-boosted trees, as opposed to neural networks, are simply better suited for this kind of feature space. CatBoost’s tree-based architecture handles the mix of categorical metadata, engagement counts, and dense embeddings without caring about scaling or whether half the features are useless. Neural networks have to learn all of that from scratch through gradient descent. The towered structure was supposed to help guide the network into better feature recognition, but the CatBoost architecture still performed better on SMPD-Image.

The SMPD-Image dataset does, however, make some assumptions that are likely false. The dataset provides no user metadata: no follower counts, no historical engagement, and no profile information beyond the user ID. This is problematic, as accounts with large followings are likely to receive more engagement than accounts with smaller followings. If our goal is to accurately predict the engagement of a post, follower information is crucial.

Future work could explore graph convolutional network (GCN) architectures using social graphs, expanding on the research in Cao et al. (2020).

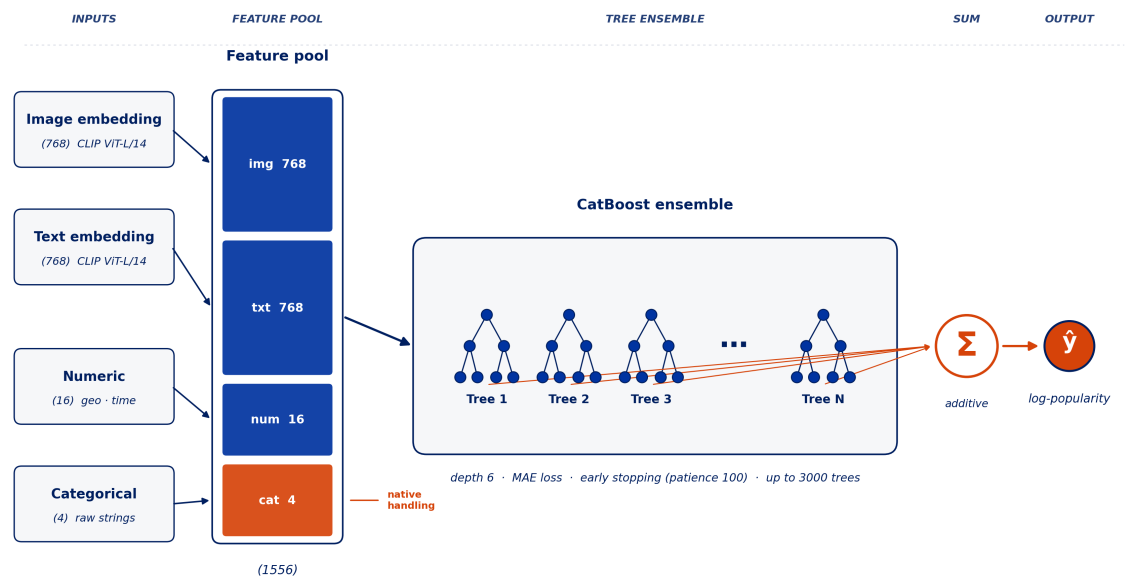


Figure 1: CatBoost model

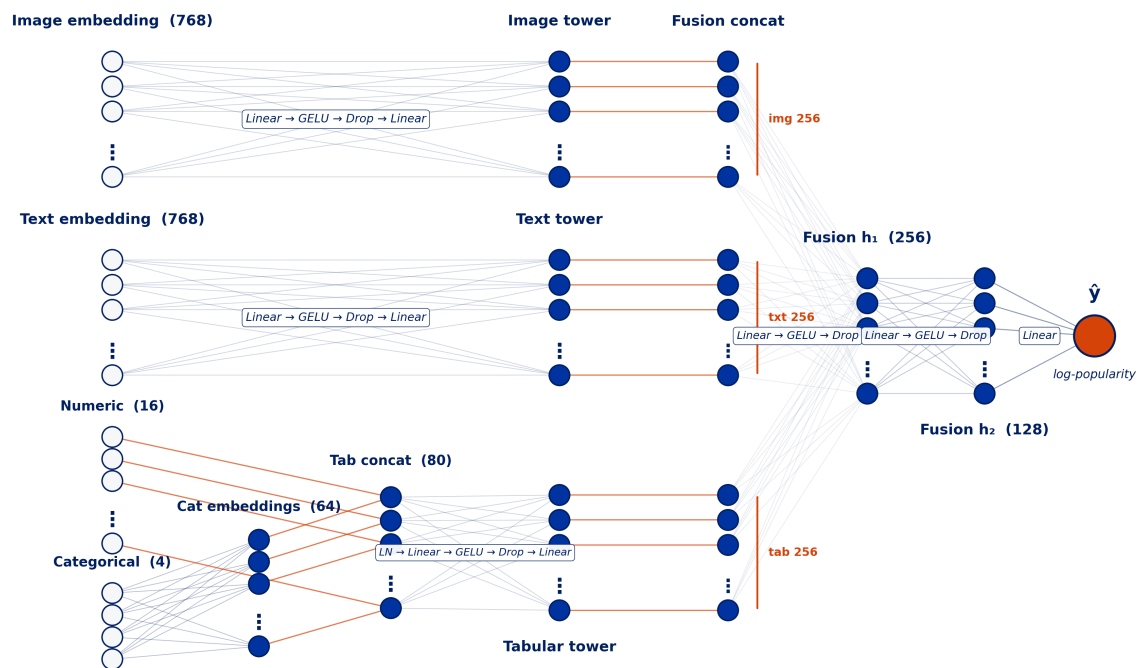


Figure 2: Towered neural network model

6 Conclusion

This paper compared a towered multimodal neural network against a CatBoost baseline on the SMPD-Image popularity prediction task, using CLIP image and text embeddings alongside tabular post metadata. CatBoost achieved a test MAE of 1.6391, outperforming the towered network (1.7148) across all five GroupKFold splits. Both models cleared the mean and median baselines by a comfortable margin, confirming that CLIP embeddings carry real predictive signal. However, the consistent gap in CatBoost’s favor suggests that gradient-boosted trees remain the better fit for this feature space, even when dense embeddings are part of the input.

The more interesting takeaway is what the dataset doesn’t include. Without follower counts or other user-level features, any content-only model is bounded by a ceiling that has little to do with model architecture. Future work should incorporate social graph structure, building on graph-based approaches like [Cao et al. \(2020\)](#), to capture the propagation dynamics that ultimately drive whether a post reaches an audience at all.

References

- Fatma S. Abousaleh, Wen-Huang Cheng, Neng-Hao Yu, and Yu Tsao. 2021. [Multimodal deep learning framework for image popularity prediction on social media](#). *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):679–692.
- Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. [The pulse of news in social media: Forecasting popularity](#). In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM '12)*.
- Qi Cao, Huawei Shen, Jinhua Gao, Bingzheng Wei, and Xueqi Cheng. 2020. [Popularity prediction on social platforms with coupled graph neural networks](#). In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*, pages 70–78. ACM.
- Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. [Image popularity prediction in social media using sentiment and context features](#). In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*, pages 907–910. ACM.
- Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS '22) Datasets and Benchmarks Track*.
- Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. [What makes an image popular?](#) In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*, pages 867–876. ACM.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML '21)*, pages 8748–8763.
- Gabor Szabo and Bernardo A. Huberman. 2010. [Predicting the popularity of online content](#). *Communications of the ACM*, 53(8):80–88.
- Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. [SMP challenge: An overview of social media prediction challenge 2019](#). In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*, pages 2667–2671. ACM.