Misinformation

13 December 2024 22:02

We're living in a post truth era → disinfo has consistently stayed ahead of research

fake news was just in 2017.
Now we look at it w/ diff perspectives → fakeness, newsworthiness, shareworthiness

Memes, satire, humor ALSO influence p well.

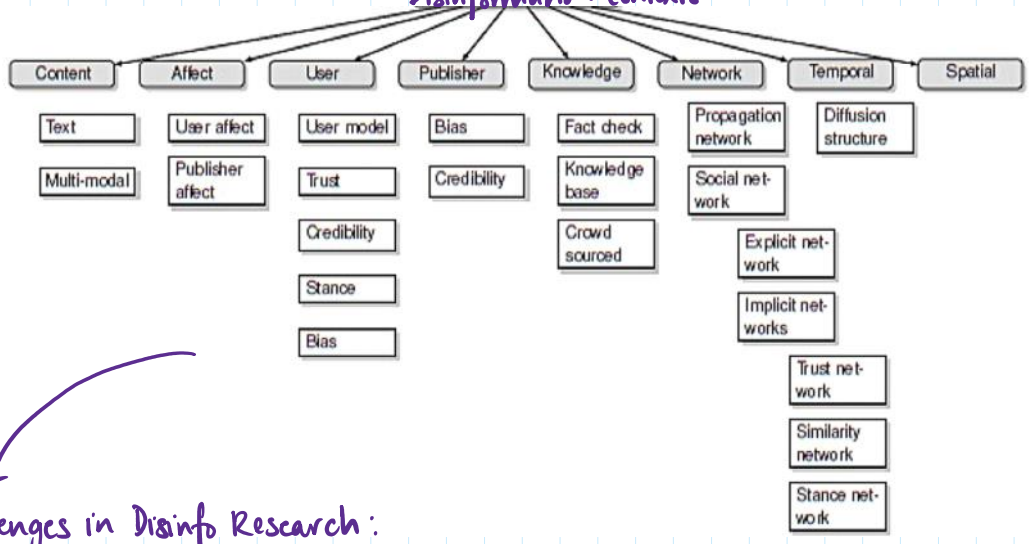
MISINFO — not to mislead (wrong/incorrect)

DISINFO — To mislead (apart/away, reversal)

MALINFO — to harm (evil)

Term	Information type	Example	Intent	Authenticity
Misinformation	False	Rumour	Not to mislead	False
Disinformation	False	Fake news	To mislead	False
Malinformation	Genuine	True news	To harm	True
Coordinated inauthentic behavior (CIB)	Can be mix of genuine and fake	Propaganda	To influence, harm and mislead	Can be mix of True and False

Disinformation Contexts.



Challenges in Disinfo Research:

1. Datasets → all context support? explainability?
2. Existing interpretability techniques
   - feature analysis
   - post hoc
   - attenⁿ + cross attⁿ
   - Sub models
   - anomaly detection
   - zero shot learning
   - fine grained reasoning
3. Early detection w/ context → content, propagation, user, knowledge, affect.
4. Weak social supervision — depends on labels derived from other models trained on limited data

④ Weak social supervision: depends on labels derived from other models trained on limited data

- User context: bias, credibility
- Affective context: emotion, cognitive words, psycholinguistic patterns
- Publisher context: bias
- Social context: homophily

Okay but how do you learn w/o labelled data?
- Unsupervised? Hell no. Too many contexts.
- Zero/few shot — works w/ typologically similar content.
- Transfer learning — catastrophic forgetting ∴ domain adaptation for disinfo?
- SSL — Deep SSL + graph based SSL.

How do you work w/ multiple contexts?
① Multiview learning
- co training
- view aggregation
- view discordance
- view ensemble
- cross view

② Multitask learning
③ Multitask multiview

LLMs + disinfo

① Weakness:
- can't reason
- factual errors, bias, + discrimination
- can't capture nuanced understanding

② Fake news detection:
- LLM makes stylistically consistent text. (models which depend on diff styles fail)
- but diff feature distributions
  → also: selection + integration of rationale — LLM cannot.

③ Rumor detection:
- not dependent on world knowledge } LLM can.
- work w/ emotion, rebuttal, support }

④ Propaganda detection:
- works w/ english. Not other langs

⑤ Dataset generation:
- adversarial prompting!

⑥ Use zero + few shot learning!
LLM for weak supervision + credibility signals.