Case Study Competition 2026

National Social Summit

Team : SSC_STRATEGY

Sparsh Parashar | Samriddhi | Pinjal Raj | Eeshaa Goel | Deepanshu Ranjan

Understanding the Digital Hate Ecosystem

Hate Speech

Hate speech is expression that attacks, disparages, or **promotes hostility** against a person or group **because of identity traits** like race, religion, gender.



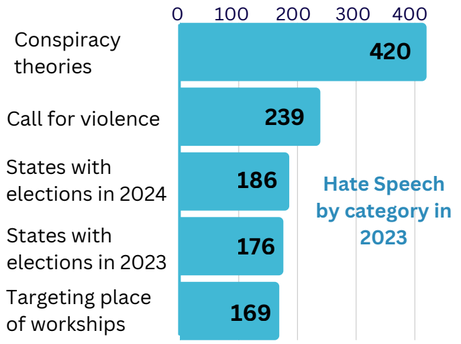
Misinformation

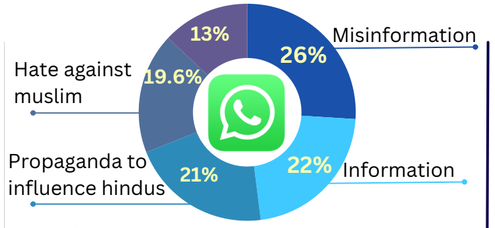
Misinformation is false, inaccurate, or **misleading information** that is shared without the clear intention to deceive the people who receive it.



The overlap exists because falsehoods are often the **"fuel" for hatred**

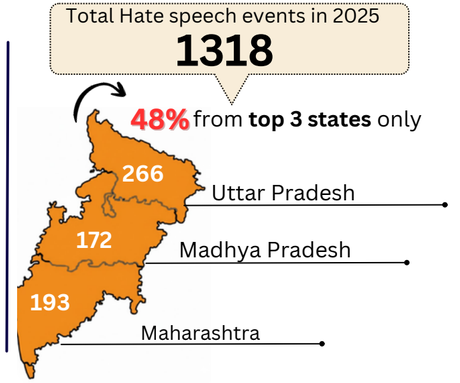
Data Insights: Patterns of Hate Speech (2023–2025)





Key takeaways

- Over **70% of misinformation** in India is spread through WhatsApp.
- Whatsapp has been transformed into a **primary theater** for digital political warfare



How Hate Speech Spreads and Escalates

Context

- Tense social or political environment
- Historical and structural group divides
- Online anonymity

Key Spreaders

Political leaders & parties

Coordinated troll networks

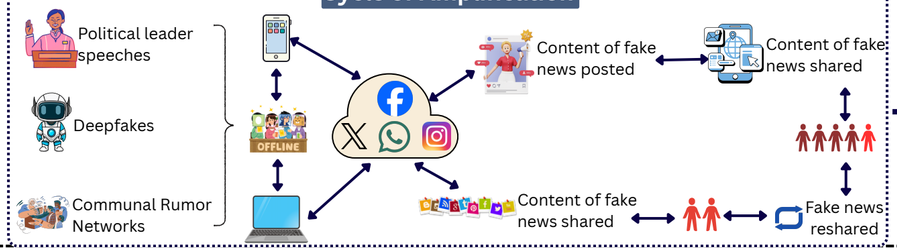
Extremists

Content

97% of tracked hate speeches are shared as videos

- **Deepfakes**: AI-generated "convincing" videos
- **Call to Action**: Direct calls for violence

Cycle of Amplification



Consequences

The average time to even register an FIR is **8.7 days.**

276 speeches called for the destruction of religious structures

Rise in calls for boycotts directly impacts the **livelihoods of minorities**



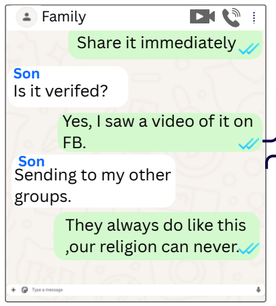
90% of eventual reach of content within 48 hours

Platform-Driven Amplification of Hate Speech

97% of tracked hate speeches originated as social media videos

Breakdown of first video upload

- 71.6%, 942 videos
- 18.7%, 246 videos
- 5.1%, 67 videos



Majority **(65–80%)** frequently encounter harmful content online.

Almost all hate speech incidents start offline and are amplified online

Why Falsehood Outpaces Facts



The Knowledge Gap

Lower factual knowledge increases resistance to correction due to weak evidence evaluation.



Social Affiliation

People are more likely to believe false information and reject corrections when it comes from a **trusted in-group source.**



Attribution Filter

When seen as "elite imposition," corrections can trigger a **backfire effect** and strengthen false beliefs.

Link Link Link

UNDERSTANDING PROBLEM

The Digital–Judicial Speed Mismatch

Enforcement & Legal Response Lag

within **48Hrs**
Misinformation reaches 90% of its total audience

However, the average time to register an FIR is **8.7 days**

Legal process only begins **200+ hours** after the damage is done.

Avg. Report to Legal Action

Germany	2-4 days
France	1-3 days
U.K.	3-7 days

The Preventive Failure

Research indicates that for legal action to have a deterrent effect, it must occur within **24 hrs**

The current 8.7-day average for FIRs means the law has "virtually no preventive effect" on content spread

Conviction Attrition



Only **16% of documented hate speech cases result in a conviction**. The total journey from viral content to a verdict takes an average of **5.3 years**

Structural Reasons for Time Lag

Infrastructure Deficits

Police Cyber Capacity

12,000
Total cyber police position

3,500
Actual filled position

187/766
District with no cyber cell

Court Infrastructure

Dedicated cyber courts: **21** (for population of 1.4 billion)

Judges trained in cyber law: **<500**

Pending cyber crime cases: **290,000+ (2022)**

Procedural Bottlenecks

Jurisdictional confusion: Average **15 days lost** determining which police station has jurisdiction for online content

Digital evidence collection: Most police stations lack equipment; **evidence collection takes 30-60 days**

Technical expertise: Courts often require expert testimony, adding **3-6 months**

Bail ease: 89% of hate speech accused **get bail quickly,** reducing urgency

A System Engineered for Outrage

Anchors are no longer neutral observers; they are "Performers."



4 out of 5 anchors uses aggressive tones

TV REACH plummeted from **94% in 2018 to 82% in 2022**



Young audiences prefer long-form YouTube content over TV news.

Young audiences reject bad journalism, not news.

Cost of lying < Profit from outrage

₹50 lakh in ads | ₹1 lakh in NBDSA fine

- PCI only warn (no fines)
- Fine is negilible for large channels
- No license cancellation threat

UK vs India Contrast

Forced public apology	Limited corrective accountability
Can fine millions	Small fines (Max of 1 lakh)
Can suspend license	No License Threat
Legally binding	Self regulating bodies

Algorithmic Amplification

Root Incentive

Engagement < Accuracy

Anger Content → High engagement

Neutral corrections spread slower

Echo Chambers

- Users see content aligned with their beliefs
- Repeated exposure reinforces bias
- Opposing views are filtered out

Speed Gap

Hate spreads in hours

Fact checking take hours

Social Polarization

- Increased distrust
- Community tensions
- Normalization of hate



Connectivity Without Critical Literacy

Core issue

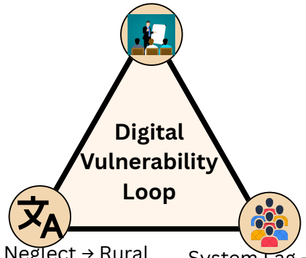
Rapid internet expansion without digital literacy training

5%
use verified fact-checking platforms.

<15%
of fact-checking efforts target vernacular language

65%
try to verify information, but mostly through informal searches

Training Deficit → Informal Verification



Language Neglect → Rural Truth Decay

System Lag → No Proactive Immunity

"Interest vs. Access" Tension



Hunger for Skill

84% of rural women expressed an active interest in receiving digital training.



Training Void

Despite this much high interest, 82% of respondents had **never received any formal digital literacy training.**

The system fails to provide accessible, community-based programs to meet this massive demand.

Strategic Interventions

Establish Community Digital Hubs

Vernacular Fact-Checking Networks

Link Link

Different User Behaviors Driving Online Hate Spread in India

COHORT 1: IDENTITY-DRIVEN AMPLIFIERS





Tier-2/3 youth, small business owners, and local community members

Expressive majority- active sharers driven by belonging.

Estimated Size : **175–200 million**

User Journey :

Receives provocative identity-based content in trusted WhatsApp community groups. → Skims headline quickly and interprets message as cultural threat. → Feels immediate anger and strong urge to react emotionally. → Forwards message instantly to friends, family, and local groups.

Underlying Behavior Patterns

- Strong identity attachment to **religion** and **community belonging**
- **Emotional triggers** of fear, pride, and perceived social threats
- **Confirmation bias** toward familiar narratives and in-group opinions
- Social **validation from peers** reinforces repeated sharing behavior

COHORT 2 : THE DISENGAGED MAJORITY





Middle-class working , homemakers, and students

the silent majority—neither amplifiers nor fact-checkers.

Estimated Size : **245–280 million**

User Journey :

Receives hate or political content in familiar WhatsApp groups. → Skims message quickly with mild internal discomfort or concern. → Avoids responding publicly due to time and conflict avoidance. → passive exposure gradually normalize such harmful content.

Underlying Behavior Patterns

- Cognitive **overload** from work and daily life pressures
- **Conflict avoidance** in social and family WhatsApp groups
- Bystander effect: assume **others will respond** instead
- Emotional fatigue from **constant negative** news cycles online

COHORT 3 : THE RATIONAL CORRECTORS





Urban students, young professionals, and digitally literate fact-checkers

Intent-driven minority—fact-checkers seeking accuracy but lacking amplification.

Estimated Size : **70–85 million**

User Journey :

Receives viral hate or misinformation across social media platforms. → Feels concern and responsibility to verify authenticity of content. → Searches multiple sources to confirm facts and context. → Attempts responding but lacks support and visible amplification.

Underlying Behavior Patterns

- Strong belief in fairness, accuracy, and responsible public discourse
- Motivation driven by civic responsibility and social harmony
- Fear of backlash or isolation reduces consistent participation
- Low visibility and support weakens long-term corrective engagement

Link

SOLUTION : OVERVIEW

Community-Driven Verification Network

- A non-hierarchical network spanning urban, semi-urban, and offline rural regions.
- Blends human trust networks and digital infrastructure to accelerate verification and correct information dissemination.
- Designed to amplify verified content through structured, community-centric channels.

Operational Framework to Contain Misinformation

Before Organization

2 DAYS TO GO

Organic rumors spread unchecked for **~72–96 hours,** especially via intimate, encrypted networks.

FACT

Fact checks often **fail to reach the same social clusters** rapidly

Before Organization

Structured channels + trained community moderators = verified correction circulation within **48Hrs**

Key Mechanism

- Rapid escalation path
- Humancentric trust nodes for amplification
- Avoidance of unstructured forwarding

Volunteer engagement & Capacity Building

Volunteer Profile

College Social Service Club Member

- Active member of college social service club
- Engaged in outreach, awareness, and local engagement
- Connected with peers and community networks

16,000+ colleges have active social-service clubs

70% of student volunteers report improved leadership skills

Value Proposition for Joining Organization X

- Hands-on Responsibility
- Practical Skills & Experience
- Network & Growth

How we stop misinformation spread

Rigorous Verification

All published content is verified against credible sources and reviewed by trained fact-checkers.

Clear Labeling

When information is unverified or may be misleading, we label it clearly so audiences understand its status

Education and Awareness

We educate communities through trusted rural opinion leaders, urban volunteers, and school awareness programs.

Responsible system design

When information is unverified or may be misleading, we label it clearly so audiences understand its status

Machine Learning (ML) : Detection & Trend Signal

Automatically identify and prioritize misinformation signals from massive information streams.

How it's useful

Detects patterns humans can't process at scale.

Flags trending misinformation early.

Helps volunteers and fact-checkers focus efforts where they matter most.

Digital Layer : WhatsApp & Online Groups

Primary Digital Infrastructure

Language-aligned regional WhatsApp groups as front-line nodes

Dedicated moderation roles ensure:

- Claim triage
- Preliminary context checks
- Escalation to verification hub

Why whatsapp

Secure Encrypted Messaging

Widespread User Access

WhatsApp

Instant Direct Communication

User & Community Engagement

Panchayat & Community Meetings:

We regularly engage village panchayats and community leaders to exchange insights and monitor emerging misinformation trends.

On-Ground Volunteer Interaction

Volunteers monitor WhatsApp and local networks daily, flagging misinformation hotspots and escalating unusual trends for verification.

Feedback Integration

Community surveys and regional feedback are used to continuously refine our verification processes and awareness strategies.

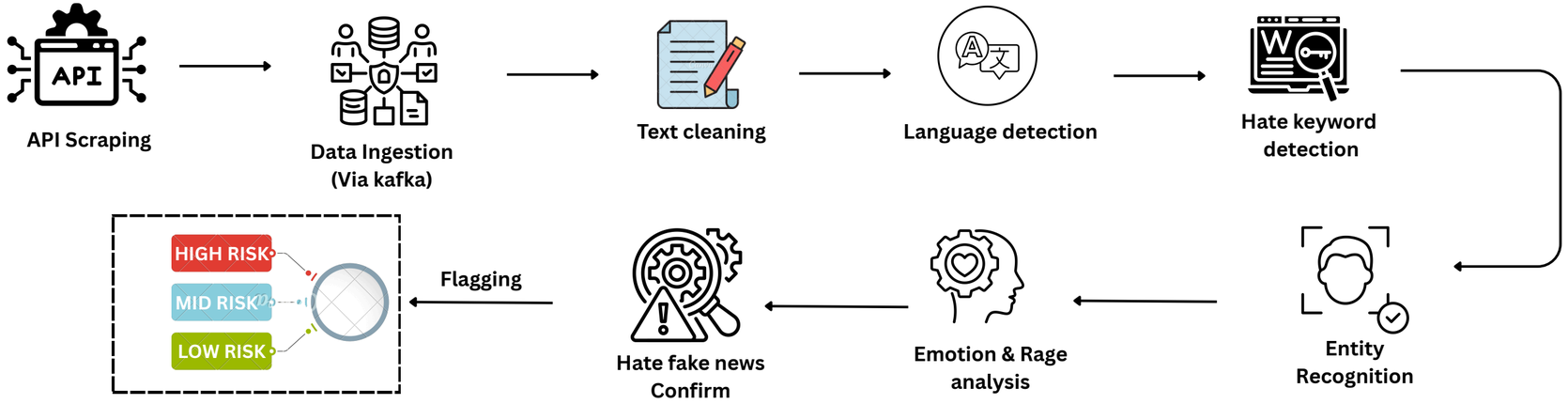
SOLUTION BRIEF

- *AI-powered backend system that scans digital media posts, comments, and news content in real time*
- *Scalable across regions, languages, and multiple digital platforms*
- *Can scan millions of posts per month, enabling large-scale monitoring*

A monitoring dashboard displays flagged content, trends, and alerts for continuous tracking and intervention.

- 🔴 Red: High-risk (violent or dangerous hate content)
- 🔵 Blue: Medium-risk (offensive or rising harmful narrative)
- 🟢 Green: Low-risk (safe or minimal concern)

SYSTEM ARCHITECHTURE



PREDICTING PARAMETERS

Hate keyword density

slurs, communal terms, violent verbs.

Toxicity score

using Detoxify, Perspective API.

Target detection

detect if text targets religion, caste, gender

Emotion & Rage

- RoBERTa emotion model

Context-Based Parameters

Virality score , User influence
Temporal spike detection

IMPACTS

2M+

DIGITAL POSTS PER REGION MONTHLY

across social media, forums, and news comments.

15K–25K

HATE/OFFENSIVE POSTS DETECTED PER MONTH

Fullfulling our trrigering parameters

1.5K-3K

POTENTIALLY VIOLENT/INCITING POSTS MONTHLY

Flagged as RED

60–70%

FASTER DETECTION VS MANUAL MONITORING

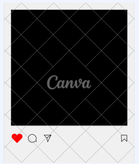
Through our local deployed models

30–40%

REDUCE SPREAD OF HIGH-RISK HATE

By Works of our volunteer

WORK FLOW



HATE SPEECH IS POSTED



ML- MODEL DETECTS



RESEARCH TEAM



CONTENT TEAM

ONLINE AND OFFLINE AWARENESS

- Gives empowerment to the 2nd user cohort to amplify their voices
- Makes an Organisation Brand image which can be later used as a certification of news
- Counters the amplication power of the hate speech
- Once this gets into circulation the algorithm will automatically get neutral leading to less provoking by giving neutral

ONLINE AWARENESS

- Online presence is most important or teir 1 - 2 places
- they will have the maximum reach over the wide online population
- this will create our brand image
- we will have an intense feedback loop to support the suastainibility andkeep growing



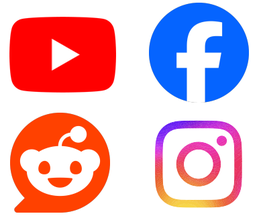
Our institute chapters will indivisually operate at the local level



Personel website for people to check out real time updates on local news



Articles and pblogs will be posted at different websites to gain engagement



Videos posting at all different social media platforms

OFFLINE AWARENESS

TIERED ECOSYSTEM

Urban and internet Connected Users

Regional , language based Whatsapp Groups

serves as hyperlocal hubs for incoming claims spreading

Moderators

- Training college students
- Social Service club members

Multi directional , rapid feedback loop

Ofline villages (No network zone)

Local Volunteers

↓

Sarpanch

↓

Local influential persons

↓

Local aware people

Link

Digital Awareness Week
(1st Week of February)

⚠️ Festive months often see a rise in hate & misinformation

What We Do


Social Media Awareness


Nukkad Natak


Group Discussions


Poster / Reel Creation


Peace & Unity Activities



Myth vs Fact Sessions

Goal

- Media literacy & responsible social media use
- Identifying misinformation & hate narratives
- Ethical sharing: Think Before You Share
- Positive content creation

Why it is effective

- Research shows misinformation spreads mainly due to lack of media literacy, not intent
- Teaching “how to verify before sharing” directly targets the decision point of hate spread
- Youth-led content increases peer credibility, making messages more likely to be accepted

📌 UNESCO highlights media literacy as a primary tool to counter online hate and misinformation.

Hate speech damages people, peace & democracy



“Think Before You Share”



Offline Community Outreach

⚠️ Where hate spreads easily, but awareness doesn’t
📢 Grassroots Awareness in Low-Connectivity Areas

Target

📍 Areas with limited internet access
👥 Students (Classes 8–12)
🔁**Frequency:** Once every month

WHY 8 - 12: 🌱 Opinions start forming & voices become stronger

To Do

- Grassroots awareness in low-connectivity areas
- Student-led nukkad nataks
- Moderated group discussions & open dialogues
- Sessions recorded and amplified online

Activities


Student-led Nukkad Natak


Debate & Group Discussions


Open Question Sessions

Online Amplification

Record sessions → Share on social media → reach wider audience



Why it is effective

- Hate often spread through informal conversations
- Face-to-face dialogue reduces polarization by allowing questions, clarification, and empathy
- Community-based interaction builds trust, which online campaigns often lack

School Based Awareness
Building values before hate takes root

Timing

📅 Twice a year
- Before Summer Break
- Before Winter Break

Summer and winter breaks are chosen because students are most digitally active and psychologically receptive during these periods, making pre-break intervention both preventive and impactful.

What to do

1️⃣ **Training Sessions**
🧠 What is hate speech?
⚖️ Its social & legal consequences
📱 Responsible digital behavior

2️⃣ **Skill Building**
🎤 Communication & dialogue
🤝 Conflict resolution
🧩 Critical thinking

Why it is effective

- Classes 8–12 are a critical belief-formation stage
- Early exposure prevents normalization of hate before opinions solidify
- Skill-building (communication, conflict resolution) creates long-term behavioral change, not just awareness

📌 UN and UNICEF stress early educational intervention as the most sustainable prevention strategy.

When youth speak for peace, hate loses its voice.

Awareness Today, Harmony Tomorrow

Link Link Link

Policy Reforms : Law shapes Behaviour

---Legal & Digital Gaps---


No Clear Definition


66A struck down


Digital vacuum


Post-harm focus

👧 SC (2015): Vague laws violated free speech (Art. 19)

📈 MIT (2018): False news spreads **6x faster** than truth

---Needed Reforms---

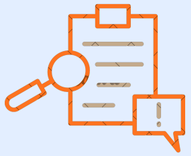

Clear Definition


Digital Framework


Platform Accountability


Fast FIR


Free Speech Safeguard


Transparent Moderation

Japan (2016 Hate Speech Act) : Law passed + Human Rights education integrated in schools → Preventive, curriculum-based hate control model

Shreya Singhal v. Union of India (2015) : Sec. 66A struck down (unconstitutional) → Created digital regulation gap needing clearer hate speech framework

Curriculum Reform: Education Shapes Mindset

---Learning Gaps---


No Media Literacy


Early Bias Formation


Textbook-only Civics


High Media Exposure

UNESCO (2023): Prejudice peaks during adolescence

OECD: Early civic education reduces extremism tolerance

---Needed Reforms---


Empathy Building


Respect for Diversity


Understanding Hate Speech Impact


Digital Behaviour Awareness


Media Literacy


Legal Understanding

🧒 Classes 1–5 (Foundation Stage)

🧑 Classes 6–8 (Awareness Stage)

🎓 Classes 9–12 (Critical Stage)

Link Link

Key Risks and Mitigation Strategies for overall success

	Key Risks				Mitigation Strategies			
Prediction	over dependance on automation	Bias against specific communities	False positives (normal content flagged)	Data drift over time	Human-in-the-loop review	Train on multilingual datasets	Periodic bias audits	continuous model retraining
Intervention	Accusations of free speech violation	Political / ideological backlash	Slow response in viral cases	Mass false reporting to target rivals	Clear public harm-based policy	Transparent enforcement guidelines	Real-time alert escalation system	Reporter trust scoring
Awareness	Low participation / lack of trust	Rural level disconnect	Awareness campaigns losing momentum	Funding constraints	Partner with schools, NGOs, Panchayat	Local-language sessions	Monthly mini-events	CSR + government digital literacy grants

Mitigation strategies Implementation

PREDICTION

Establish a **multilingual data pipeline** that collects public content across regions to ensure **inclusive model training** and reduce systemic bias

Deploy **context-aware NLP models** that detect harmful narratives , supported by a **human-in-the-loop review** system for sensitive case

Implement **confidence thresholds** before takedown actions to **minimize false positives** and protect legitimate speech

Continuously **retrain models** and conduct **bias audits** to adapt to evolving language patterns and **prevent model drift.**

INTERVENTION

Activate a **real-time alert escalation system** to prioritize viral and high-risk content before it spreads widely.

Publish a **clear harm-based policy framework** distinguishing hate speech from opinion to maintain transparency and public trust.

Introduce a **reporter trust-scoring** mechanism with rate limits to **prevent coordinated misuse** or mass-reporting abuse.

Apply a **graduated enforcement model —** warning, education, then restriction — focusing on behavior correction before punitive action.

AWARENESS

Build **partnerships with schools, NGOs, and Panchayat leaders** to conduct localized awareness sessions in regional languages.

Run **digital campaigns and social media outreach** using short videos, and myth-busting content to reach youth audiences

Introduce **gamified learning modules** and real-life simulations to encourage long-term behavioral change.

Develop a **network of student ambassadors and community volunteers** to sustain monthly events and grassroots engagement

Link Link

Thank You

~One viral lie can destroy a life.
One responsible system can save thousands.

Let’s choose responsibility.


ViralPredictor
AI-powered viral potential analysis for social media content
Content
Unpopular opinion: a quick thread with steps on AI. Save this.
Reach (1h)
25000
Likes (1h)
1800
Predict Viral Potential
10
🔥 Potential Mega-Viral Content!

```
Click to add a breakpoint
import pandas as pd

from sklearn.model_selection import train_test_split
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import mean_absolute_error, accuracy_score, confusion_matrix, classification_report

class ViralOrdinalTrainer:
    """
    Ordinal classification via cumulative link idea:
    Train 9 binary models for: P(rating > k), k=1..9
    Predict rating = 1 + count_k( P(r>k) >= threshold )
    """

    def __init__(self, csv_path: str, seed: int = 42):
        self.csv_path = csv_path
        self.seed = seed
        self.preprocess = None
        self.threshold_models = {}  # k -> Pipeline

    def _build_preprocess(self):
        return ColumnTransformer(
            transformers=[
                ("text", TfidfVectorizer(ngram_range=(1, 2), max_features=6000), "content"),
                ("nums", Pipeline([("scaler", StandardScaler())]), ["reach_1h", "likes_1h"]),
            ],
            remainder="drop"
        )

    def fit(self):
        df = pd.read_csv(self.csv_path)

        # Basic checks
```

```python
# Basic checks
expected_cols = {"content", "reach_1h", "likes_1h", "rating_viral_1to10"}
missing = expected_cols - set(df.columns)
if missing:
    raise ValueError(f"CSV missing columns: {missing}")

X = df[["content", "reach_1h", "likes_1h"]].copy()
y = df["rating_viral_1to10"].astype(int).copy()

X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=self.seed,
    stratify=y
)

self.preprocess = self._build_preprocess()
self.threshold_models = {}

for k in range(1, 10):
    yk_train = (y_train > k).astype(int)

    clf = LogisticRegression(
        solver="saga",
        max_iter=4000,
        n_jobs=-1,
        class_weight="balanced",
```

```
            random_state=self.seed
        )

        pipe = Pipeline([
            ("preprocess", self.preprocess),
            ("clf", clf)
        ])

        pipe.fit(X_train, yk_train)
        self.threshold_models[k] = pipe

    # ---- Evaluate ----
    y_pred = self.predict_batch(X_test)

    # Metrics
    mae = mean_absolute_error(y_test, y_pred)
    exact_acc = accuracy_score(y_test, y_pred)

    # within-1 accuracy: prediction off by at most 1
    within1 = float(np.mean(np.abs(y_pred - y_test.values) <= 1))

    print("=== Test Metrics (20% split) ===")
    print("MAE (lower is better):", round(mae, 4))
    print("Exact accuracy:", round(exact_acc, 4))
    print("Within ±1 accuracy:", round(within1, 4))

    print("\n=== Confusion Matrix (rows=true, cols=pred) ===")
```

```
    labels = list(range(1, 11))
    cm = confusion_matrix(y_test, y_pred, labels=labels)
    print("Labels:", labels)
    print(cm)

    print("\n=== Classification Report ===")
    # Note: with imbalance, some high classes may have low support -> report is still useful
    print(classification_report(y_test, y_pred, labels=labels, zero_division=0))

def predict_one(self, content: str, reach_1h: int, likes_1h: int, threshold: float = 0.5) -> int:
    if not self.threshold_models:
        raise RuntimeError("Train first by calling fit().")

    X_new = pd.DataFrame([{
        "content": content,
        "reach_1h": int(reach_1h),
        "likes_1h": int(likes_1h),
    }])

    rating = 1
    for k in range(1, 10):
        p = self.threshold_models[k].predict_proba(X_new)[0, 1]  # P(rating > k)
        if p >= threshold:
            rating += 1

    return int(np.clip(rating, 1, 10))
```

```python
    def predict_batch(self, X_df: pd.DataFrame, threshold: float = 0.5) -> np.ndarray:
        if not self.threshold_models:
            raise RuntimeError("Train first by calling fit().")

        # Compute P(r>k) for all k, then count thresholds exceeded per row
        probs = []
        for k in range(1, 10):
            p = self.threshold_models[k].predict_proba(X_df)[..., 1]  # shape (n,)
            probs.append(p)

        probs = np.vstack(probs).T  # shape (n, 9)
        exceeded = (probs >= threshold).sum(axis=1)
        pred = 1 + exceeded
        return np.clip(pred, 1, 10).astype(int)


if __name__ == "__main__":
    trainer = ViralOrdinalTrainer(
        csv_path="viral_social_media_dataset_1500_imbalanced.csv",
        seed=42
    )
    trainer.fit()

    # Example inference
    print("\nExample prediction:")
    print(trainer.predict_one(
        content="Unpopular opinion: a quick thread with steps on AI. Save this.",
        reach_1h=25000,
        likes_1h=1800
```