35TH USENIX SECURITY SYMPOSIUM

Cross-National Information Attacks: A Two-Decade Analysis of Troll Behavior in Korea

Jaehong Kim, Hyeonseung Kim, Jiseon Kim

Alice Oh, Thorsten Holz, Wonjae Lee, Meeyoung Cha

KAIST

MAX PLANCK INSTITUTE FOR SECURITY AND PRIVACY

Foreign influence trolls are a growing threat to democracy

Reuters, February 2026

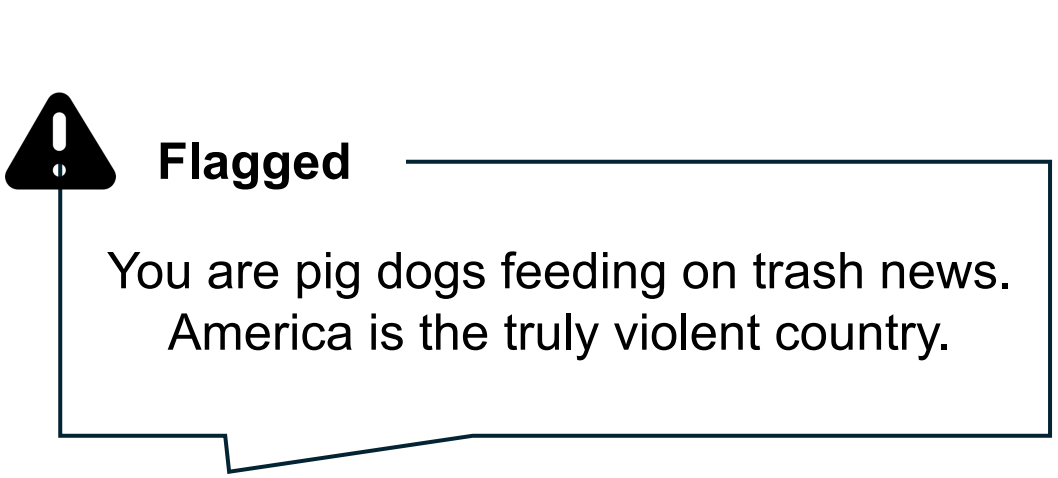


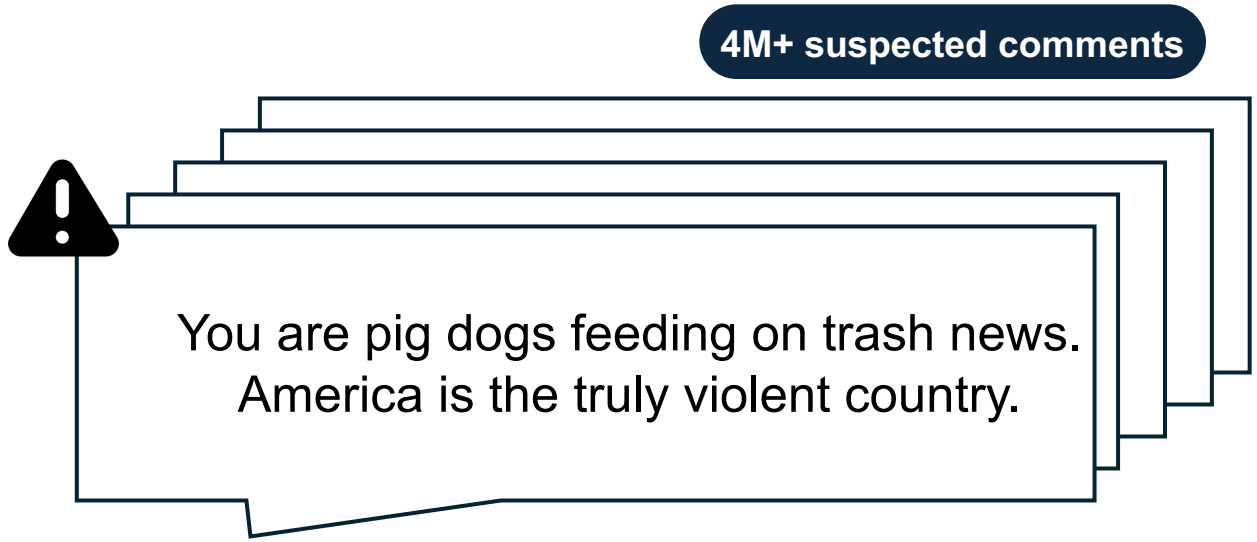
OpenAI report, June 2026



Source: Reuters (Feb 2026); OpenAI (Jun 2026). Country names replaced with [Foreign]/[State] for neutrality.

Detecting state-linked trolls is important.
But detection alone is not enough for stakeholders.





Platform moderators ask:
"Why is it flagged?"

Influence operation analysts ask:
"Which messages are critical?"

From explainable detection to strategy analysis

① Explainable content detection

Foreign suspected → Moral Emotion → Target Country

Flagged

Writing region: [KR, MNS] ⟨Foreign-suspected⟩

Pig dogs feeding on trash news. ⟨Condemning → Korea⟩

America is the violent country. ⟨Condemning → Rival⟩

② User Extraction

20 Years
112M Comments
4.3M Users

Suspected 24K Users
4M Comments

③ Strategy Analysis

Longitudinal Trend Analysis

Engagements and Visibility

Targeted Entities 🎯

From explainable detection to strategy analysis

① Explainable content detection

Foreign suspected → Moral Emotion → Target Country

Flagged

Writing region: [KR, MNS] ⟨Foreign-suspected⟩

Pig dogs feeding on trash news. ⟨Condemning → Korea⟩

America is the violent country. ⟨Condemning → Rival⟩

② User Extraction

20 Years
112M Comments
4.3M Users

Suspected 24K Users
4M Comments

③ Strategy Analysis

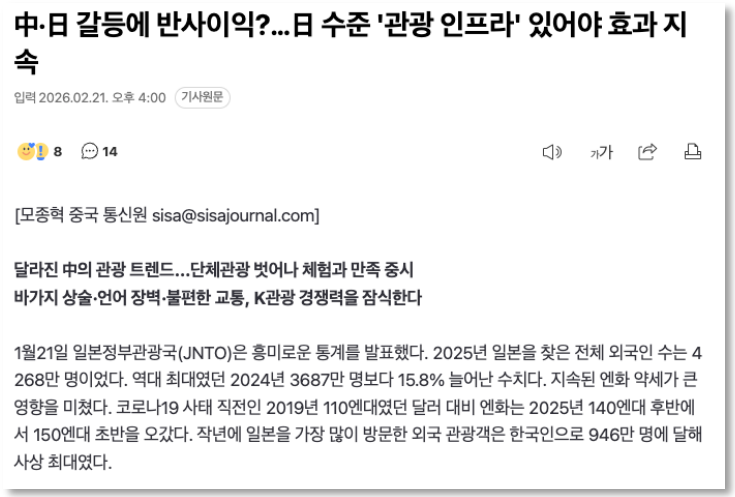
Longitudinal Trend Analysis

Engagements and Visibility

Targeted Entities 🎯

Data Scope: Naver News Platform

- INSS(Institute for National Security Strategy) identified **70 troll accounts** linked to a **Major Neighboring State** (**MNS**, anonymized): our **Known Trolls**
- Target for Naver News where Korea's main news platform, used by 63% of adults



中·日 갈등에 반사이익?...日 수준 '관광 인프라' 있어야 효과 지속

입력 2026.02.21. 오후 4:00 기사원문

8 14

[모종혁 중국 통신원 sisa@sisajournal.com]

달라진 中의 관광 트렌드...단체관광 벗어나 체험과 만족 중시
바가지 상술·언어 장벽·불편한 교통, K관광 경쟁력을 잠식한다

1월21일 일본정부관광국(JNTO)은 흥미로운 통계를 발표했다. 2025년 일본을 찾은 전체 외국인 수는 4 268만 명이었다. 역대 최대였던 2024년 3687만 명보다 15.8% 늘어난 수치다. 지속된 엔화 약세가 큰 영향을 미쳤다. 코로나19 사태 직전인 2019년 110엔대였던 달러 대비 엔화는 2025년 140엔대 후반에서 150엔대 초반을 오갔다. 작년에 일본을 가장 많이 방문한 외국 관광객은 한국인으로 946만 명에 달해 사상 최대였다.



현재 댓글 14 | 작성자 삭제 0 | 규정 미준수 0

댓글을 입력해주세요

순공감순 최신순 **공감비율순**

클린봇이 악성댓글을 감지합니다. 설정

real**** 2026.02.21. 10:52 팔로우

저것들이 오는게 이익인거나? 어후

답글 4 1

digi**** 2026.02.21. 09:14 팔로우

지금일본인데 줄긴 줄어도 짱고라 목소리 많이 들림..

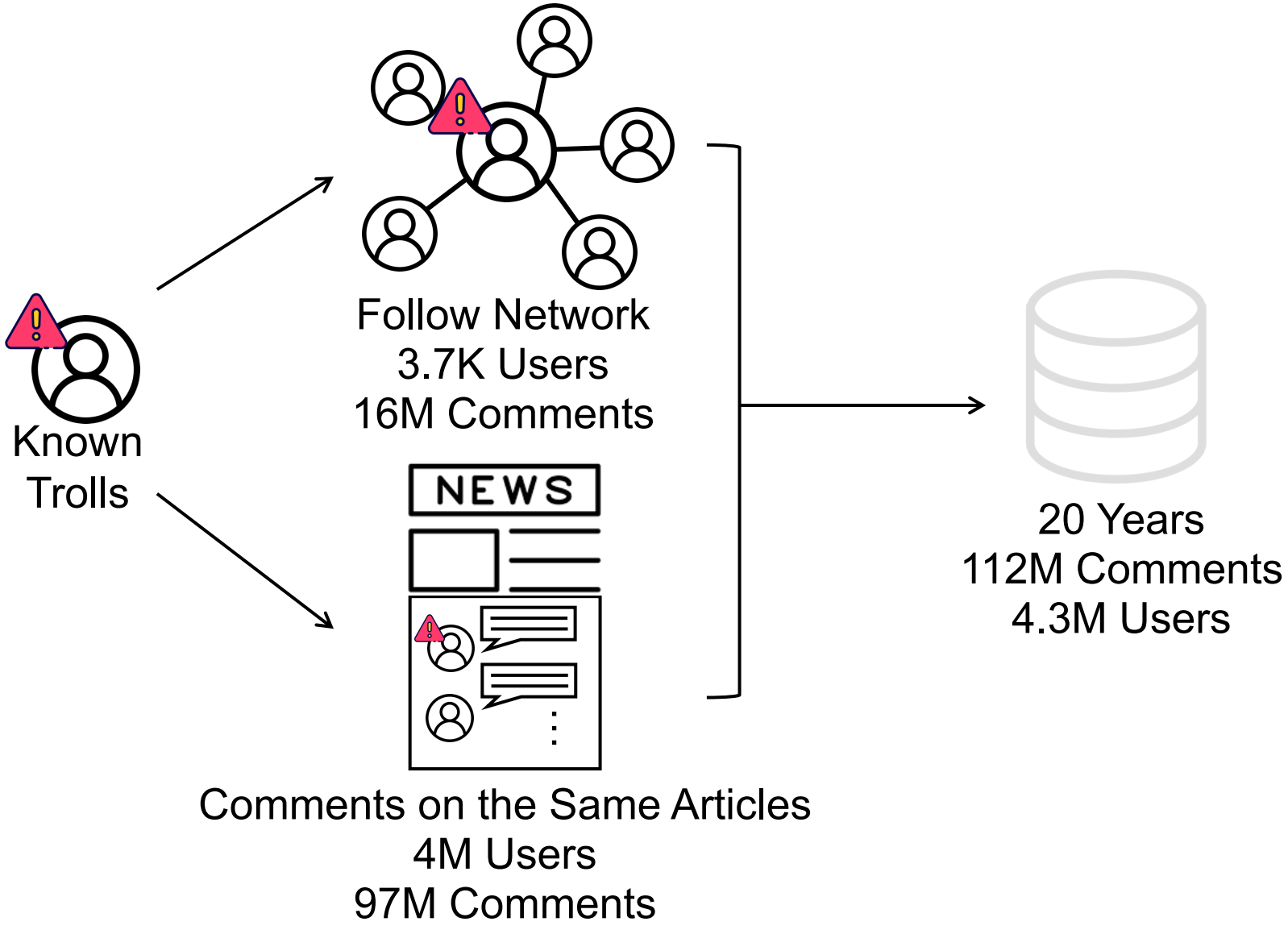
답글 3 11 3

News

Comments

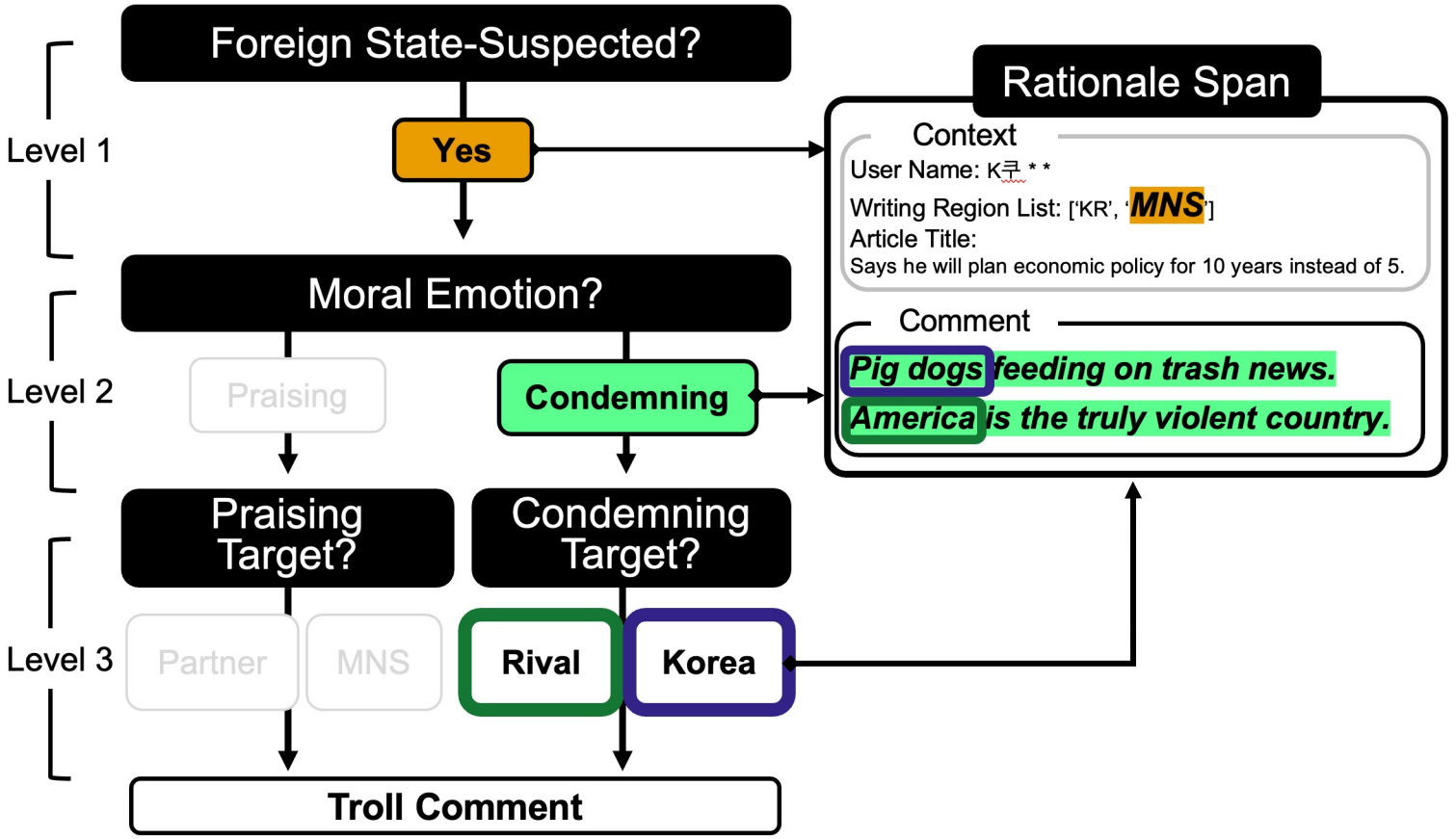
- Ranked by **Like Ratio**
- Like Ratio = #Like / (#Like + #Dislike)

Data Collection: 112M Troll-Linked Comments over 20 Years



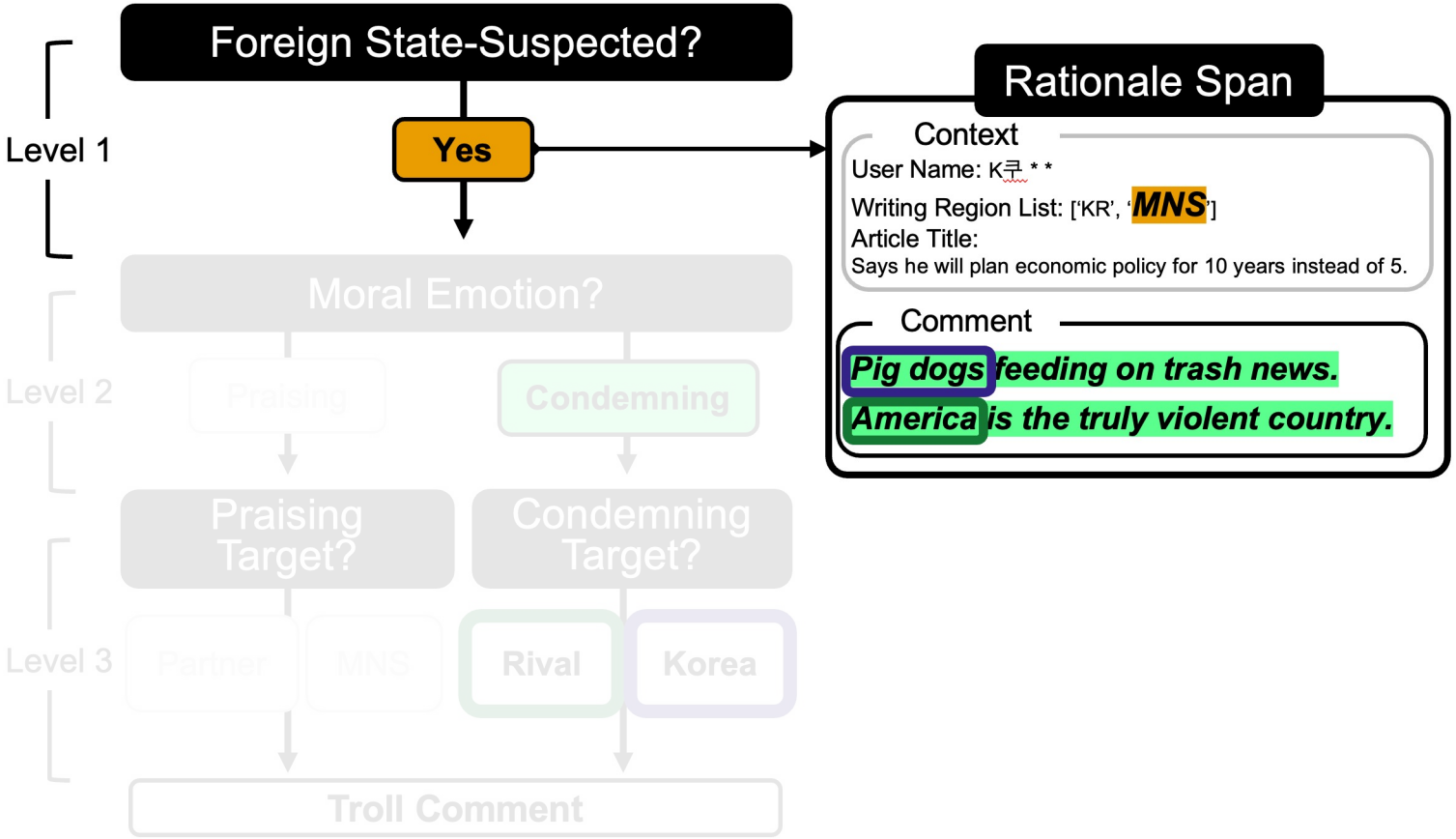
Explainable content detection: annotation structure

Three-level hierarchy, each level backed by a **rationale span**



Explainable content detection: annotation structure

Three-level hierarchy, each level backed by a **rationale span**

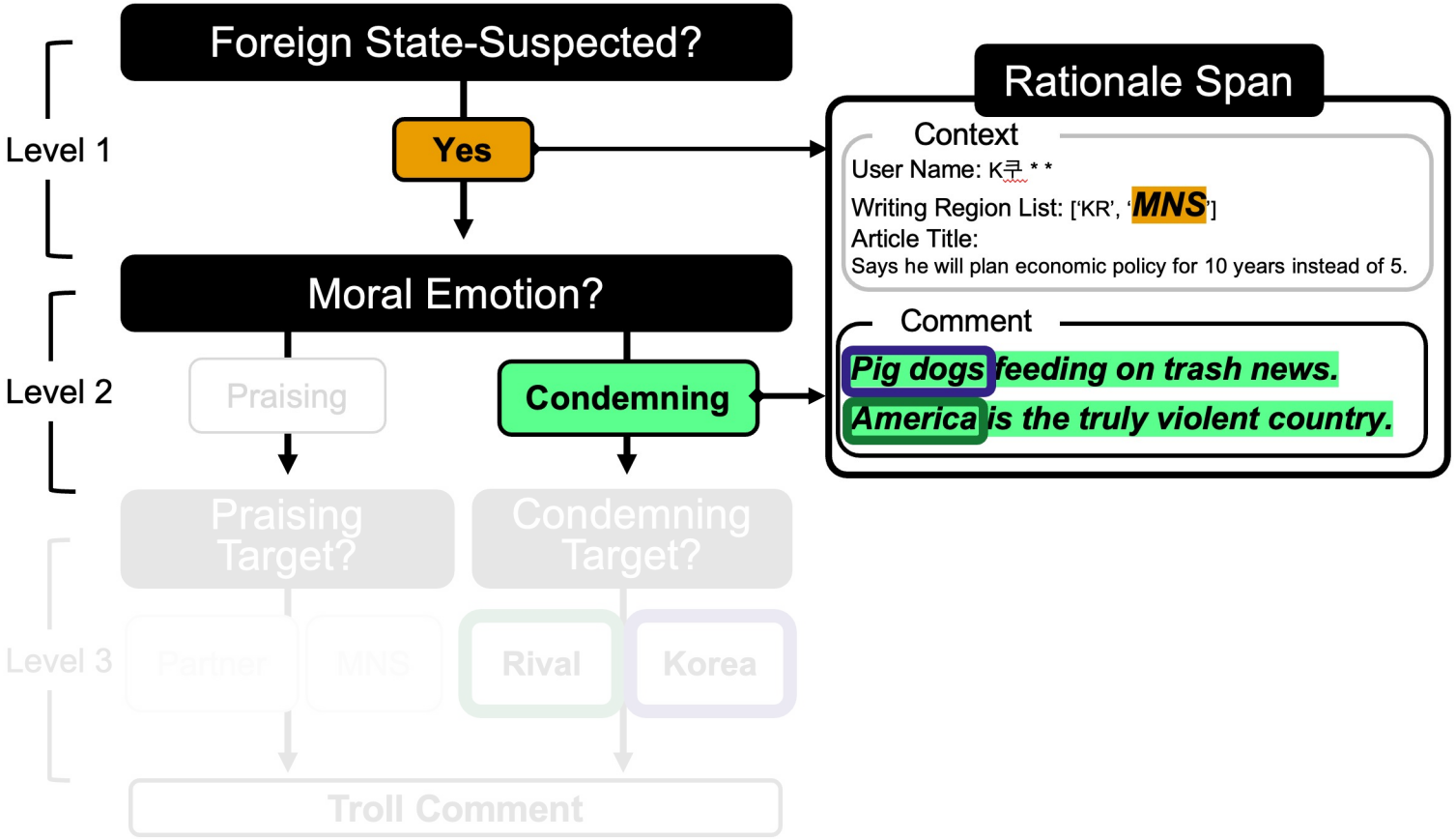


Foreign State-Suspected

- Posted from abroad
- Foreign-derived slurs
- MNS-linked terms in usernames/comments

Explainable content detection: annotation structure

Three-level hierarchy, each level backed by a **rationale span**



Foreign State-Suspected

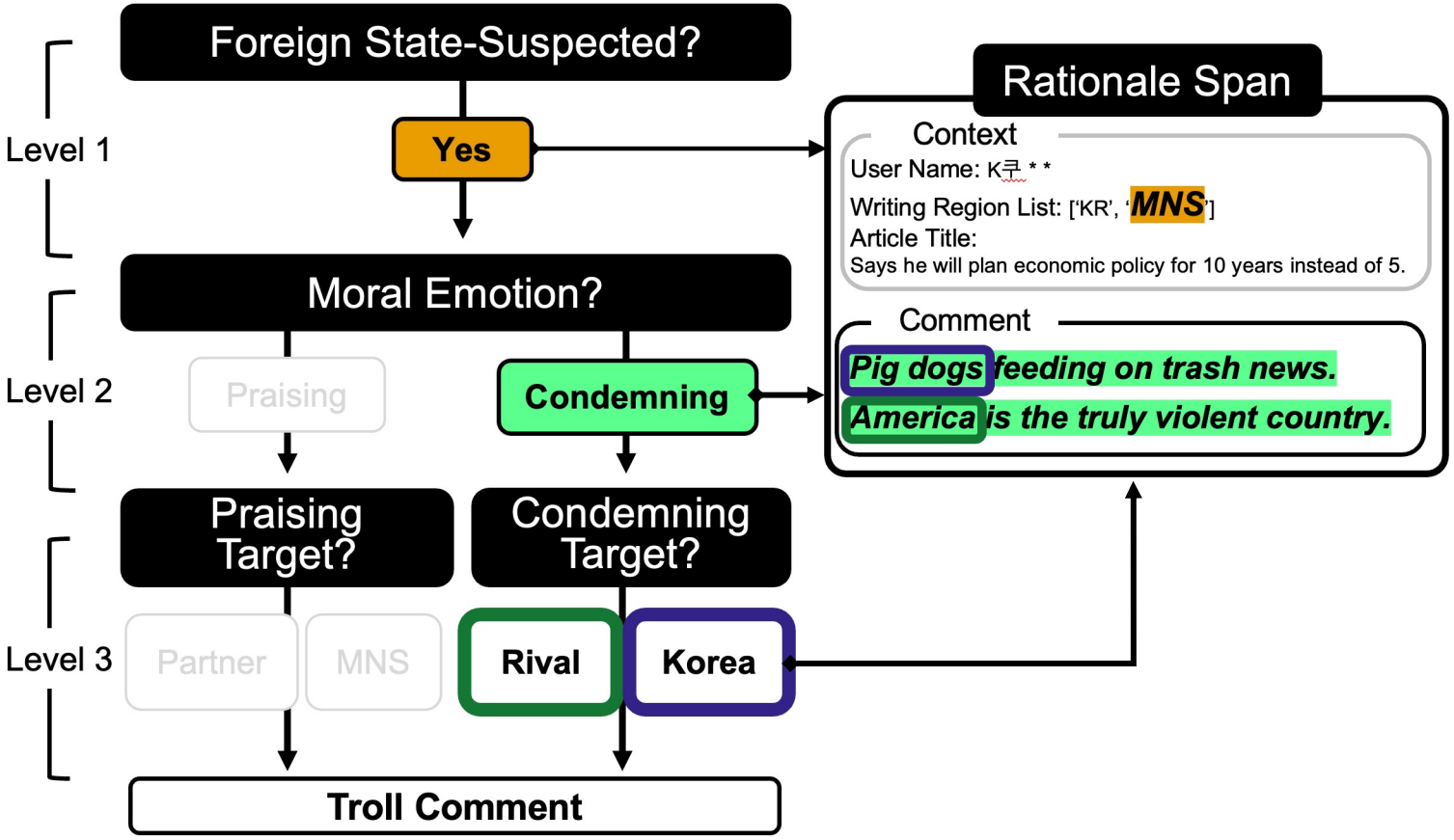
- Posted from abroad
- Foreign-derived slurs
- MNS-linked terms in usernames/comments

Moral Emotion: fuels polarization

- Other-condemning
- Other-praising

Explainable content detection: annotation structure

Three-level hierarchy, each level backed by a **rationale span**



Foreign State-Suspected

- Posted from abroad
- Foreign-derived slurs
- MNS-linked terms in usernames/comments

Moral Emotion: fuels polarization

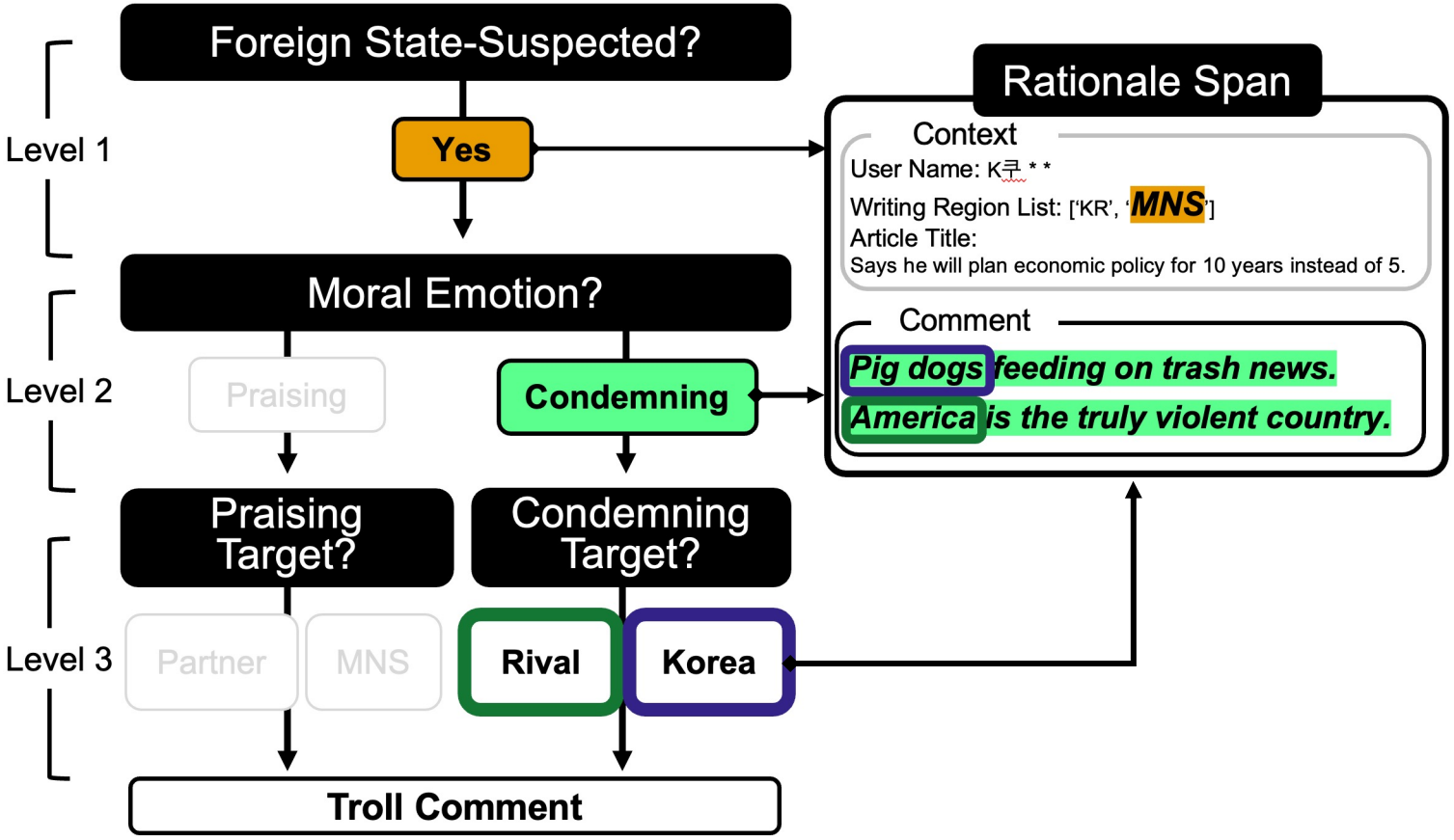
- Other-condemning
- Other-praising

Target Country of the Emotion

- Condemning or praising a country's people, leaders, or issues

Explainable content detection: annotation structure

Three-level hierarchy, each level backed by a **rationale span**



Foreign State-Suspected

- Posted from abroad
- Foreign-derived slurs
- MNS-linked terms in usernames/comments

Moral Emotion: fuels polarization

- Other-condemning
- Other-praising

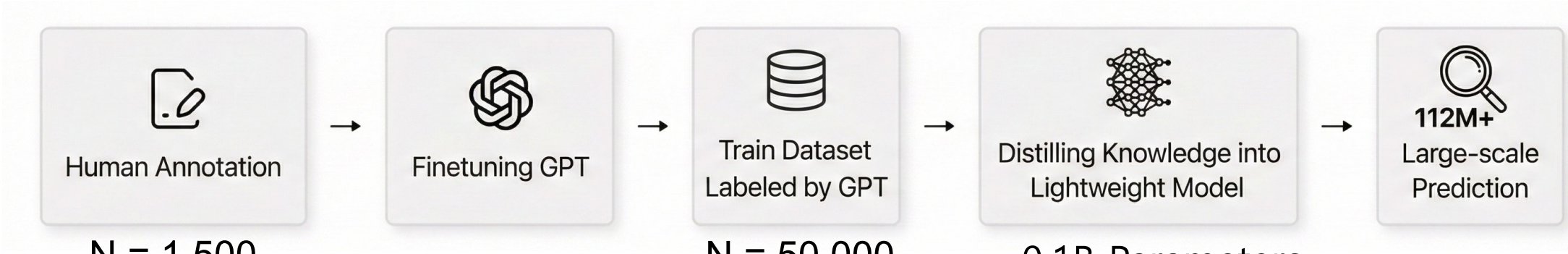
Target Country of the Emotion

- Condemning or praising a country's people, leaders, or issues

Troll label requires **all three levels** to align:

- Foreign → Praising → MNS(Major neighboring state of Korea) or MNS's Partners
- Foreign → Condemning → Korea or MNS's Rivals

Explainable content detection: implementation



N = 1,500
3 Researcher

N = 50,000

≈ 0.1B Parameters
- Class F1 = 0.84
- Text Span F1 = 0.58

Gold Label

English translation

User Name: zhanlie****
Writing Region List: ['KR' 'MNS']
Article Title: Explanation from former Girls' Generation member Jessica's side regarding the rumor about her appearance in an MNS entertainment program as an 'MNS minority group'.
Comment: I am an MNS minority group living in Japan, and an MNS minority group have the highest university graduation rate and personal income among the 56 ethnic minorities in MNS.
MNS is a powerful country, the second largest in the world. From Jessica's perspective, there's no need to be an MNS minority group, but it's still a hundred times better than a Korean who only listens to Americans and has no future. Haha.

Foreign-State Suspected

Praising MNS

Condemning Korean

Model Prediction

English translation

User Name: zhanlie****
Writing Region List: ['KR' 'MNS']
Article Title: Explanation from former Girls' Generation member Jessica's side regarding the rumor about her appearance in an MNS entertainment program as an 'MNS minority group'.
Comment: I am an MNS minority group living in Japan, and an MNS minority group have the highest university graduation rate and personal income among the 56 ethnic minorities in MNS.
MNS is a powerful country, the second largest in the world. From Jessica's perspective, there's no need to be an MNS minority group, but it's still a hundred times better than a Korean who only listens to Americans and has no future. Haha.

User-level detection: behaving like a known troll

One troll-like comment is not enough: it produces too many false positives.

Task: classify **70 known trolls** vs. **81 verified non-trolls** with 24 features

Content

- Content-level model's output
- Reposts identical text

Metadata

- Likes and dislikes
- Activity time
- Span between first and last comment
- Share of repeated posts

Network

- Interacts with known troll accounts
- Praises MNS in the same pattern

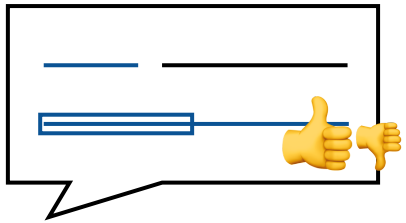
Result: XGBoost (**F1 = 0.94**) on all 4.3M users → **24K suspected trolls** (0.59%)
Flagged users match known trolls in activity timing
These users wrote **~4M troll comments**, the basis for all following analyses

Strategy analysis: 24K suspected trolls over 20 years

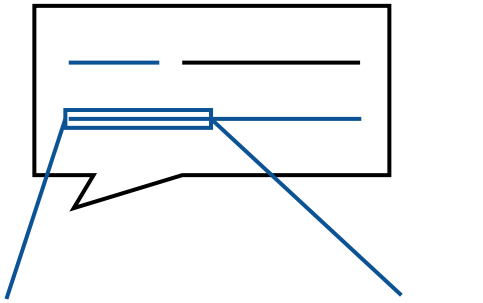
Three analyses on 4M comments from the suspected accounts



Longitudinal Trend Analysis



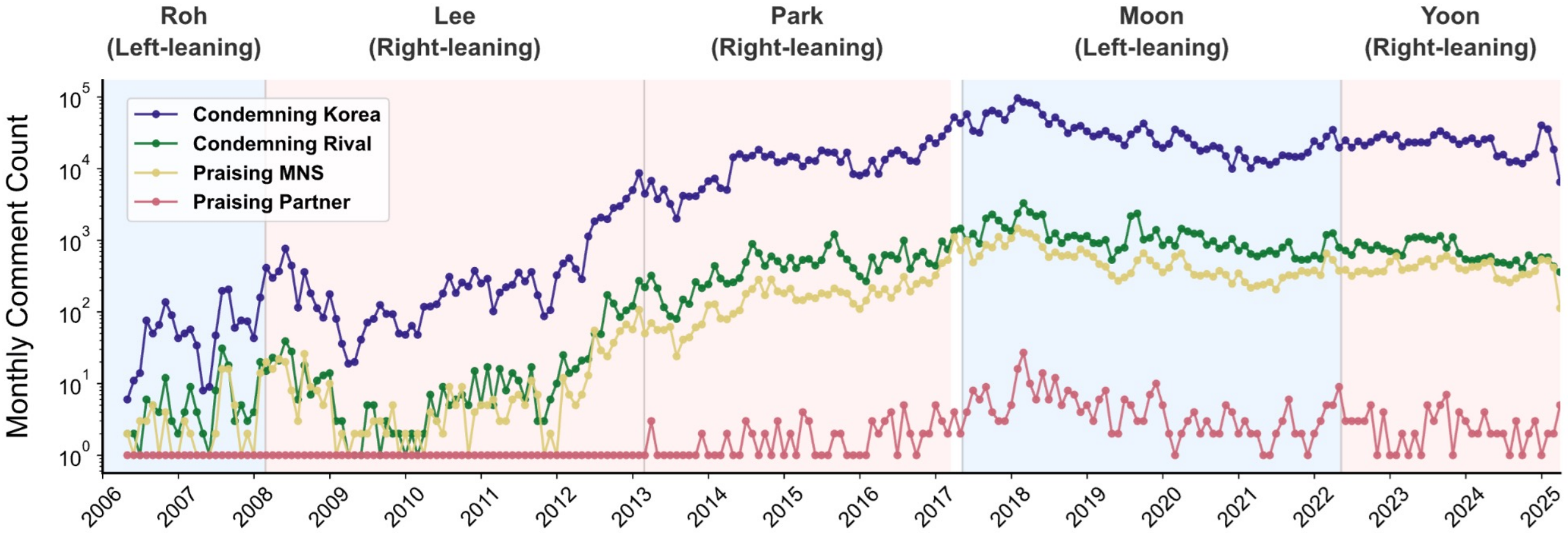
Engagements and Visibility



Targeted Entities🎯

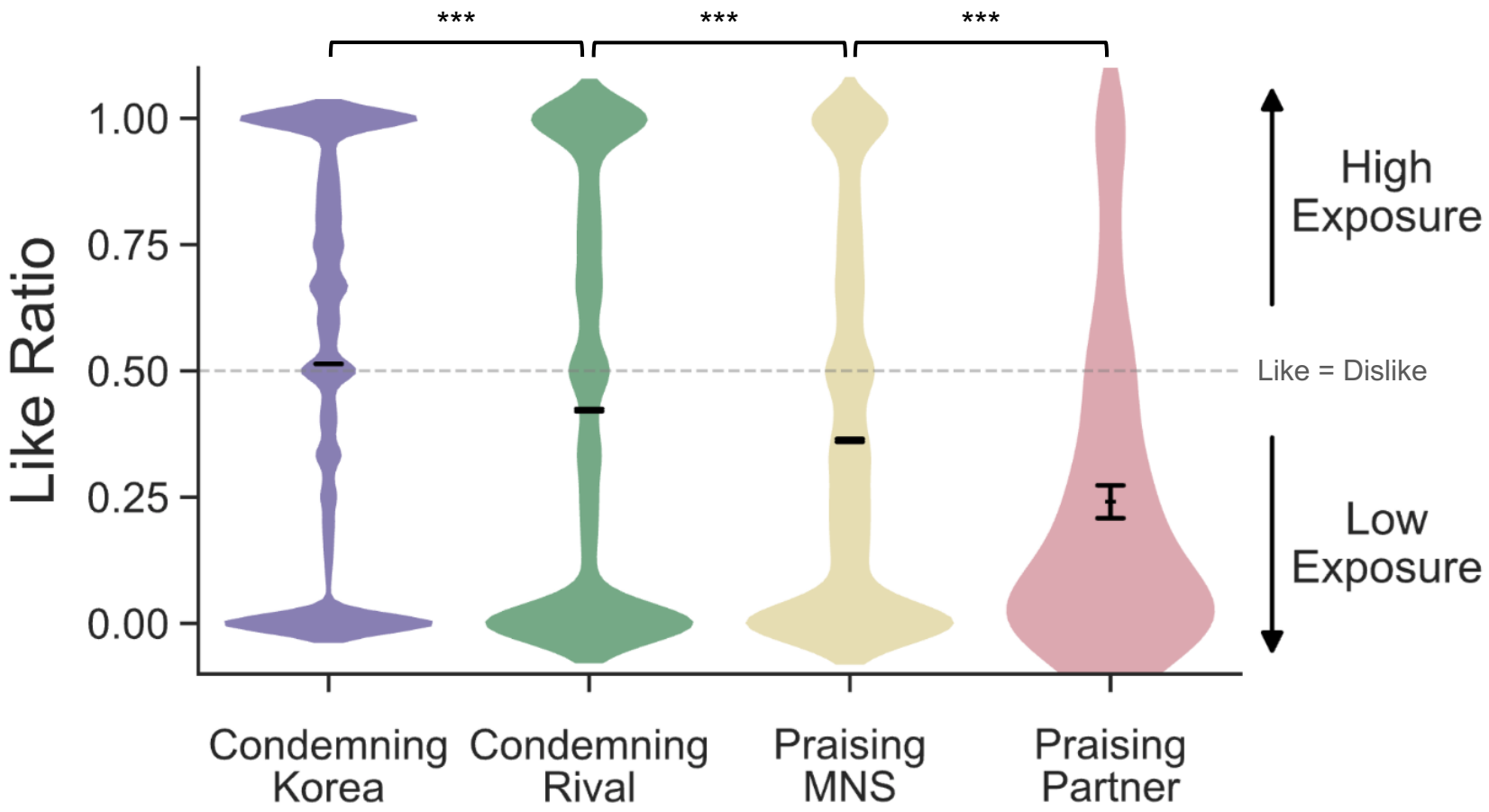
A 20-year longitudinal view: condemning rhetoric dominates

- Condemning the target country **dominates**, not promoting their own state.
- Holds under **every administration**, and the gap **widens** in recent years.
- Entry of suspected troll accounts rises **51%** around major elections.



Exploiting outrage: condemnation gains visibility

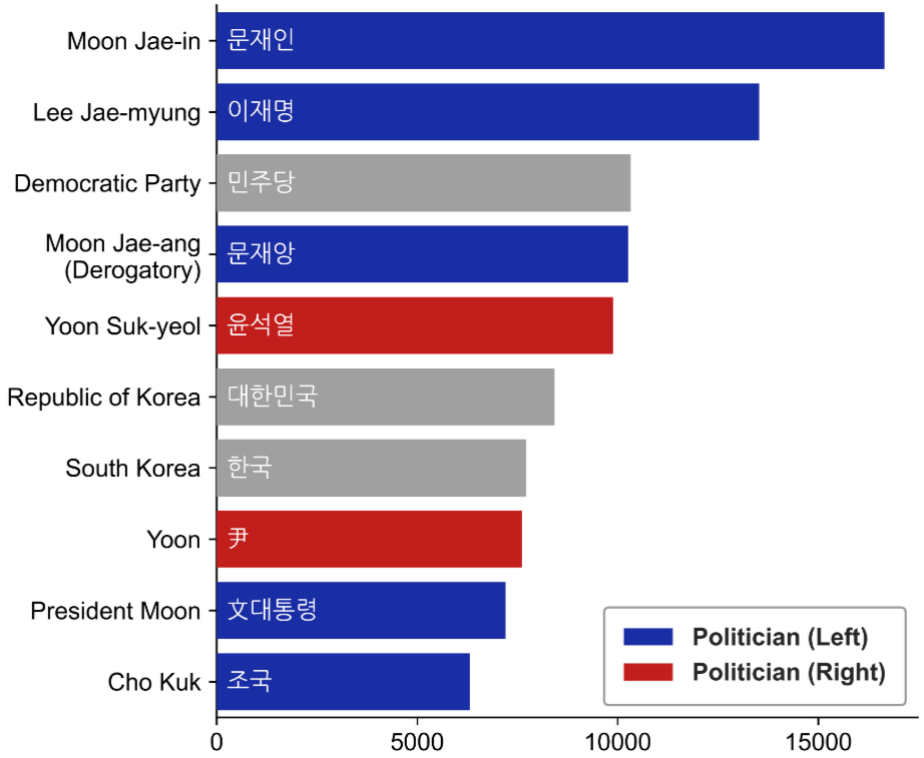
- Only **Condemning Korea** receives more likes than dislikes
- **Praising** narratives receive lower visibility.
- **Stronger condemnation** predicts greater visibility.



They condemn both sides: a pattern consistent with polarization

- Seven of the top ten targets are **political figures from across the political spectrum.**
- **The sitting president** ranks among the top targets **under all five administrations.**

→ The pattern points to **amplifying division**, not backing one side.



Top targets **over 20 years**

Roh **Moo-hyun** 2006 – 2008

Lee **Myung-bak** 2008 – 2013

Park **Geun-hye** 2013 – 2017

Moon **Jae-in** 2017 – 2022

Yoon **Suk-yeol** 2022 – 2025

Top targets **by administration**

From explainable detection to actionable insight

1. Explainable detection supports **evidence-based moderation**.
2. Defensive prioritization: **when and what to watch**.
3. Suspected accounts shared with **stakeholders** for verification.

Two-Decade Dataset
112M comments and models

luke.4.18@kaist.ac.kr