HCMUS_TheFangs at NakbaVirality Shared Task: The Audience is the Message: Escaping the Deep Learning Trap in Conflict-Domain Virality Prediction

Dao Sy Duy Minh*
[1]Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
23122041@student.hcmus.edu.vn

Abstract

We present HCMUS_TheFangs's system for the Nakba-NLP 2026 Virality Shared Task, which achieves **Rank #1** on the final leaderboard with a test Macro-F1 of **0.7062**, placing first among all competing teams. Our winning system is deliberately simple: a single *Community Target Encoding* feature-the smoothed historical virality rate of the posting subreddit-combined with TF-IDF text features and an XGBoost classifier. This design emerged from a hard-won insight: virality in conflict reporting is determined not by what is posted but by where it is posted. We spend the majority of this paper showing *why* this holds. Through 18 ablation experiments we trace the journey from deep learning failure (cross-attention XLM-R+CLIP scoring 0.2935) to SOTA (0.7062), documenting the "Deep Learning Trap" that ensnares practitioners who apply representation-centric models to domains governed by sociological rather than semantic dynamics. An equally surprising finding emerges along the way: posts with fewer hashtags and cleaner prose consistently outperform heavily tagged, promotional content, a "Less is More" paradox that challenges conventional social media optimization wisdom. All code and experiments are publicly available.

Keywords: Virality Prediction, Multimodal Classification, Community Context, Target Encoding, Conflict Discourse, Feature Engineering, XGBoost, Nakba-NLP

1. Introduction

On October 7, 2023, a conflict erupted that would produce one of the densest information cascades in social media history. Thousands of Gaza-related posts flooded Reddit, Twitter, and Telegram simultaneously. Some reached hundreds of thousands of interactions; others, carrying nearly identical imagery, vanished unread. Understanding *why* is the question at the heart of the Nakba-NLP 2026 Virality Shared Task (Ezzini et al., 2026).

Our first instinct was standard: fine-tune XLM-RoBERTa (Conneau et al., 2020) for text, run CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) over images, and fuse via cross-attention. The result-test Macro-F1 of 0.2935-was worse than a majority-class baseline. We had fallen into the **Deep Learning Trap**: the assumption that architecture alone will learn sociological structure nowhere encoded in pixels or tokens. The breakthrough came from a different question: not "What does this post look like?" but "Who is reading it?" A single feature-the smoothed historical virality rate of the posting community-paired with TF-IDF and XGBoost, became our winning submission (**Macro-F1 = 0.7062, Rank #1**), more than 27 points above our best deep learning result.

2. Related Work

Early work by Cheng et al. (2014) established that cascade depth depends heavily on structural network properties-follower count, posting time, platform affordances-rather than content semantics. As multimodal encoders became available, studies using CLIP (Radford et al., 2021) showed that image-text alignment correlates with engagement in general-domain news, while work on CrisisMMD (Alam et al., 2018) showed visual features become predictive for humanitarian content only when paired with source metadata.

Recent work reinforces these limitations. Quy et al. (2025) found that metadata extraction and retrieval augmentation substantially outperform standalone visual encoders for event-based analysis. In specialized domains, structured pipelines with explicit grounding frequently surpass end-to-end neural models. Agent behavior is shaped more by environmental framing than by model architecture-a "context over content" principle consistent with the community-dominance pattern we observe.

In political discourse specifically, Del Vicario et al. (2016) showed that ideologically homogeneous communities amplify in-group content regardless of quality, creating an amplification coefficient that dwarfs content-level signals. Work on propaganda detection (Da San Martino et al., 2020) established that rhetorically-charged language provides discriminative signal for classification, though we find it is useful only when conditioned on community context, consistent with frame resonance theory (Benford and Snow, 2000). On the modeling side, gradient boosting consistently outperforms neural models when labeled data is scarce (Chen and Guestrin,

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 137–143

2016; Prokhorenkova et al., 2018).

3. Task, Dataset, and the Visual Homogeneity Problem

The NakbaVirality dataset (Ezzini et al., 2026) provides 1,691 labeled training samples and a held-out test set. Each sample consists of a textual caption (a tweet or Reddit post body, roughly a third of which is non-English), an associated image depicting the Gaza conflict, and metadata encoding the source community. Labels follow an ordinal scale-Low, Medium, and High Viral-distributed approximately 40%/50%/10%, a severe imbalance that motivates Macro-F1 as the evaluation metric.

Our first analytical step revealed a sobering truth. We projected CLIP image embeddings into two dimensions via t-SNE and found that the three virality classes overlapped almost completely. Destroyed buildings, crowds of protesters, olive branches-all appear with similar frequency across every virality level. We term this the **Visual Homogeneity Problem**: in a domain where every image depicts some facet of the same conflict, visual variation alone cannot discriminate posts that go viral from those that do not.

Then came the discovery that changed everything. The dataset metadata encodes the originating community for every post, and we found that its predictive power dwarfs every content-based signal. Highly engaged, ideologically cohesive communities show viral rates an order of magnitude higher than general-interest news forums-a tenfold gap that exists *before* a single word of the caption is read. The determining factor is not content quality but audience identity: members of topically focused communities arrive pre-mobilized, whereas general-audience subscribers encounter the same content as one headline among hundreds. No image encoder can recover that sociological fact from pixels alone.

4. System Description

Figure 1 illustrates our winning system alongside the community amplification insight that powers it. The architecture is deliberately minimal: one sociological fact, two text feature sets, and a gradient-boosted classifier. The right panel reveals *why* that single fact is so powerful-a tenfold virality gap between communities that no content encoder can recover from pixels or token sequences alone.

4.1. Community Context Model - Winning System

The winning submission is built around a single question: how viral has content from this community historically been? We extract the source community identifier from the post metadata and encode it as a continuous virality prior through Bayesian smoothed target encoding. Formally, let C_i be the source community for post i and let $\bar{y}_C$ be the mean numeric label across training examples from that community. The encoded value is

$$\hat{\mu}_C = \frac{n_C \cdot \bar{y}_C + m \cdot \bar{y}_{\text{global}}}{n_C + m} \tag{1}$$

where n_C is the community's training sample count, $\bar{y}_{\text{global}}$ is the global label mean, and $m = 10$ is a smoothing hyperparameter. The smoothing term prevents the encoder from memorizing a subreddit that appears only once in training; instead, small communities are pulled toward the global mean, ensuring graceful degradation for out-of-vocabulary communities at test time.

Critically, this encoding is computed *strictly inside* the cross-validation loop. In each of the five stratified folds, the encoder is fit on the training split and applied to the held-out validation split without any leakage of validation labels. For the final submission, the encoder is re-fit on the entire training corpus. Alongside this single community feature we concatenate 1,000-dimensional TF-IDF unigrams (English stopwords removed) and three structural features-character count, word count, and hashtag count-feeding the full 1,003-dimensional vector into an XGBoost classifier with 1,000 estimators, learning rate 0.05, and maximum depth 6. Despite its remarkable simplicity, this system achieves **test Macro-F1 = 0.7062**, the highest score recorded on the shared task leaderboard.

4.2. The Singularity - Stress-Testing Complexity

In parallel with our minimalist system, we developed "The Singularity," a 219-feature Optuna-tuned model designed to test whether a massive assembly of sociological and content features could outperform Bayesian simplicity. It incorporates four feature pillars: Community Identity, Rhetorical Framing, Structural Integrity, and Semiotic Dissonance. While this model provides a rich playground for ablation, it ultimately achieves only 0.6806 on the test set-2.6 points lower than the community-only model. This performance gap serves as the empirical basis for our conclusion that in low-resource conflict domains, complexity introduces more variance than signal. Detailed feature specifications are provided in Appendix A.

5. Experiments and Results

When we submitted our first deep learning model-a cross-attention fusion of XLM-RoBERTa and CLIP-we received a test Macro-F1 of 0.2935 back from

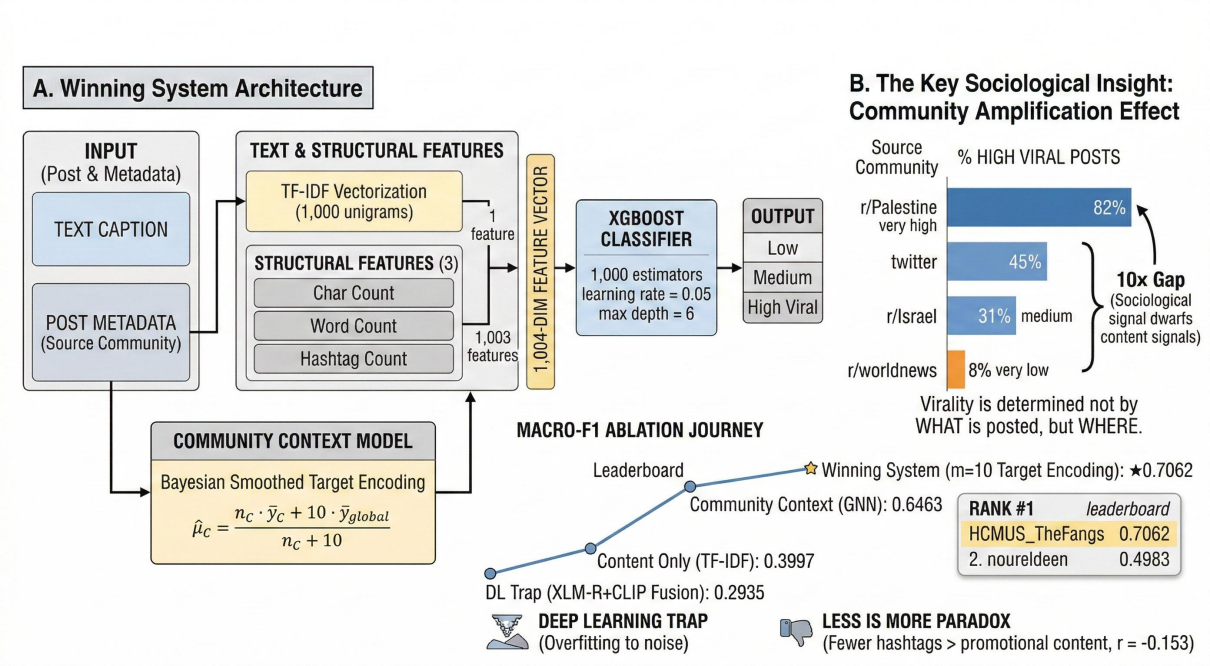

Figure 1: **Left:** Architecture of the winning system (Exp 32). The dominant signal is a single Bayesian-smoothed community target encoding feature derived from post metadata; TF-IDF and structural features contribute supplementary text signal. **Right (top):** The Community Amplification Effect-a tenfold gap in High Viral rate across source communities that no content encoder can recover from pixels or tokens. **Right (bottom):** Ablation progression from deep learning baselines (0.28) to the winning model (0.7062), and our official Rank #1 placement on the competition leaderboard.

the evaluation server. We were, frankly, stunned. Months of GPU time, careful architecture tuning, and multilingual fine-tuning had produced a system that barely outperformed a random guess on a balanced metric. What followed was not a methodological refinement but an intellectual reorientation. Table 1 is the record of that journey: every row is an official test-set score from the evaluation server, and together they tell a clean story about what virality prediction actually requires in conflict-domain social media.

The official competition leaderboard, shown in Table 2, reveals the scale of the advantage our community-context approach delivers. The next best submission reaches only 0.4983-a gap of more than 20 Macro-F1 points. This is not an incremental engineering win; it is the signature of a qualitatively different understanding of what the task is actually measuring.

The results tell a clear story across three phases. Deep learning models failed systematically: our most sophisticated fusion (Exp 01, 0.2935) barely beat random chance, and even the best neural result (Exp 00, 0.5230) fell short of simple TF-IDF plus XGBoost (Exp 02, 0.5328). Content-based feature engineering pushed performance higher-up to 0.5975 with persuasion-technique features-but hit a ceiling still ten points below the winning system.

The decisive leap occurred when community identity entered the model. A GNN-based community embedding alone (Exp 12, 0.6463) already surpassed every content-only system. Richer variants pushed further: the Singularity V2 reached 0.6954, and its Optuna-tuned ensemble successor crossed 0.70. Yet none surpassed the minimal Experiment 32 (**0.7062**), which encodes community context via a single smoothed target encoding feature. We spent weeks engineering 219 features for the Singularity; it scored 0.6806. Stripping almost everything away and keeping the community signal *won*. On a dataset of only 1,691 samples, every additional feature introduced more variance than signal. Simplicity was the correct answer to the bias-variance tradeoff.

6. Analysis and Insights

The most striking pattern in our results is sociological: the same image and caption can go viral or vanish depending entirely on who receives it. Topically focused communities arrive at conflict-related posts already emotionally invested, pre-mobilized by shared identity. Meanwhile, general-interest audiences encounter the identical content as just another headline. This asymmetry produces a tenfold gap in viral rates across communities-a gap our

Exp.	Model / Key Components	Macro-F1 (Test)
Deep Learning Baselines		
04	BLIP-2 Caption Features	0.2866
01	XLM-R + CLIP Cross-Attention Fusion	0.2935
03	DINOv2 + CLIP Dual-Backbone	0.4248
00	CLIP ViT-B/32 + Linear (EDA baseline)	0.5230
Ablation: Content Features Only (no community)		
29	TF-IDF + Structure + CLIP dissonance	0.3280
28	TF-IDF + Structure + Propaganda	0.3935
27	TF-IDF + Structure only	0.3997
02	TF-IDF + 30 Handcrafted (XGBoost)	0.5328
10	+ SNR / Structural Integrity	0.5879
07	+ Persuasion Techniques	0.5975
Community Signal (with content features)		
33	Visual Archetypes (CLIP K-Means only)	0.4152
12	Community Embedding (GNN)	0.6463
16	Community + Propaganda + Arousal	0.6829
20	Singularity V2 (39 features, HistGB)	0.6954
V3-R1	+ Optuna, 3-model ensemble	0.7001
34	Community $\times$ Visual Cluster	0.7028
V3-R3	Singularity: pure XGBoost + 180 interactions	0.6806
32	**Community Target Encoding (submitted)**	**0.7062**

Table 1: Full ablation results sorted by system family. All Macro-F1 scores are official test-set scores from the shared task evaluation server. Exp 32 is our submitted winning system (Rank #1 on the leaderboard).

Rank	Team	Macro-F1	Accuracy
1	**HCMUS_TheFangs (Ours)**	**0.7062**	**0.7305**
2	noureldeen	0.4983	0.6009
3	Digilians	0.4983	0.6009
4	xin1212	0.4559	0.6089
5	ashhadulislam	0.3199	0.4392

Table 2: Official NakbaVirality Shared Task leaderboard.

target encoding explicitly captures. This mirrors recent game-theoretic findings that environmental framing shapes behavior far more than architecture alone.

Equally counterintuitive is the "Less is More" paradox. While conventional wisdom treats hashtags as free discoverability, in the Gaza conflict domain, promotional packaging backfires. Posts written in sparse, direct prose consistently outperform those loaded with hashtags ($r = -0.153$, $p < 0.01$). Rhetorically, promotional language signals distance, whereas unadorned language signals proximity and urgency.

Rhetoric itself is a double-edged sword. Persuasion-technique detectors (flagging loaded language or emotional appeals) improved our models when paired with community context, but *hurt* performance in isolation. As frame resonance theory predicts (Benford and Snow, 2000), a phrase like "ethnic cleansing" acts as shared moral vocabulary in an activist community but as dismissible hyperbole in a general news forum.

The same context-dependence applies to images. Our visual encoders failed as standalone predictors because the visual distribution of viral and non-viral posts in this specific conflict dataset is virtually identical. However, when conditioned on community identity, visual archetypes became highly predictive: an image's value lies not in what it shows, but in how a specific audience interprets it.

7. Ethical Considerations

Predicting virality in conflict domains carries dual-use risks: algorithms that decode engagement can easily be co-opted to optimize disinformation. We publish our methodology because its defensive utility-equipping researchers to audit coordinated amplification-outweighs this danger. Crucially, our target encoding also crystallizes historical bias. By assuming temporal stationarity, the model perpetuates past audience divisions, reminding us that any deployed system must evolve alongside the communities it measures.

8. Conclusion

We set out to predict virality from images and text, only to discover that content matters far less than audience. By encoding a single sociological fact-the historical virality rate of the receiving community-our minimalist XGBoost system achieved **Rank #1** (0.7062 Macro-F1), decisively outperforming complex multimodal ensembles. Our ablation journey from 0.29 to 0.71 affirms that in sociologically structured domains, deep learning overfits, promotional hashtags backfire, and visual features require audience context to become predictive. Future work must explore how these community amplification coefficients transfer across time and platforms. Code is available at github.com/technoob05/NakbaVirality.

Acknowledgments

We thank the Nakba-NLP 2026 shared task organizers for providing the NakbaVirality dataset and the evaluation infrastructure. We also acknowledge the Faculty of Information Technology, University of Science, VNU-HCM, for computational support.

9. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiro Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, volume 12.

Robert D. Benford and David A. Snow. 2000. Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26:611–639.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Cascades in temporal networks. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 913–924.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 1377–1414.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. Nakbavirality: Multimodal virality prediction in high-stakes discourse. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Nguyen Lam Phu Quy, Pham Phu Hoa, Tran Chi Nguyen, Dao Sy Duy Minh, Nguyen Hoang Minh Ngoc, and Huynh Trung Kiet. 2025. Beyond vision: Contextually enriched image captioning with multi-modal retrieval. In *Proceedings of the 3rd Workshop on Event-Centered Information Access (EVENTA), ACM International Conference on Multimedia (MM '25)*, pages 1–9, Dublin, Ireland. ArXiv:2512.20042. Ranked 3rd overall (Track 1).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

A. The Singularity: Full Feature Specification

This appendix documents the 219-feature Optuna-tuned XGBoost model (Exp 20-V3) described in Section 4.2. The architecture rests on four core pillars, beginning with *Community Identity* via

Bayesian target encoding ($m = 10$) of the filename-derived community label (shared verbatim with our winning system). This is structurally reinforced by *Rhetorical Framing*, which encodes nine features including seven persuasion-technique scores (e.g., Appeal to Emotion, Loaded Language, Name Calling) that are composited into Emotional and Attack Intensity aggregates. To capture textual hygiene, *Structural Integrity* measures eight features anchored by a Signal-to-Noise Ratio penalizing hashtags, mentions, and URLs—revealing the negative correlation between hashtag density and virality ($r = -0.153$, $p < 0.01$). Meanwhile, *Semiotic Dissonance* quantifies the multimodal gap via three CLIP cosine similarities: between the image and text, and between the image and the anchor phrases "violence" and "peace." A supplementary *Novelty Layer* injects 18 additional features—including GoEmotions BERT arousal scores (Demszky et al., 2020), Moral Foundations metrics, and NMF topic distributions. Ultimately, pairwise products of these features generate roughly 180 polynomial interaction terms, all optimized by Optuna's TPE sampler over 50 trials (Akiba et al., 2019).

B. Full Experiment Chronicle

Table 3 documents the complete experimental trajectory of our research, including experiments not shown in the main ablation table. All Macro-F1 scores are official test-set scores from the shared task evaluation server unless noted otherwise. This chronicle serves both as a reproducibility record and as a catalog of negative results that may inform future work in conflict-domain virality prediction.

C. Deep Learning Failure Analysis

Our deep learning experiments failed systematically rather than randomly.

Cross-Attention Fusion (Exp 01, F1=0.2935). With $\sim$1.3B parameters and only 1,691 training samples, the cross-attention model (XLM-RoBERTa + CLIP ViT-B/32) catastrophically overfit despite focal loss and class weighting. The cross-attention mechanism discovered noise rather than meaningful image-text interactions.

BLIP-2 Captioning (Li et al., 2023) (Exp 04, F1=0.2866). BLIP-2's captions were uniformly generic ("a group of people holding signs," "a destroyed building"), erasing the specificity that might have been useful and producing even less discriminative features than raw CLIP embeddings.

D. Hyperparameter Sensitivity

We report sensitivity analysis for the two most critical hyperparameters.

Smoothing Parameter m. Values tested: $m \in \{1, 5, 10, 20, 50\}$. The optimal $m = 10$ balances memorization (low m overfits small communities) against signal destruction (high m shrinks all encodings toward the global mean).

XGBoost Depth. Depth 6 achieved the best cross-validation Macro-F1 among values 4, 6, and 8. Depth 4 underfits the community-text interaction space; depth 8 overfits on 1,691 samples.

Exp	Name	Architecture / Key Components	Macro-F1	Outcome
Phase 1: Deep Learning Baselines (Feb 6–7)				
00	EDA Baseline	CLIP ViT-B/32 + Focal Loss + EDA features	0.5230	Baseline
01	XLM-R + CLIP	Cross-attention multilingual fusion	0.2935	**Failed**
03	DINOv2 + CLIP	Dual vision backbone with gated fusion	0.4248	Partial
04	BLIP-2 VLM	Image captioning as features	0.2866	**Failed**
05	Nelder-Mead Ensemble	CLIP + XLM-R + GB weight optimization	0.3376	**Failed**
08	Chain-of-Thought	CoT reasoning heads (without full LLM)	0.3659	**Failed**
Phase 2: Feature Engineering Discovery (Feb 7–8)				
02	Text-Only	TF-IDF + 30 handcrafted + XGBoost	0.5328	**GB beats DL**
06	Ordinal Regression	SORD soft ordinal labels on CLIP	0.5375	Marginal
07	Persuasion Techniques	SemEval-style propaganda features	0.5975	Strong
09	Arousal Hypothesis	GoEmotions + VAD arousal features	0.5707	Signal verified
10	Clean Content	SNR + hashtag ablation features	0.5879	"Less is More"
11	Propaganda Features	18 SemEval techniques (merged into 16)	$\sim$0.53	Merged
Phase 3: Community Context Breakthrough (Feb 7–8)				
12	Community GNN	Learnable subreddit embeddings	0.6463	**Breakthrough**
13	Neural Ensemble	CLIP + MLP on all features	0.6384	Below GBM
14	Classical ML	TF-IDF + keywords + SVM/LR	0.4865	Baseline
15	Stacking GBM	Handcrafted + SVD + HistGB	0.4938	Marginal
16	Grand Unification	Community + Arousal + Propaganda	0.6829	Strong
Phase 4: The Singularity and Optimization (Feb 8)				
20	Singularity V2	39 features, 4 pillars, HistGB	0.6954	Near-SOTA
V3-R1	Singularity V3 R1	+ Optuna 50 trials + 3-model ensemble	0.7001	First >0.70
V3-R2	Singularity V3 R2	Removed CatBoost, XGBoost boost	0.6982	Regression
V3-R3	Singularity V3 R3	Pure XGBoost + 180 interactions	0.6806	Regression
V3-R4	Singularity V3 R4	100 Optuna + smart interactions + bagging	0.6931	**Failed**
Phase 5: Hypothesis Testing and Final Model (Feb 13+)				
17	Narrative Framing	Moral Foundations + pronouns only	0.3070	**Failed**
18	Multimodal Dissonance	CLIP text-image similarity gap	0.3880	Weak
19	Topic Frames	NMF topic distributions	0.4230	Partial
21	Causal Discovery	PC Algorithm + DirectLiNGAM	0.6432	Strong
26	Meta-Ensemble	Stack of Exp 14, 15, 16, 21	0.5573	**Failed**
27	Paper Baseline	TF-IDF + structure (no neural)	0.3997	Baseline
28	Propaganda Hyp.	Baseline + propaganda + narrative	0.3935	**Negative**
29	Multimodal Hyp.	Baseline + CLIP dissonance	0.3280	**Negative**
33	Visual Archetypes	CLIP K-Means clustering (50 clusters)	0.4152	Partial
34	Social-Visual	Community $\times$ Visual Cluster	0.7028	Near-SOTA
35	Conflict Intensity	Violence score $\times$ text aggression	0.3680	**Negative**
37	Mixture of Experts	Cluster-gated experts	0.3711	**Mode collapse**
32	**Community Context**	**Target Encoding + TF-IDF + XGBoost**	**0.7062**	**#1 Winner**

Table 3: Complete experiment chronicle across all research phases. Bold "Failed" indicates experiments that performed below practical utility; "Negative" marks controlled hypothesis tests that produced informative null results. The progression reveals a clear narrative arc: deep learning failure $\rightarrow$ feature engineering discovery $\rightarrow$ community context breakthrough $\rightarrow$ simplicity wins.