Detecting and Understanding Information Disorder

Lin Ai

Senior Applied Scientist, Microsoft
PhD work with Prof. Hirschberg, Columbia University

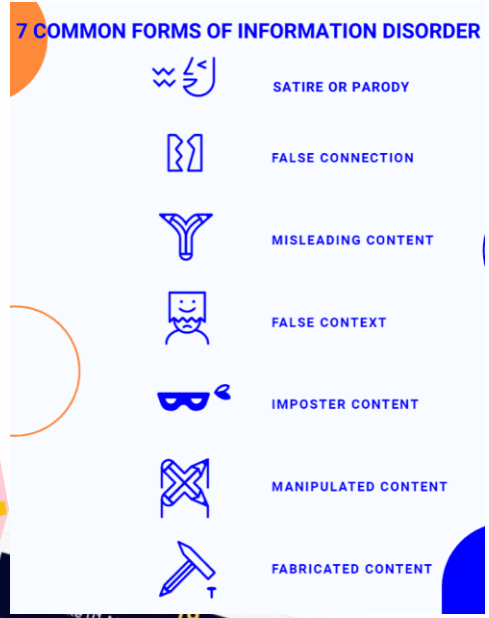
What is information disorder?

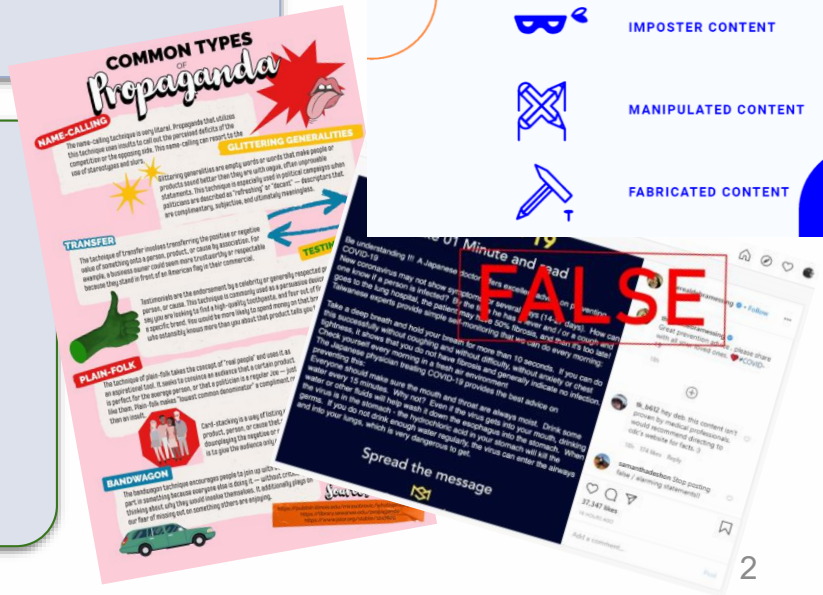
Content types

- Misinformation • Disinformation • Propaganda
- Manipulative persuasion

How it looks

- False/unsupported claims
- Context stripping & selective framing
- Fabricated/edited media
- etc. …

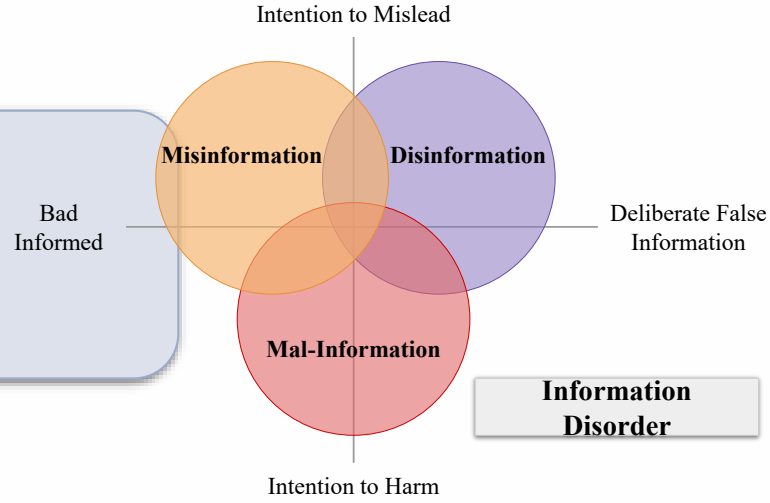




What is information disorder?

Intent dimension

- Intent to Mislead • Intent to harm



How it works

- Emotional appeals, authority mimicry, etc
- Agenda-driven narratives
- Target specific tactics

Who uses it—and who it targets?

Actors

- State-affiliated media & campaigns
- Influencers & politicians
- Issue networks & for-profit operations
- Emerging: LLM agents

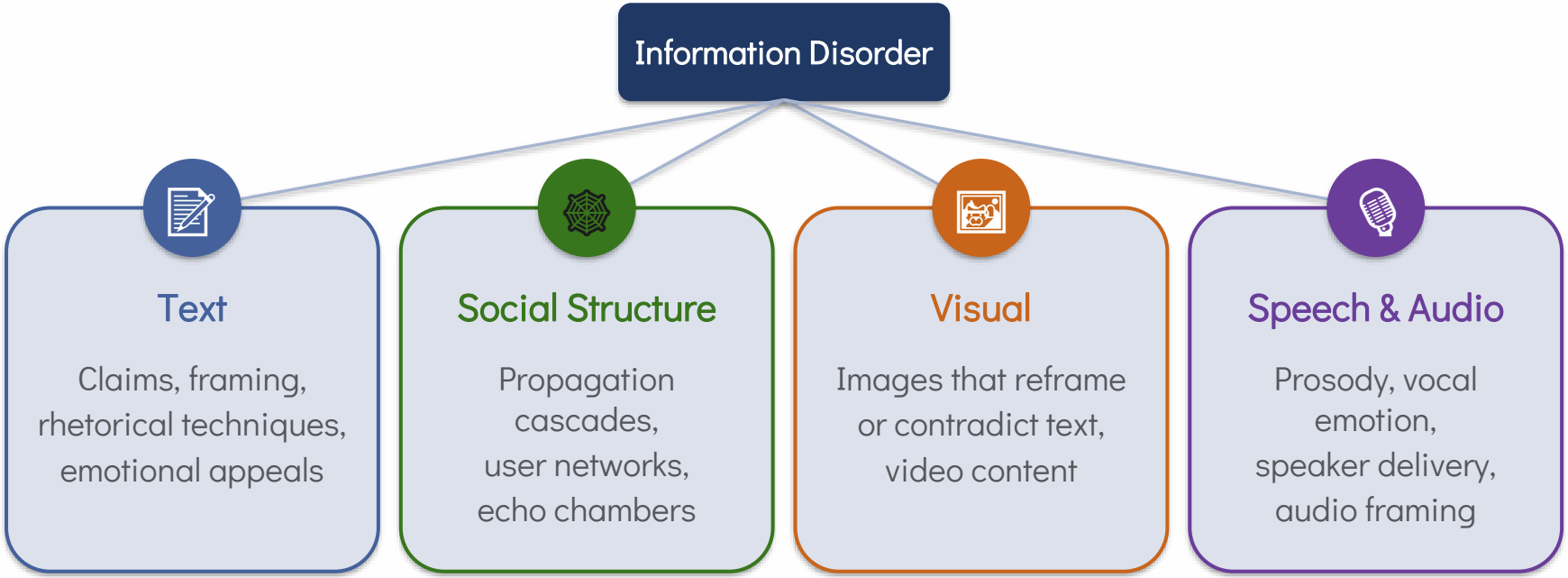
Audiences

- General public segments
- High-susceptibility cohorts
- Issue-aligned communities
- Decision makers

Intents

- Mislead or Deceive
- Propaganda
- Radicalize
- etc. ...

Information disorder leaves traces across modalities

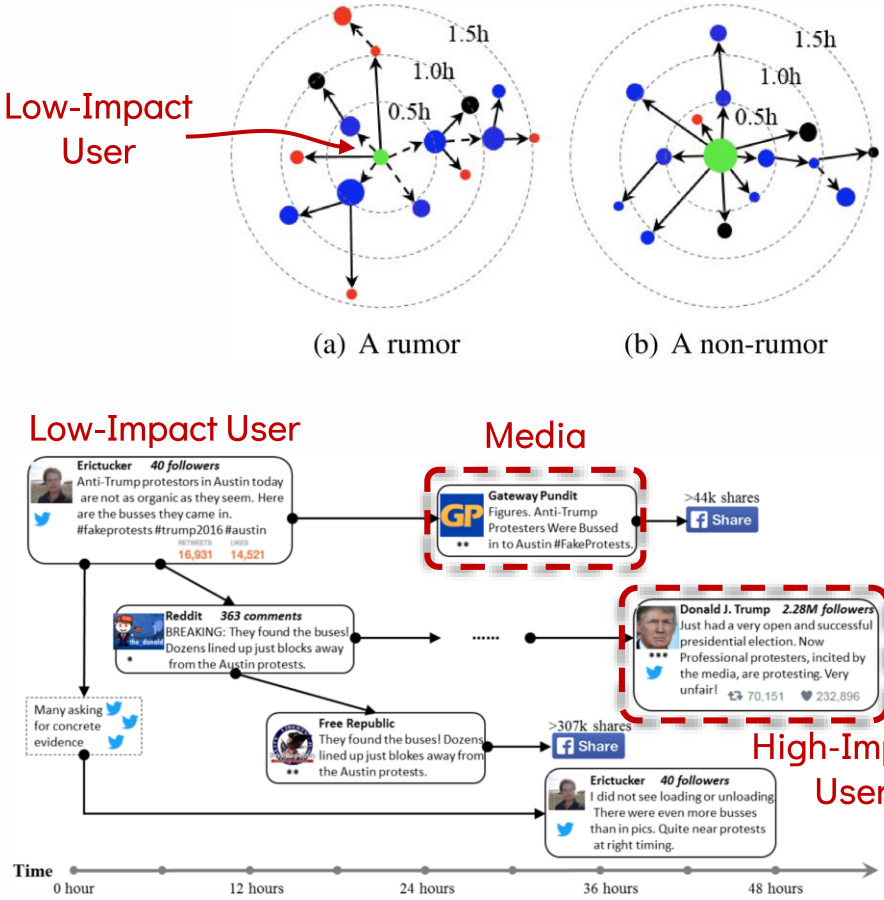


Today's talk draws on all four — including speech prosody and vocal emotion in the audience-perception study

Background

- **Early work** on information disorder detection rely primarily on **linguistic features**
- **Propagation-based approaches** make use of tree-structured propagation patterns of microblog posts and learn contextual representations

→ We tackle the challenge by incorporating linguistic, social context, and tweet propagation patterns



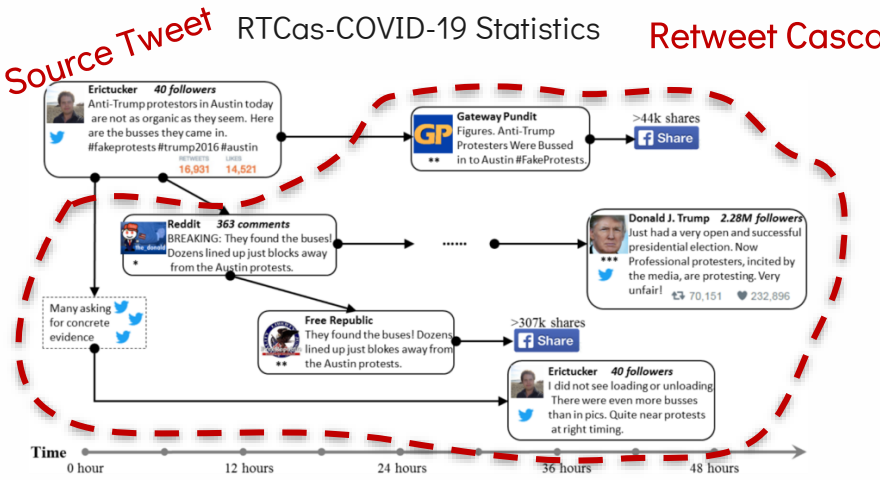
RTCas-COVID-19 Corpus

- We collect and clean a COVID-19 corpus, **RTCas-COVID-19**, using two publicly available datasets
- Cleaned corpus: **35M tweets** (10M source tweets and retweet cascades)
- **Retweet Cascade:** Tree-like structure representing how tweets spread through retweets and replies
- We weak-label a subset of tweets using URLs source credibility
  - Weak-labeled corpus: **2M tweets**

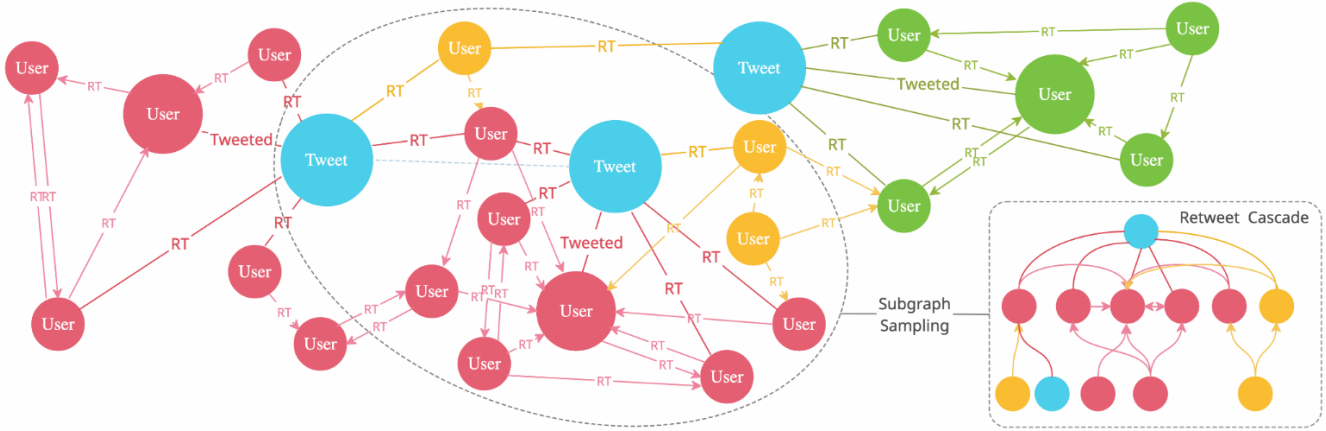
	Total	Source	Retweets
Full Corpus	35M	10M	25M

	Total	Trust	Untrust
Weak-Labeled	2M	1.64M	360K
Human-Annotated	380	215	165

RTCas-COVID-19 Statistics

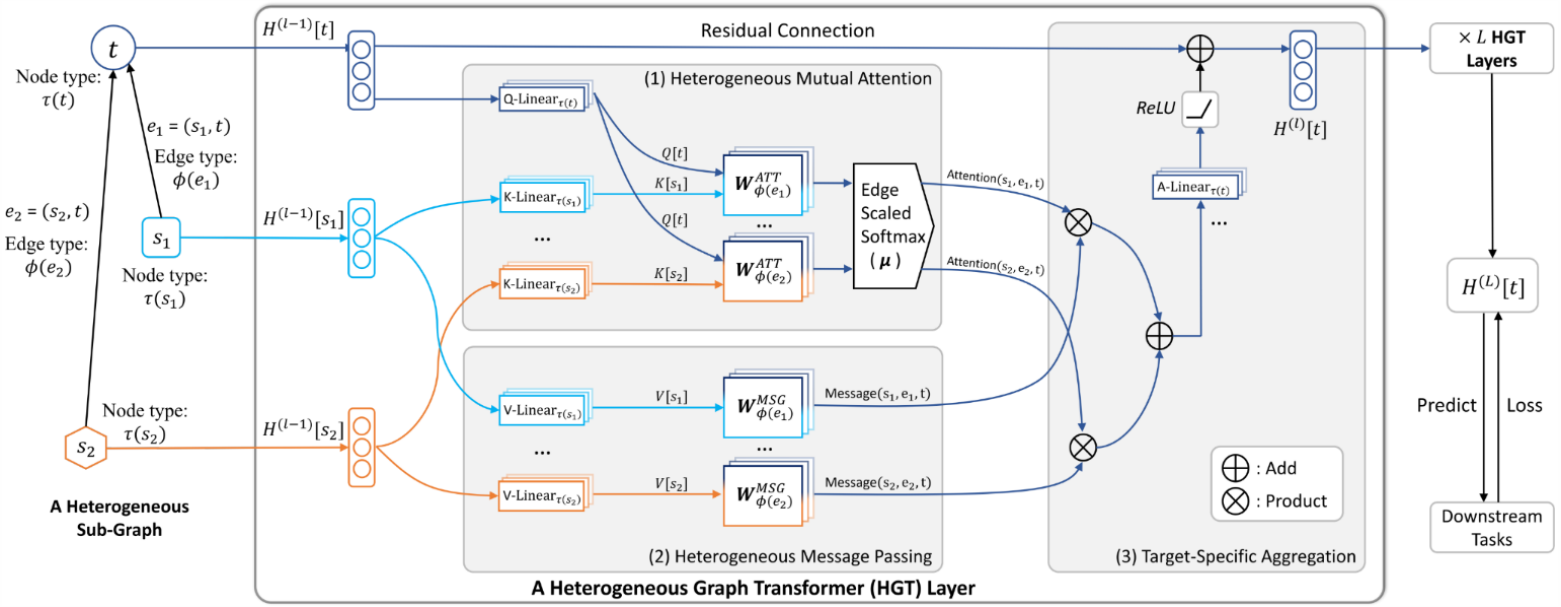


Tweet-User Heterogeneous Graph



- **Echo chamber:** users tend to interact more with like-minded people
- Nodes are connected by "tweeting" and "retweeting"

RTCS-HGT (Retweet Cascade Subgraph Sampling Heterogeneous Graph Transformer)



We apply an inductive **heterogeneous graph transformer (HGT)** for node representation learning

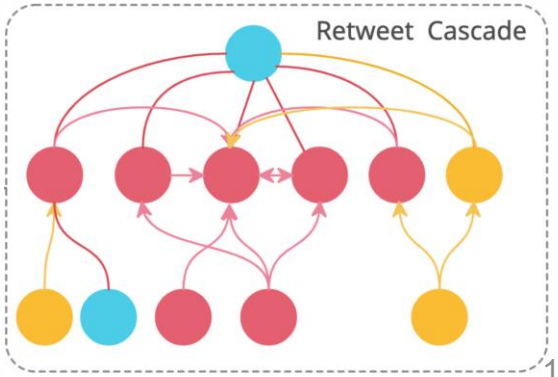
RTCS-HGT (Retweet Cascade Subgraph Sampling Heterogeneous Graph Transformer)

- **Supervised Tweet Classification Loss:** A negative log-likelihood classification loss
- **Unsupervised User Proximity Loss:**

$$proxloss = \mathbb{E}[-\mathbb{E}\log\sigma(z_u^\top z_{v_p}) - Q \cdot \mathbb{E}\log\sigma(-z_u^\top z_{v_n})]$$

where P_u and N_u represent neighboring and non-neighboring nodes of user u, respectively, and z_v denotes the node representation

- To scale the model to large graphs and reduce training time, we introduce a tweet-centered **retweet cascade subgraph sampling (RTCS)** approach



Experiments

- RTCS-HGT model outperforms all baselines on the weak-labeled RTCas-COVID-19 subset
- CT-BERT model performs comparably, it is limited to COVID-19 tweets and requires re-training for other topics
- HGATRD is a strong baseline but lacks scalability and cannot infer on unseen data

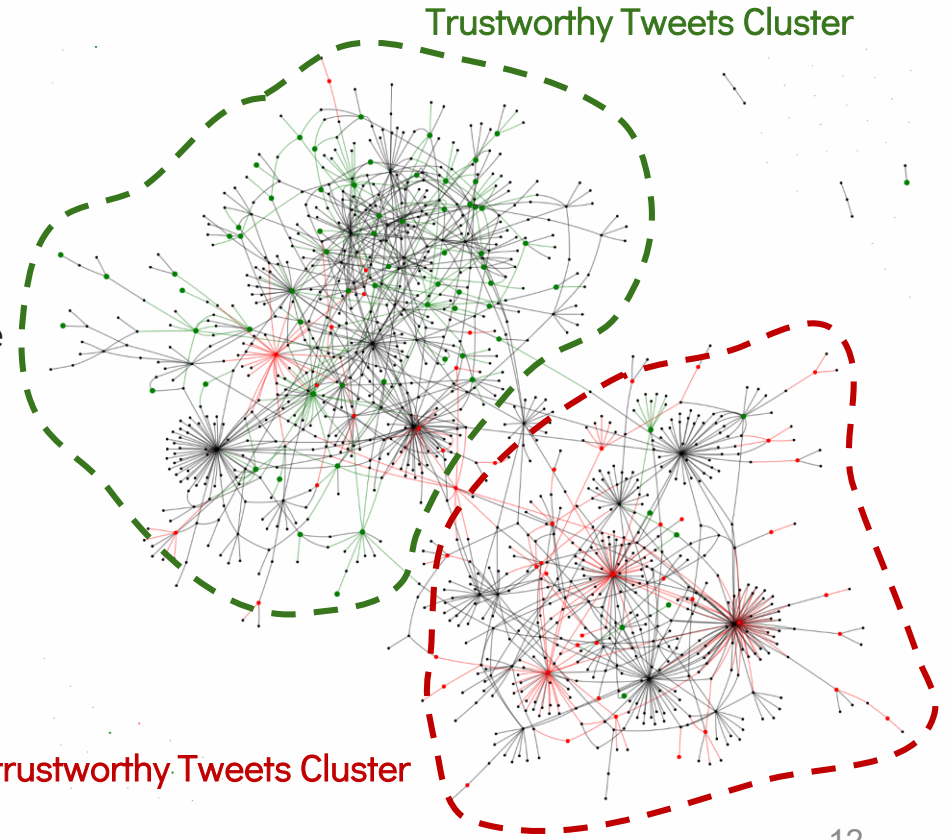
Model	Test Acc.	Macro F_1	Trust F_1	Untrust F_1
RCNN-LSTM	0.844	0.844	0.846	0.842
RCNN-GRU	0.844	0.843	0.848	0.839
BERTweet	0.847	0.847	0.850	0.843
CT-BERT	0.893	0.893	0.894	0.893
HGATRD	0.894	0.894	0.894	0.895
HGT	0.908	0.908	0.910	0.906
RTCS-HGT	0.913	0.913	0.913	0.912
RTCS-HGT (no *proxloss*)	**0.918**	**0.918**	**0.918**	**0.918**

RTCS-HGT vs. Baselines on Weak-Labeled RTCas-COVID-19 Test Sets

Analysis – Social Context Information

- Nodes naturally form 2 clusters
- Users within each cluster tend to interact more frequently with other users in the same cluster rather than with users from the other cluster

Echo Chamber Effect



Multimodal Intent Detection in Social Media -- Background

- While previous framework detects untrustworthy information using linguistic and social context, **visual content** plays a key role in social media posts

Happy Earth 🌎 Day! #EarthDay #HappyEarthDay
#ClimateChange #GlobalWarming

BENIGN

Happy Earth 🌎 Day! #EarthDay #HappyEarthDay
#ClimateChange #GlobalWarming

MALICIOUS

Multimodal Intent Detection in Social Media -- Corpus

- We collect a multi-modal social media corpus of approximately 13K posts
  - Text + image posts from Twitter and Facebook
- Topics: COVID-19, Climate Change
- Each post is weak-labeled as malicious/benign with 5 fine-grained intent categories: *polarization*, *call-to-action*, *virality*, *sarcasm*, *humor*
  - Posts are weak-labeled using hashtags
  - Viral posts are those among the top 1,500 most retweeted

	Data →	Covid-19	Climate Change
Malicious	Polarizing	1938	1725
	Call-To-Action	2454	-
	Viral	1500	56
	Total	*5892*	*1781*
Benign	Sarcasm	2361	323
	Humor	1500	1497
	Total	*3861*	*1820*
	Total	***9753***	***3601***

Detailed statistics of the 13K posts

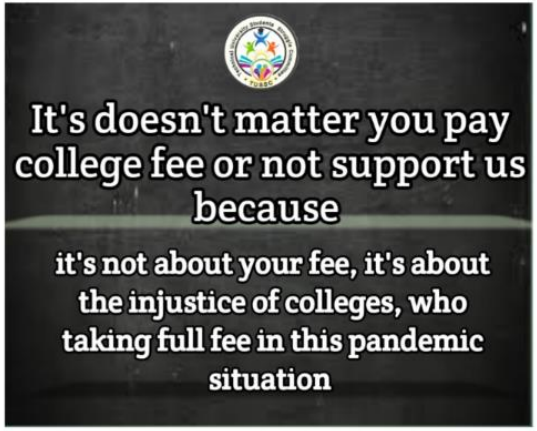
Multimodal Intent Detection in Social Media -- Corpus

⚠ **Content Disclaimer:** *examples shown contain harmful/misleading content involving real public figures and do **NOT** reflect the views of the presenter or co-authors, nor any factual claims about the individuals depicted.*

Pro vaxxers are always demanding proof then dismiss it when you give it to them. They say vaccines have been “proven” safe but when asked to show said “proof” they give 🤙🤙🤙🤙🤙 Then they bow down and worship their lord and savior #BillGatesBioTerrorist #sad

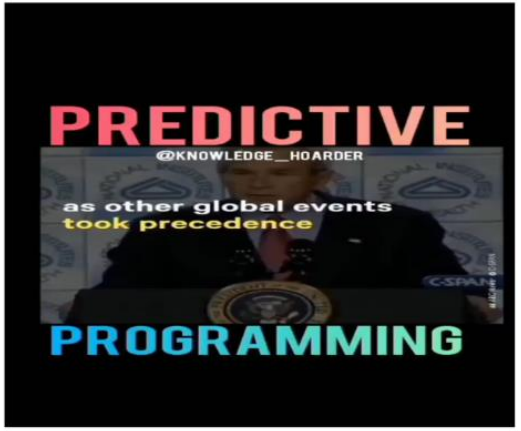
Polarizing

Support us on Twitter campaign Speakup against fee hike in colleges #tussc #ReduceCollegeFees #StudentProblems #Standup #COVID19 #supportus #ThursdayThoughts



Call-to-Action

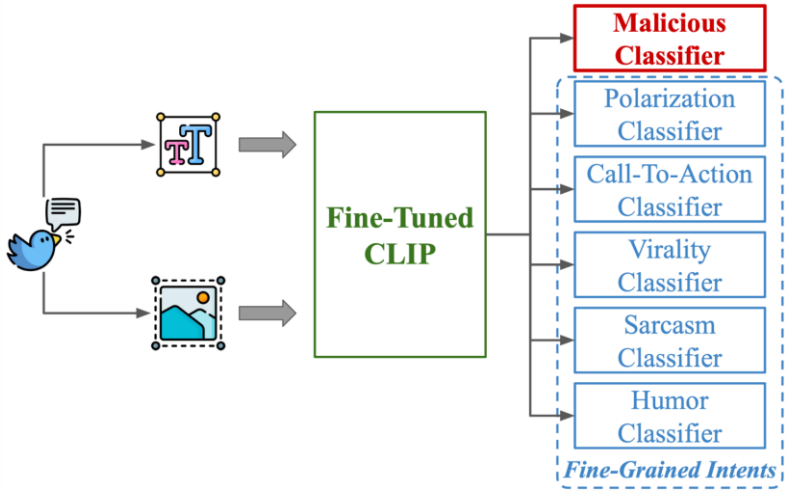
You do not understand what a global reset is.....it is a plan to kill you and your family and your generations to come..... It is a depopulation plan... It is a global slavery and surveillance plan.... It is a plan to create a one world government.... Covid was only a trigger..



Viral

CLIP-MTL

- We fine-tune CLIP's pre-trained text and image encoders to extract linguistic and visual features
- We train the model using **multi-task learning (MTL)**, with shared CLIP weights and separated intent classifiers
- Each task benefits from the shared knowledge of the others



Experiments and Analysis

- The model achieves a test accuracy of 0.978 on the maliciousness classification task
- It attains an average test accuracy of 0.977 across the five intent classifiers

➡ Future improvements include addressing potential overfitting and enhancing generalizability

Task ↓	Accuracy	Neg Macro F1	Pos Macro F1
Polarization	0.970	0.981	0.932
Call-To-Action	0.979	0.988	0.923
Virality	0.985	0.992	0.914
Sarcasm	0.983	0.990	0.934
Humor	0.966	0.979	0.905
Maliciousness	**0.978**	**0.979**	**0.977**

Performance on malicious classification and each specific intent classification

Warning from Heterogeneous Signals

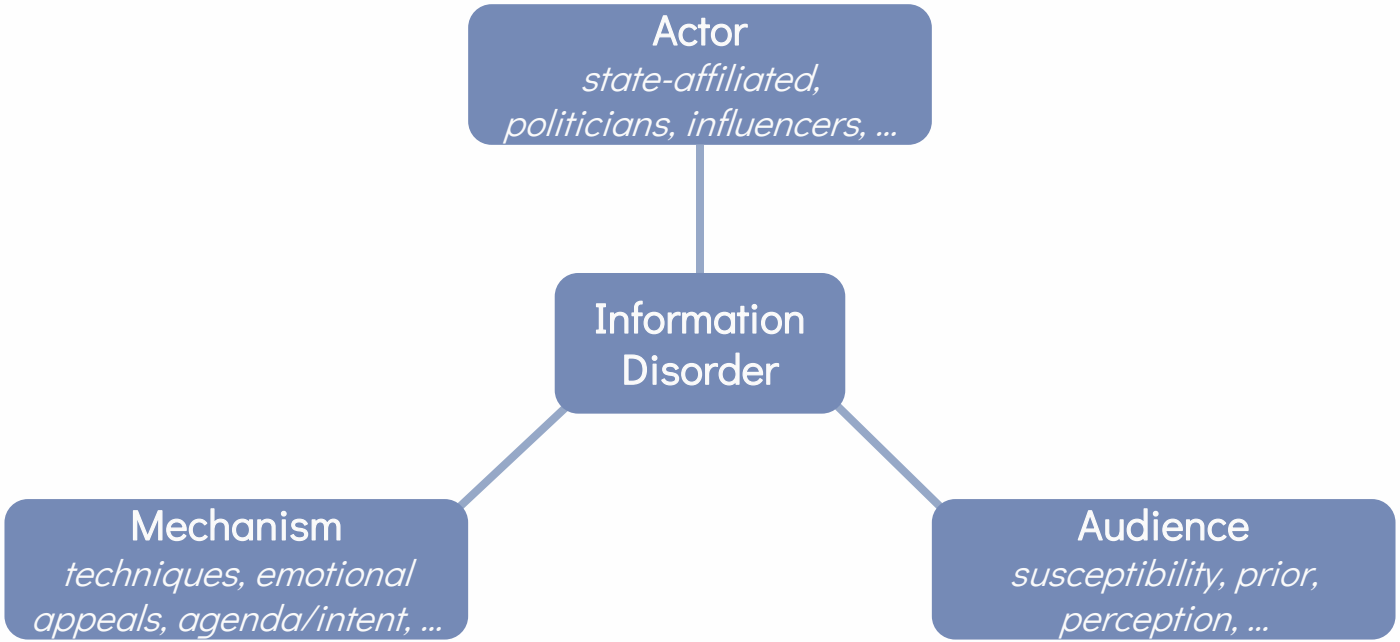
- Combining textual cues with propagation lifts early detection and reveals community patterns missed by text only
- Visual features aid to uncover malicious narrative intent
- High-recall first line of defense at scale

Findings / Takeaways

- Detection ≠ persuasion model – we still don't know why the content works
- We don't identify who is most susceptible (audience traits, perceptions)
- Labels hint at intent, but we need actor-level narratives and technique/appeal structure

Gaps → Next

Fine-Grained Understand



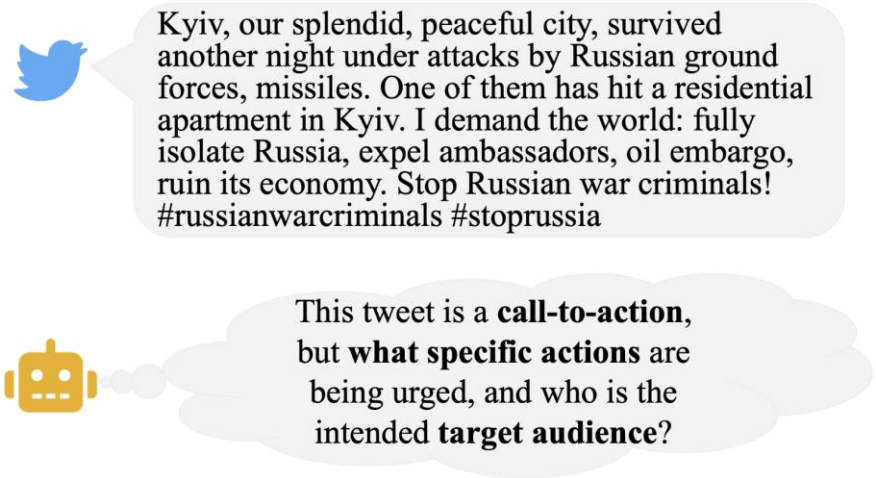
Manipulative Narrative of State Actors

Actor
state-affiliated, politicians, influencers, ...

- We dive deeper into **why** and **how** specific content is spread to convey actors' messages

- Researchers have also focused on analyzing **propaganda** campaigns, notably from Russia

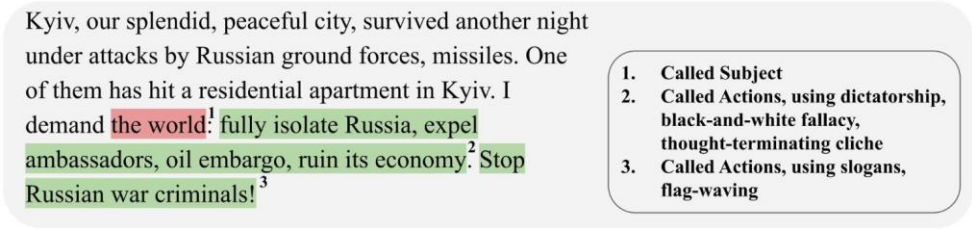
→ Our work aims to dive deeper into this, focusing on narratives from **both Russia and Ukraine**



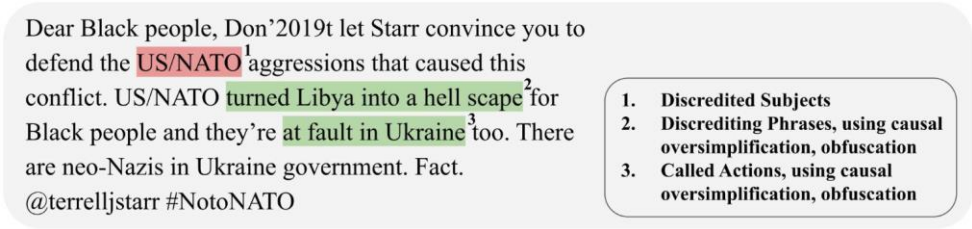
⚠ **Research Disclaimer:** *this work analyzes state-affiliated messaging from both Russian and Ukrainian government accounts as a computational case study. It does **NOT** take a political position on the conflict, endorse any actor's narrative, or equate the actions of the parties involved.*

Fine-Grained Intent

- Call To Action (*CTA*)
  - *Called Subjects:* the intended audience
  - *Called Actions:* specific actions
- Discredit Entity (*DE*)
  - *Discredited Subjects:* the entities being undermined
  - *Discrediting Phrases:* specific language used to discredit these entities



Example Tweet Annotated as ***CTA***



Example Tweet Annotated as ***DE***

Ai, L., et al. *TweetIntent@Crisis: A Dataset Revealing Narratives of Both Sides in the Russia-Ukraine Crisis.* ***AAAI ICWSM 2024***

TweetIntent@Crisis

Collect
- 67 RU / 12 UA gov–affiliated Twitter accounts

Seed human labels
- 5,000 pre-filtered by GPT-4 → 3,691 valid
- CTA: 93 (≈1.9%) DE: 411 (≈8.8%)

Scarce Positives

Model Selection
- *(a) Intent* *(b) Interrogative span localization*
- Fine-tuned GPT-3.5-Turbo > GPT-4 / GPT-4+ICL

Machine annotation
- 14K tweets auto-labeled with FT GPT-3.5-Turbo

Full dataset #Tweets	17,854	
	Human Annotated	**Machine Annotated**
#Tweets	3,691	14,163
#CTA Tweets	93	307
#CTA Text Spans	196	537
#DE Tweets	411	767
#DE Text Spans	1,292	2,447

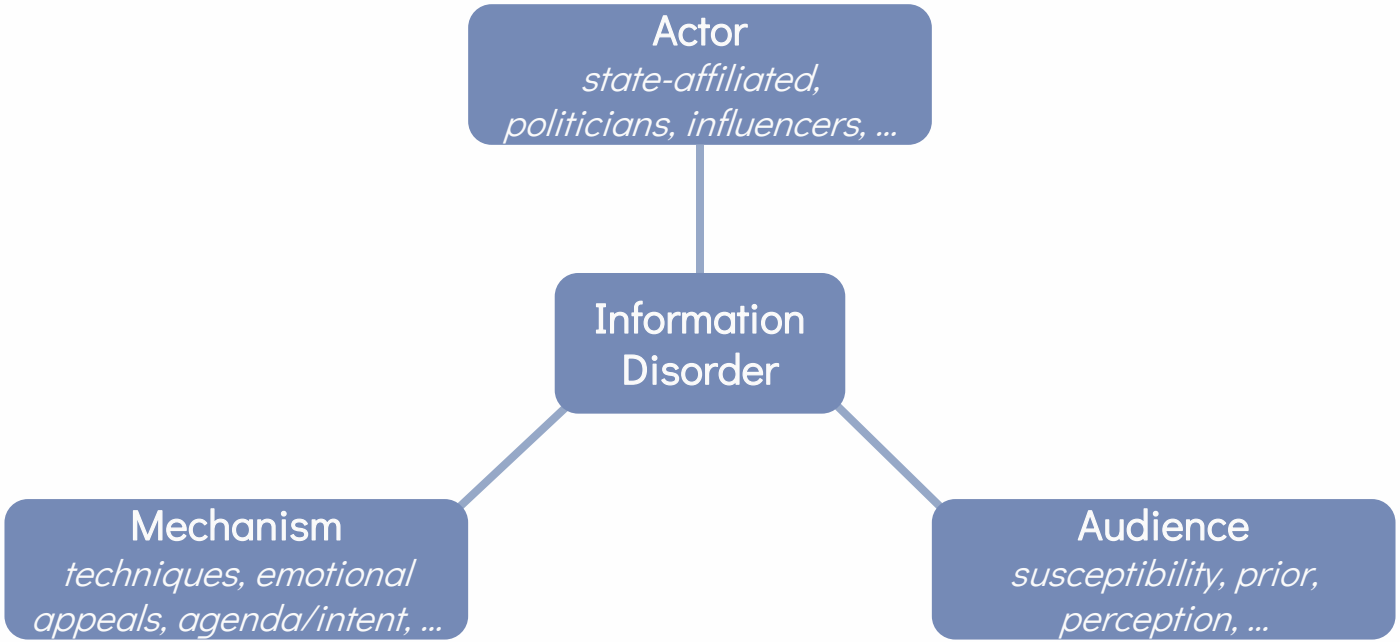
Dataset Statistics

Granular Narrative Analysis

- **TweetIntent@Crisis** dataset provides a detailed view of narratives from both Russian and Ukrainian perspectives
- Russian narratives frequently discredit Ukraine, NATO, and Western governments, while Ukrainian accounts discredit Russia for the invasion
- Ukrainian accounts emphasize rallying support, with more *CTA* tweets advocating for sanctions and military aid, often targeting international audiences like the UN

Most Frequent Discredited Subjects (Top) and Discrediting Phrases (bottom) in Tweets from Russian Accounts

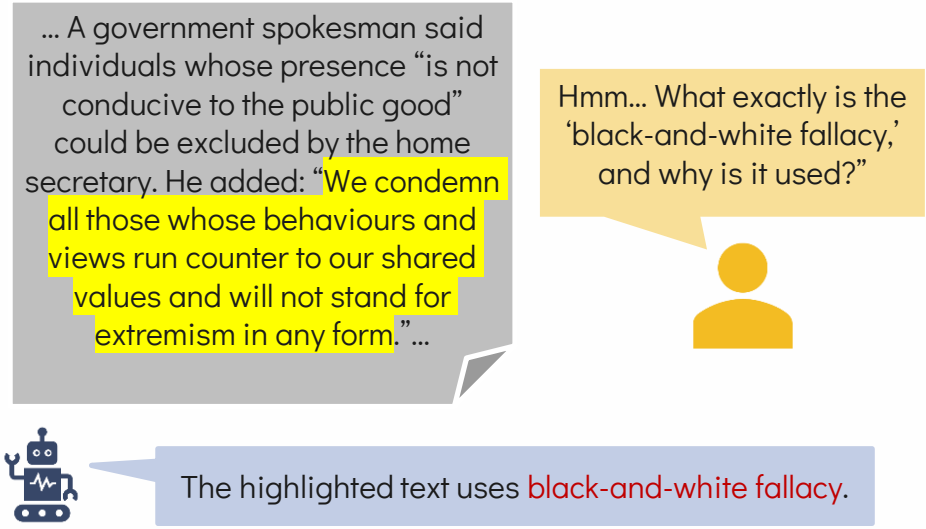
Fine-Grained Understand



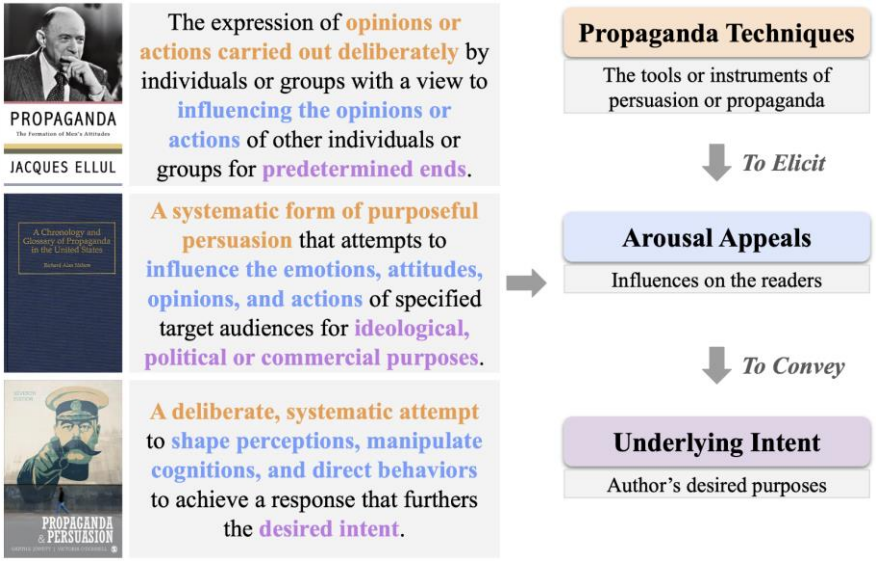
Background

- Moving forward, we propose a **generalizable framework for analyzing propaganda and the specific intent** driving each propaganda attempt

- Telling non-expert readers only what techniques are used is far from satisfying
- More research is needed to **understand the intent behind these techniques**



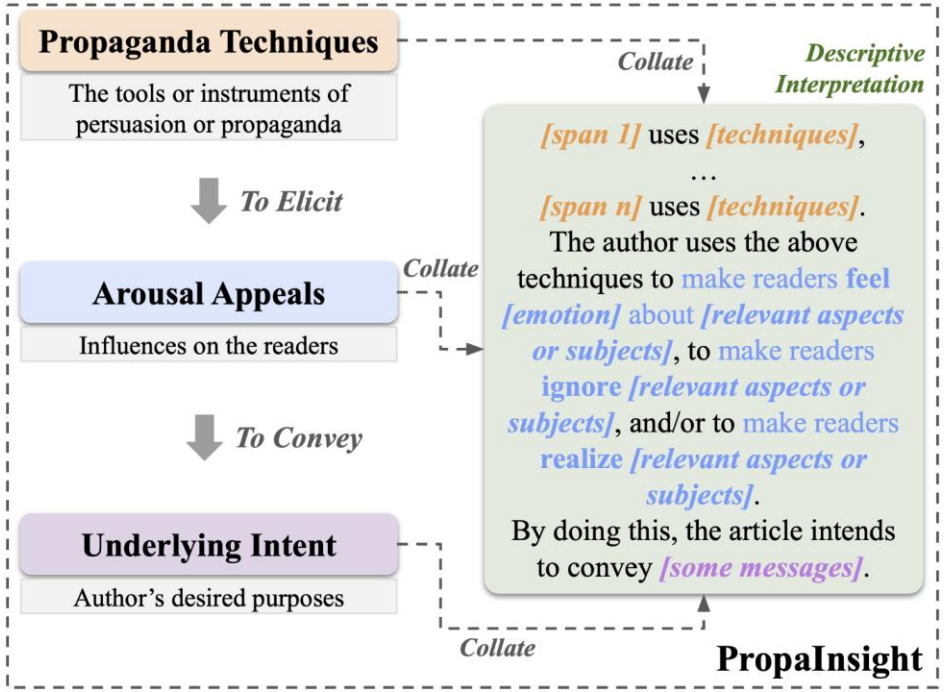
Established Research on Propaganda – Social Science



- Drawing from science research on propaganda, we identify three key elements of each propaganda attempt:
  - Propaganda Techniques (instrument)
  - Arousal Appeals (persuasion strategy)
  - Underlying Intent (purpose)
- We incorporate the elements into a *conceptual framework of propaganda analysis*

Liu, J.*, Ai, L.*, Liu, Z., Karisani, P., Hui, Z., Fung, M., ... & Ji, H. *PropaInsight: Toward Deeper Understanding of Propaganda in Terms of Techniques, Appeals, and Intent.* ***COLING 2025***

PropaInsight: Framework of Propaganda Analysis



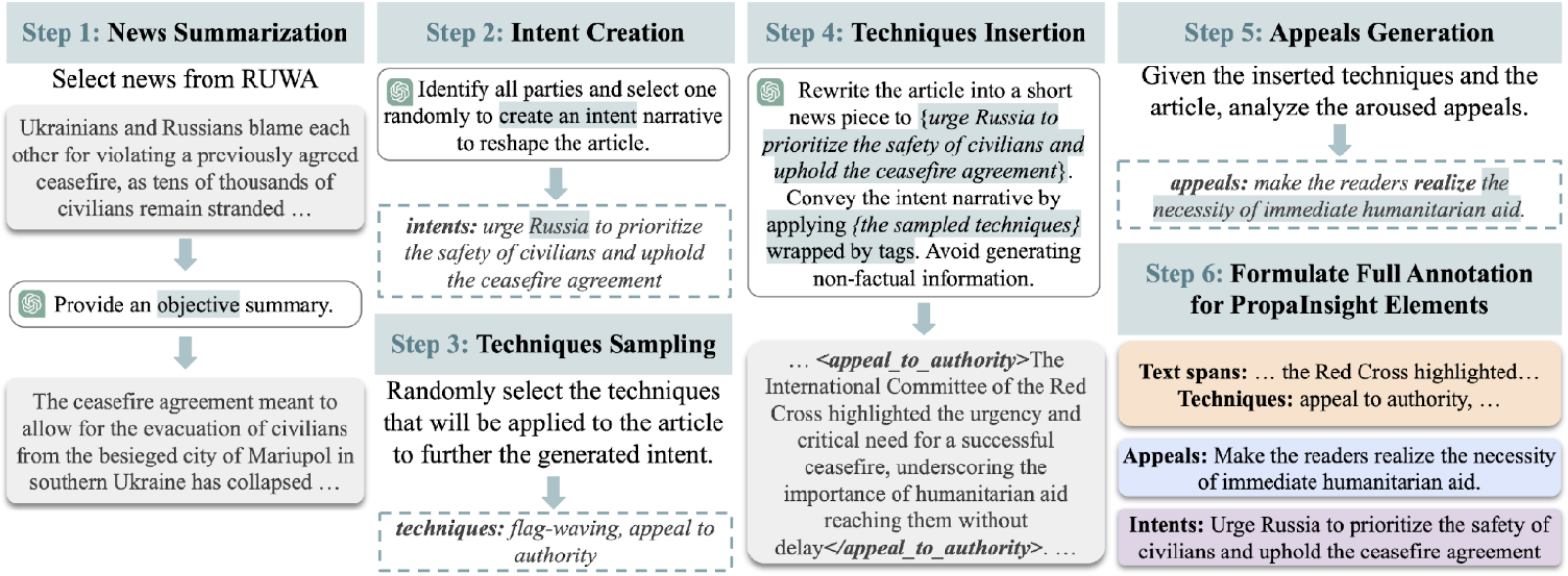
Step 1: Identify and classify techniques

Step 2: Infer evoked arousal appeals

Step 3: Deduce underlying intent

Step 4: Consolidate elements into structured natural language paragraph

PropaGaze: A Novel Dataset for Propaganda Analysis



- We construct *RUWA-Gaze* and *Politifact-Gaze* using a controlled data generation pipeline
- Rewrite articles into objective summaries, and then insert sampled propaganda techniques to shape the article's intent

Off-the-Shelf LLMs Performance

- Zero-shot LLMs struggle with propaganda analysis
- Zero-shot LLMs struggle to pinpoint techniques and to infer arousal appeals
- Models perform better at inferring intent
- Few-shot prompting improves LLM performance in analyzing propaganda elements

Data Setting ↓	Dataset → / Model ↓	*RUWA-Gaze* Span Avg. IoU	*RUWA-Gaze* Techniques Macro F1	*Politifact-Gaze* Span Avg. IoU	*Politifact-Gaze* Techniques Macro F1	*PTC-Gaze* Span Avg. IoU	*PTC-Gaze* Techniques Macro F1
No Training Data	GPT-4-Turbo$_{0s}$	0.073	0.097	0.152	0.226	0.124	0.068
	GPT-4-Turbo$_{1s}$	0.132	0.145	0.183	0.269	0.165	0.171
Data-Sparse Training	MGNN	0.089	0.139	0.160	0.159	0.140	**0.206**
	Llama-7B-Chat$_{ft}$	0.230	0.210	0.253	0.281	**0.179**	0.191
Data-Rich Training	MGNN	**0.545**	0.591	**0.449**	**0.461**	-	-
	Llama-7B-Chat$_{ft}$	0.506	**0.607**	0.409	0.453	-	-

Model Performance on *Propaganda Technique Identification* Subtask

Dataset → / Model ↓	*RUWA-Gaze* Appeals BertScore	*RUWA-Gaze* Intents BertScore	*Politifact-Gaze* Appeals BertScore	*Politifact-Gaze* Intents BertScore	*PTC-Gaze* Appeals BertScore	*PTC-Gaze* Intents BertScore
GPT-4-Turbo$_{0s}$	0.282	0.849	0.298	0.863	0.228	0.869
GPT-4-Turbo$_{1s}$	0.324	**0.879**	0.345	**0.875**	**0.331**	**0.881**
Llama-7B-Chat$_{ft}$ *(Data-Sparse)*	0.313	0.851	0.342	0.860	0.249	0.843
Llama-7B-Chat$_{ft}$ *(Data-Rich)*	**0.612**	0.861	**0.495**	0.864	-	-

Model Performance on *Appeal Analysis* and *Intent Analysis* Subtasks

PropaGaze Effectiveness

- Models: Llama-7B-Chat and Multi-Granularity Neural Networks (MGNN)
- PropaGaze substantially improves the overall propaganda analysis performance

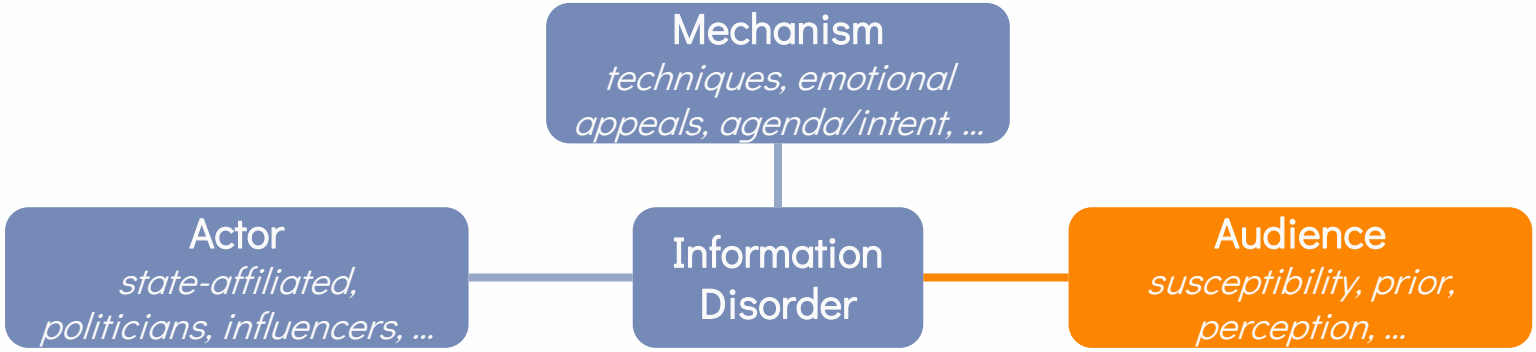
Data Setting ↓	Dataset → / Model ↓	*RUWA-Gaze* Span Avg. IoU	*RUWA-Gaze* Techniques Macro F1	*Politifact-Gaze* Span Avg. IoU	*Politifact-Gaze* Techniques Macro F1	*PTC-Gaze* Span Avg. IoU	*PTC-Gaze* Techniques Macro F1
No Training Data	GPT-4-Turbo$_{0s}$	0.073	0.097	0.152	0.226	0.124	0.068
	GPT-4-Turbo$_{1s}$	0.132	0.145	0.183	0.269	0.165	0.171
Data-Sparse Training	MGNN	0.089	0.139	0.160	0.159	0.140	**0.206**
	Llama-7B-Chat$_{ft}$	0.230	0.210	0.253	0.281	**0.179**	0.191
Data-Rich Training	MGNN	**0.545**	0.591	**0.449**	**0.461**	-	-
	Llama-7B-Chat$_{ft}$	0.506	**0.607**	0.409	0.453	-	-

Model Performance on *Propaganda Technique Identification* Subtask

Dataset → / Model ↓	*RUWA-Gaze* Appeals BertScore	*RUWA-Gaze* Intents BertScore	*Politifact-Gaze* Appeals BertScore	*Politifact-Gaze* Intents BertScore	*PTC-Gaze* Appeals BertScore	*PTC-Gaze* Intents BertScore
GPT-4-Turbo$_{0s}$	0.282	0.849	0.298	0.863	0.228	0.869
GPT-4-Turbo$_{1s}$	0.324	**0.879**	0.345	**0.875**	**0.331**	**0.881**
Llama-7B-Chat$_{ft}$ *(Data-Sparse)*	0.313	0.851	0.342	0.860	0.249	0.843
Llama-7B-Chat$_{ft}$ *(Data-Rich)*	**0.612**	0.861	**0.495**	0.864	-	-

Model Performance on *Appeal Analysis* and *Intent Analysis* Subtasks

Multimodal Cues in Radicalization: An Audience-Centric Study



1. **RQ1:** What viewer traits, such as personality traits and media consumption, are associated with their video preferences?
2. **RQ2:** What video characteristics, such as speaker traits, video quality, and arousing emotions, are correlated with viewers' perception?
3. **RQ3:** Which modality features affect viewers' perception the most?

Ai, Lin, et al. *What Makes A Video Radicalizing? Identifying Sources of Influence in QAnon Videos.* ***IC²S² 2023***

Multimodal Cues in Radicalization: An Audience-Centric Study

⚠ **Content Disclaimer:** *examples shown contain harmful/misleading content and do **NOT** reflect the views of the presenter or co-authors.*

Study at a glance

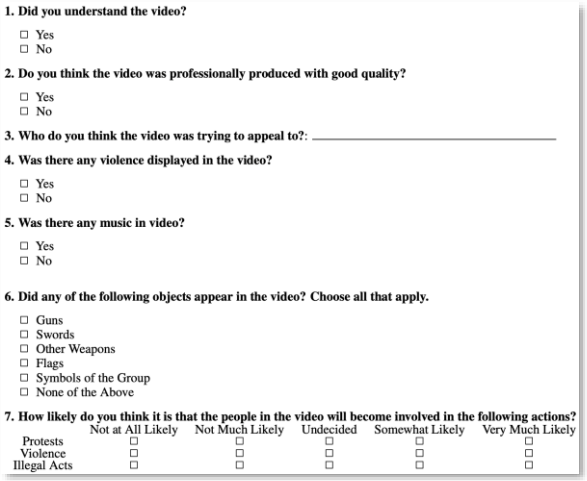
- 6 exemplar videos on QAnon
- 46 participants (questionnaire)
- **Metrics:** Enjoyment, Content, Actions

- We focus on **QAnon**, a conspiracy-driven radicalization group
- We select 6 videos based on relevance, duration, diversity of styles, content quality, and popularity

Multimodal Cues in Radicalization: An Audience-Centric Study

Study at a glance

- 6 exemplar videos on QAnon
- 46 participants (questionnaire)
- **Metrics:** Enjoyment, Content, Actions



1. Did you understand the video?

☐ Yes
☐ No

2. Do you think the video was professionally produced with good quality?

☐ Yes
☐ No

3. Who do you think the video was trying to appeal to?: _____

4. Was there any violence displayed in the video?

☐ Yes
☐ No

5. Was there any music in video?

☐ Yes
☐ No

6. Did any of the following objects appear in the video? Choose all that apply.

☐ Guns
☐ Swords
☐ Other Weapons
☐ Flags
☐ Symbols of the Group
☐ None of the Above

7. How likely do you think it is that the people in the video will become involved in the following actions?

	Not at All Likely	Not Much Likely	Undecided	Somewhat Likely	Very Much Likely
Protests	☐	☐	☐	☐	☐
Violence	☐	☐	☐	☐	☐
Illegal Acts	☐	☐	☐	☐	☐

Enjoyment Score

- How much viewers enjoy watching each video

Content Score

- Whether viewers think a video is persuasive, trustworthy, logical, and professionally created

Actions Score

- posting a criticizing comment [score -2]
- disliking the video [score -1]
- liking the video [score 1]
- posting a supporting comment [score 2]
- considering joining the group [score 3]

Multimodal Cues in Radicalization: An Audience-Centric Study

CBS NEWS · abc · FOX NEWS channel · The New York Times · NBC · FOX · The Washington Post · msnbc · CNN

Enjoyment on Pro-QAnon Videos			Enjoyment on Anti-QAnon Videos		
Feature	Corr	p-value	Feature	Corr	p-value
Opinion_CNN	0.329	2.55E-02	Opinion_Antifa	0.368	1.19E-02
Opinion_WSJ	0.298	4.40E-02			
Content of Pro-QAnon Videos			**Content of Anti-QAnon Videos**		
Feature	Corr	p-value	Feature	Corr	p-value
Opinion_Fox	0.487	5.92E-04	Researved	0.339	2.13E-02
Opinion_NPR	-0.376	1.00E-02			
Opinion_AP	-0.33	2.53E-02			
Actions after Pro-QAnon Videos			**Actions after Anti-QAnon Videos**		
Feature	Corr	p-value	Feature	Corr	p-value
Opinion_OathKeepers	0.37	1.14E-02	Disorganized	0.318	3.12E-02
Opinion_Fox	0.358	1.45E-02	Sympathetic	-0.317	3.21E-02
Opinion_CNN	0.298	4.42E-02			

Significant viewer traits

Viewer traits → Perception/Actions

- ↑ Media trust/outlet alignment → ↑Enjoyment/Actions
- ↑ Partisan priors → ↑ Actions toward aligned videos
- ↑ "Reserved" → ↑ Content agreement (anti-videos)

Takeaway: Susceptibility is not uniform across audiences

Multimodal Cues in Radicalization: An Audience-Centric Study

- **Video characteristics:** arousal emotions, speaker characteristics of the videos, etc.
- **Textual features:** LIWC, Grievance Dictionary, VADER
- **Acoustic features** *(This is where the speech signal comes in 😉)*
  - Acoustic-prosodic features, such as pitch and intensity
  - SpeechBrain's emotion detection model
- **Visual features**
  - Pre-trained FER mode
  - Clarifai's weapon detector model

Video cues → Perception/Actions

- ↑ Perceived trust/persuasion/valid points → ↑ Actions
- ↑ Perceived trust/persuasion/valid points → ↑ Actions
- ↑ Weapons / negative facial expressions → ↓ Perception

Takeaway: Presentation/style shifts perception

Multimodal Cues in Radicalization: An Audience-Centric Study

Study at a glance

- 6 exemplar videos on QAnon
- 46 participants (questionnaire)
- **Metrics:** Enjoyment, Content, Actions

Viewer traits → Perception/Actions

- ↑ Media trust/outlet alignment → ↑Enjoyment/Actions
- ↑ Partisan priors → ↑ Actions toward aligned videos
- ↑ "Reserved" → ↑ Content agreement (anti-videos)

Takeaway: Susceptibility is not uniform across audiences

Video cues → Perception/Actions

- ↑ Perceived trust/persuasion/valid points → ↑ Actions
- ↑ Perceived trust/persuasion/valid points → ↑ Actions
- ↑ Weapons / negative facial expressions → ↓ Perception

Takeaway: Presentation/style shifts perception

Why these matter

- **Targeted interventions:** tailor counterspeech
- **Evidence selection:** down-weight high-arousal/low-trust cues; up-weight credible clips
- **Persona-aware systems:** feed signals into steerable pipelines

Conclusion: Toward a fuller picture of information disorder

ACTOR

Who & why

- RTCS-HGT: untrustworthy tweet detection via text + propagation
- TweetIntent@Crisis: narrative-level CTA / DE intents in RU–UA conflict

MECHANISM

How it persuades

- CLIP-MTL: visual cues uncover malicious intent
- PropaInsight: techniques + appeals + intent in propaganda

AUDIENCE

Who it reaches

- QAnon study: viewer traits shape susceptibility
- Multimodal cues — including speech prosody & vocal emotion — shift perception

What's next

Persona-aware, steerable interventions • Tighter integration of acoustic & visual signals for model reasoning • Real-time, scalable defense systems

Thank you — questions?