Fake News 2.0: Combating Neural False Information

Penn State University, USA

Dongwon Lee, Ph.D.

dongwon@psu.edu

Apr. 14, 2021 @ MIND Workshop

False Information

Definitions of False Information

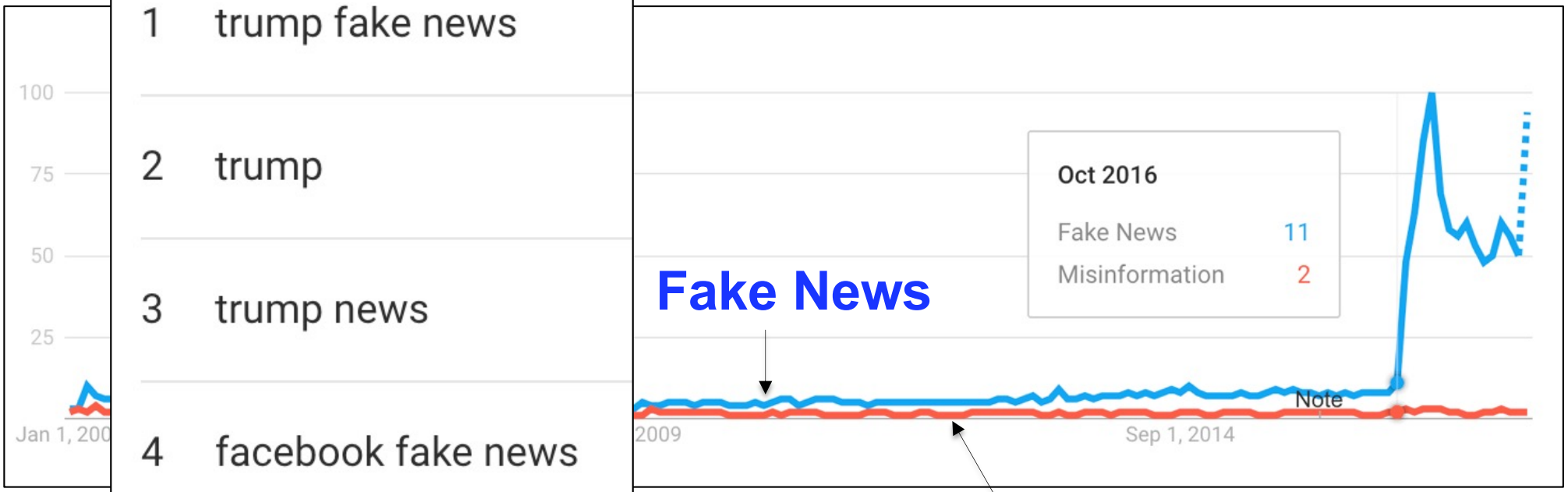
	Authenticity	Intention	News?
Fake news	False	Bad	Yes
False news	False	Unknown	Yes
Satire news	Unknown	Not bad	Yes
Disinformation	False	Bad	Unknown
Misinformation	False	Unknown	Unknown
Rumor	Unknown	Unknown	Unknown

[Zhou et al., WSDM Tutorial 2019]

Surge of “Fake News”

Related queries

1	trump fake news
2	trump
3	trump news
4	facebook fake news
5	fake news awards

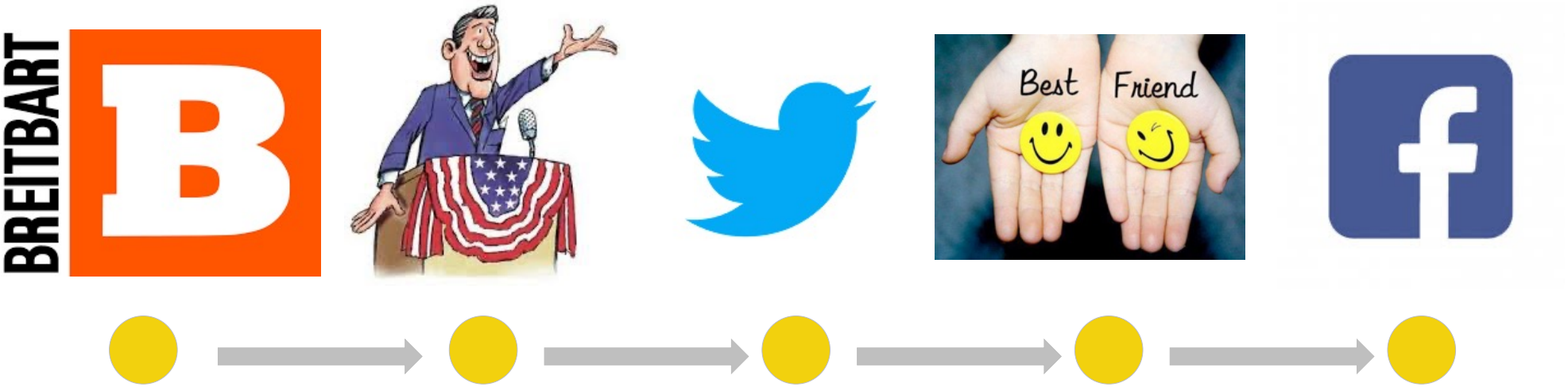


More Problems in Social Media?

1. Fundamental shift in communication: Consumer as producer [Sundar and Nass, 2001]
2. Monetary incentives: Ads by Google/Facebook

More Problems in Social Media?

3. Source Layering

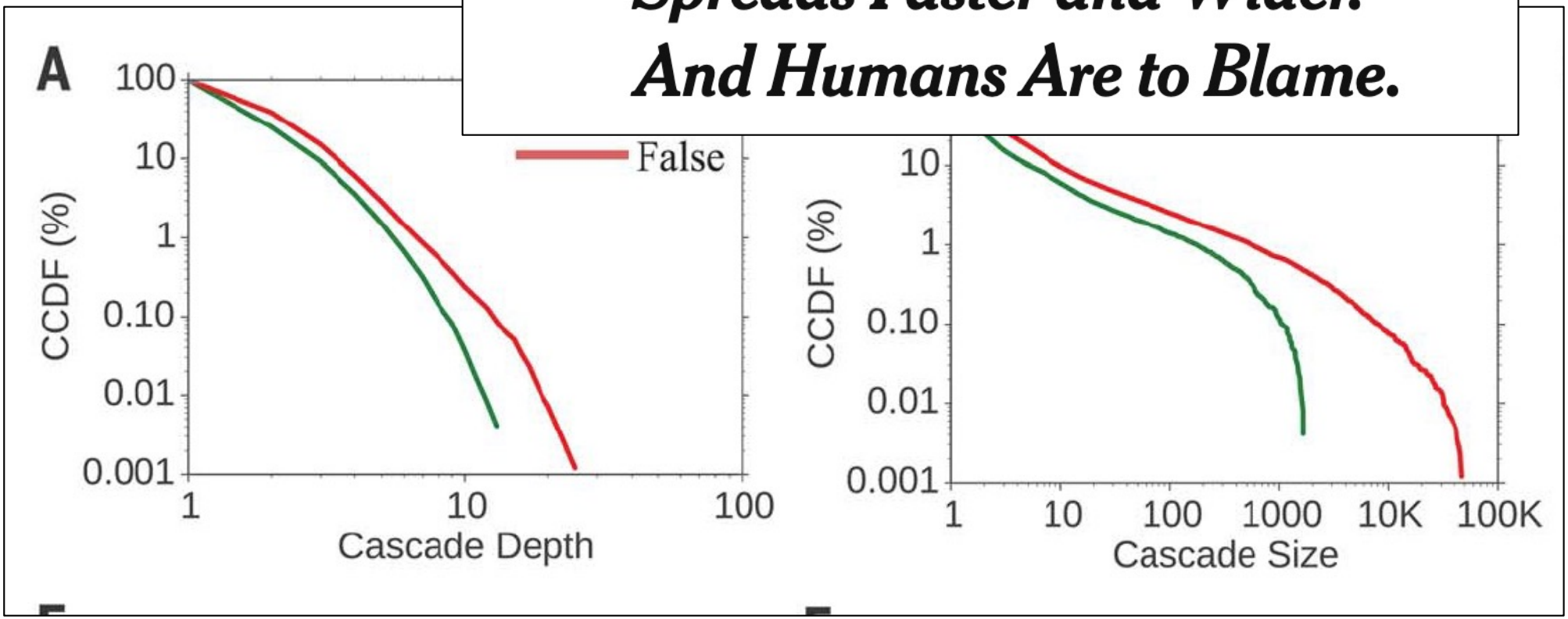


More Problems in Social Media?

4. Virality

The New York Times

It's True: False News Spreads Faster and Wider. And Humans Are to Blame.



[Vosoughi et al., 2018]

Popular Computational RQs

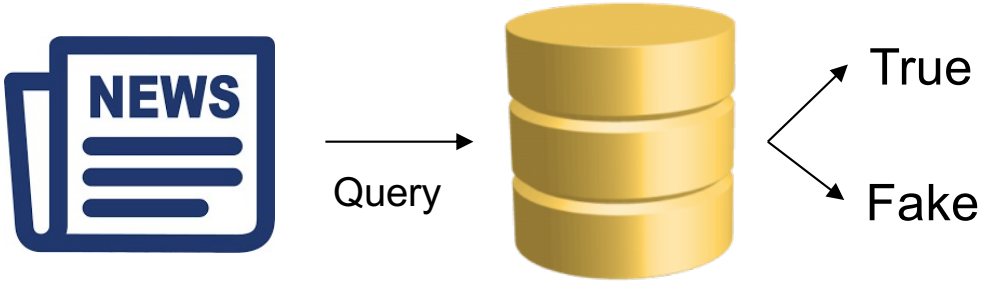
- What is false information?
  - Operational definition
  - Multi-faceted, multi-types
  - How to label?
  - Understand how it spreads?

- How to detect false information?
  - Which one to fact-check (among many)?
  - Detect accurately and early?
  - Can explain verdict?

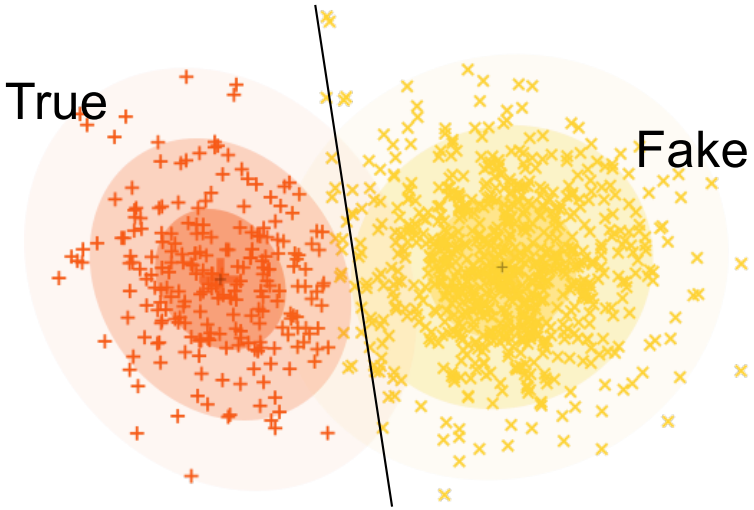
To Detect False Information

- Human Based
  - Manual fact-checking
  - Crowdsourcing based

- Machine Based
  - DB approach



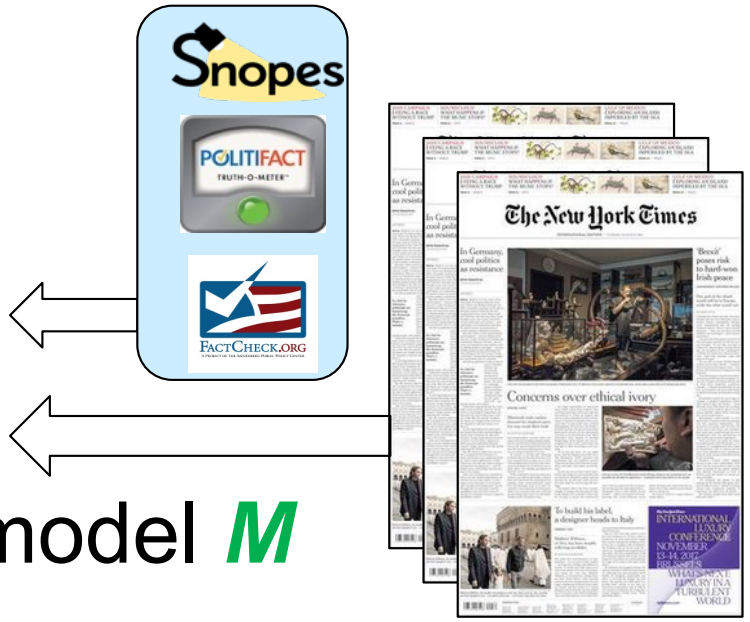
  - AI approach



AI: Supervised Learning Approach

In Training

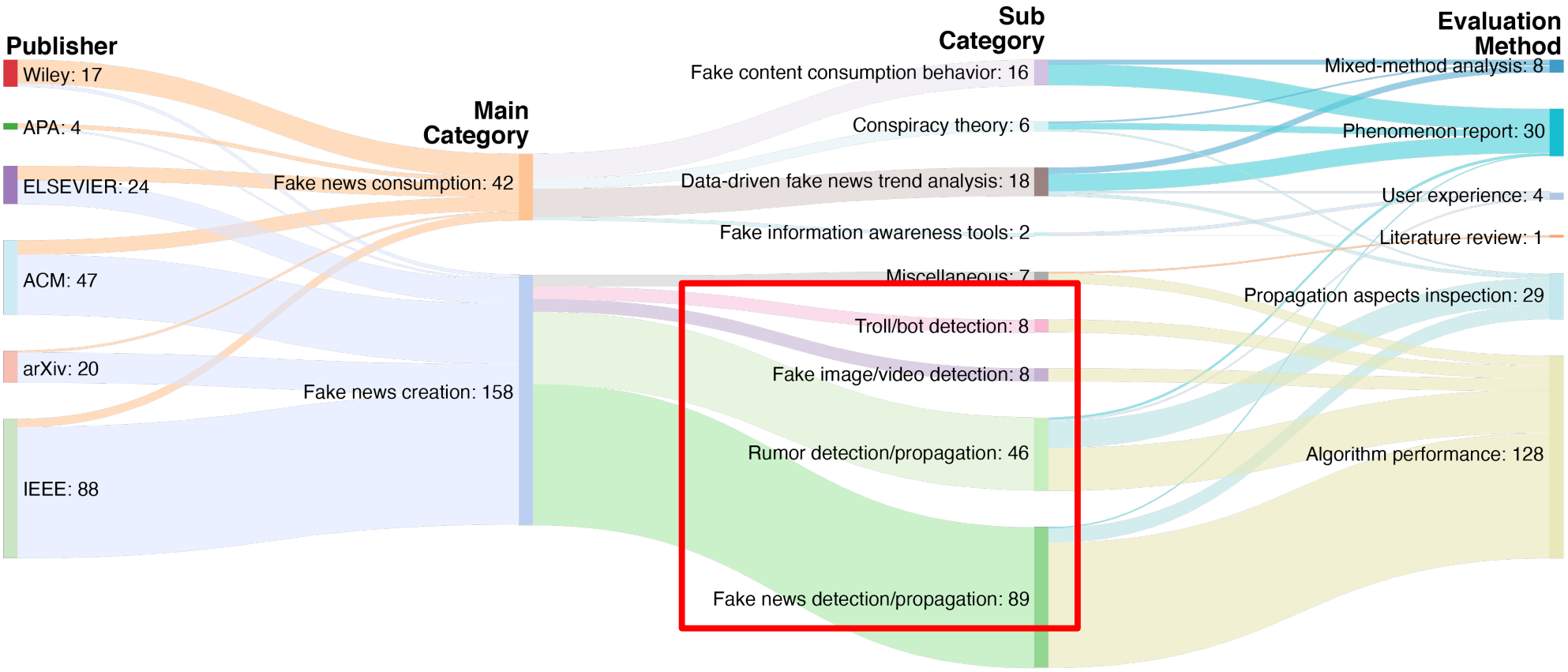
- Learning
  - **P**: Features from “fake” news
  - **N**: Features from “true” news
- Feed (**P**, **N**) to ML to build a model ***M***



In Deployment

- Feed a news story *A* to ***M***
- ***M*** determines if *A* is fake or true news story

Extensive Research (2010-2020)

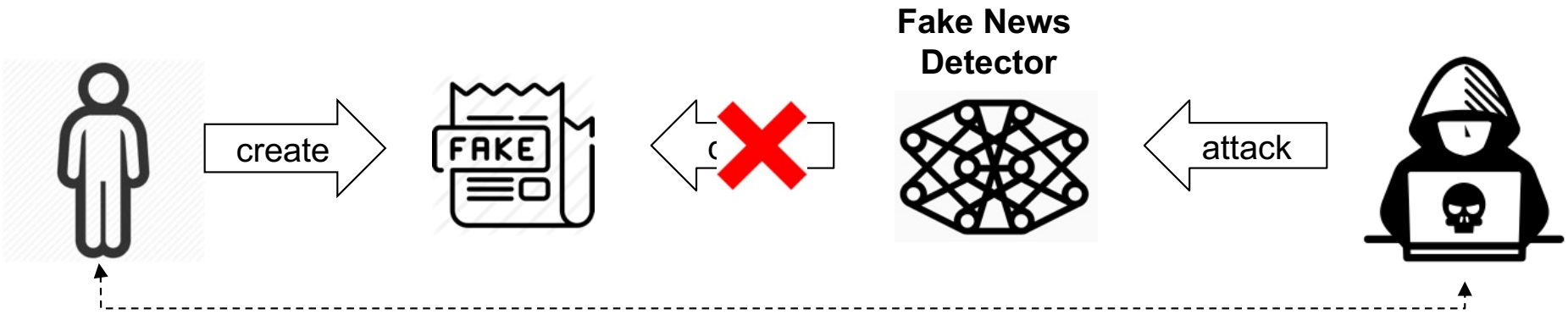


A visualization of 200 fake news-related papers published in IEEE, ACM, ELSEVIER, arXiv, Wiley, APA from 2010 to 2020

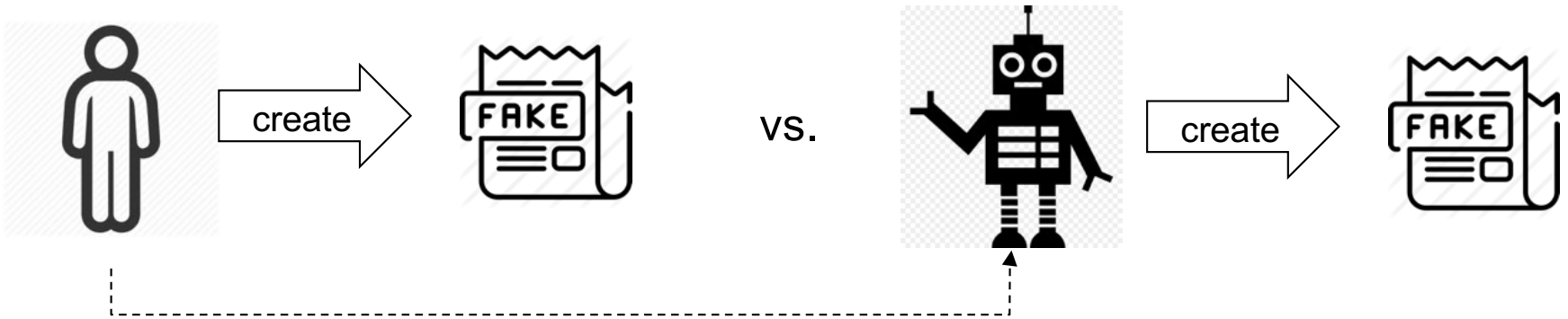
[Kim et al., 2021, Tech Report]

Fake News 2.0: New Challenges

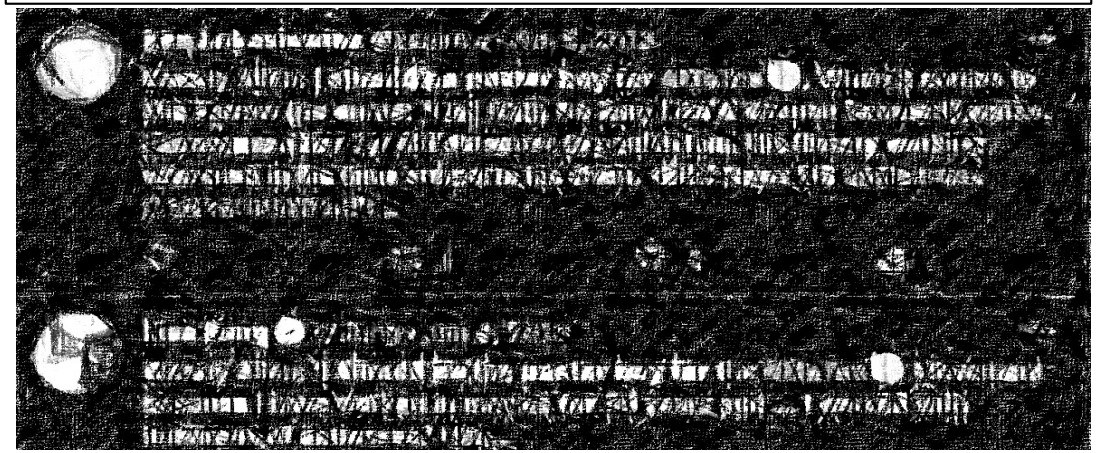
- Challenge #1

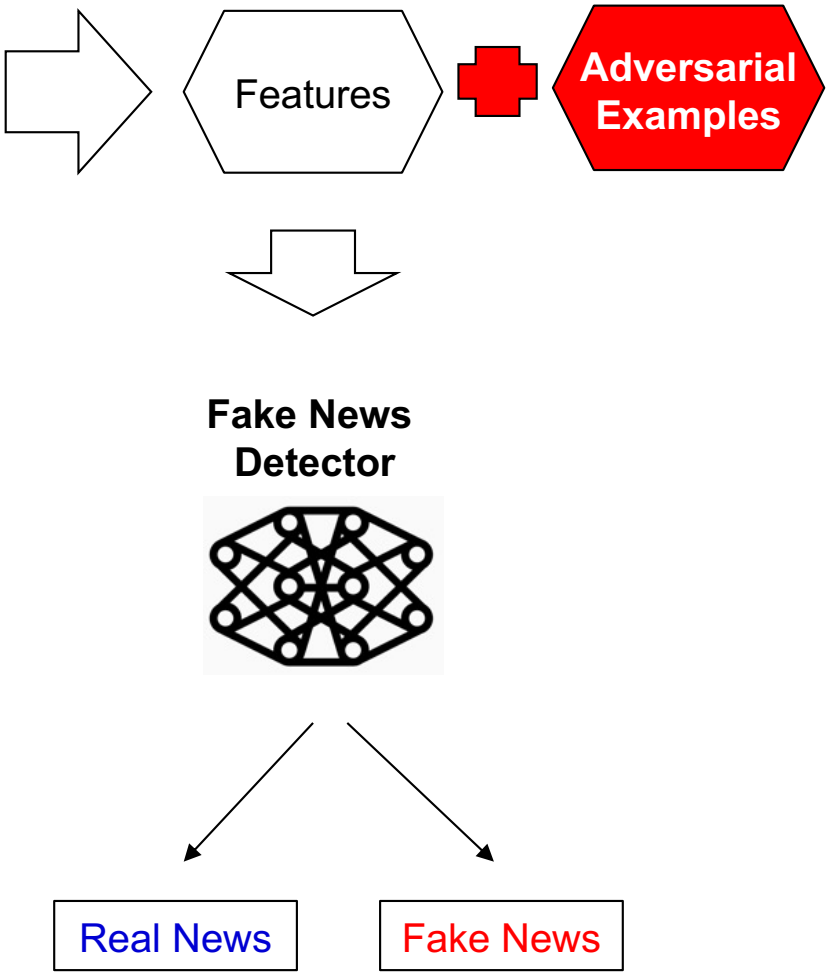


- Challenge #2

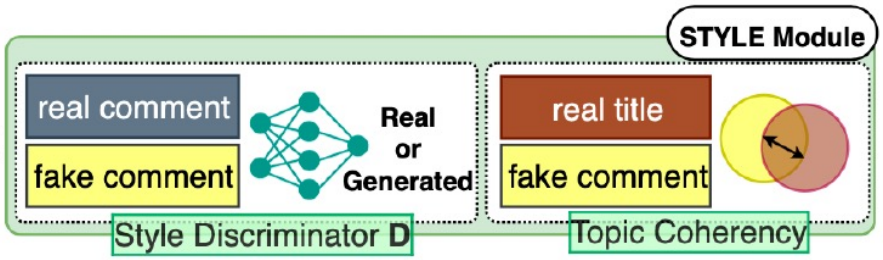


C1. Attack the Detector





C1. Attack the Detector



1. Should look like regular user comment

2. Contents should be relevant

Real Comment: admitting i'm not going to read this (...)
Malcom: *he's a conservative from a few months ago*
Prediction Change: Real News ⟶ Fake News

Whitebox: ~94%
Blackbox: ~90%

[Le, Wang & Lee, ICDM 2020]

C2. Neural Fakes: Text

Prompt

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

Machine

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others th nto the valley. "By the time we reached the water looked blue.

GPT-2 by OpenAI

PROMPT	Putin and Xi are using the coronavirus crisis to extend their control. Across the world, Trump is struggling to keep up
Human	. . . Donald Trump's shock announcement shutting down travel from much of Europe came with the customary back-pedaling common to many of his hastily announced initiatives. At first, Trump said trade and travel would be targeted, in his latest attempt to stem the spread . . .
CTRL	. . . . President Vladimir Putin, meanwhile, says there was no reason why Moscow would spread such diseases. But both leaders have been unable so far to . . . [read more] what do you think about this story? do we need another Ebola scare or two? . . .
GPT	. . . with them in a war that has been going on for decades now - but he's still not strong enough. " " i'm sorry about your family. it must be hard being so far away from home. " she said softly. " yes, we're all very lonely here. there wasn't much time before . . .
GPT2	. . . Putin's paranoia over the contagion of Ebola is a common theme. As his recent visit to Washington shows, he's playing a dangerous game. And here's what Putin wants Putin wants to use Ebola as a pretext for imposing further Russian sanctions on the West. . . .
GROVER	. . . hours after world cup workers hand-picked by Putin to fix problems faced by workers in turkey and Venezuela, a new threat to trump is receding.According to Wikileaks, the Coronavirus outbreak is receding and the Russian leader, Putin,. . .
XLM	. . . with Russia alone ruling over Europe while china under Mao Zedong rules Asia in turn which makes it difficult for him politically but financially as well being president / CEO of America first bank & chairman's nominee at CIA . . .
XLNET	. . . with china or Russia in his efforts at de-destructuring America from the financial crisis that has taken place over the past two years of this century as well a global economic downturn that will continue for many more months into next year . . .
PPLM	. . . the latest developments in Syria, Ukraine, Russia's intervention in the middle east and china's rise to global power. In Moscow, a group of Russian diplomats, who did not want their real names to be reported on the record. . .
FAIR	.. He has ordered his federal agencies to step up efforts to combat it and to warn of new threats. in Washington, Trump's national security adviser, h.r. mcmaster, said on Monday the U.S. government had seen signs of the bug and that countries in the Middle East were on high alert. . . .

C2. Neural Fakes: Text

- Comparison among 8 neural generative language models (LMs)
- Turing Test: is this text written by human or machine?

Turing Tester | 8 SOTA Neural Language Models

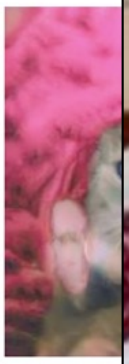
Model	CTRL	GPT	GPT2	GROVER	XLM	XLNET	PPLM	FAIR	AVG
Embedding	0.9768	0.9838	0.4044	0.6628	0.6535	0.6551	0.8449	0.5178	0.7124
RNN	**1.0**	0.9930	0.6329	**0.9977**	0.9977	**1.0**	0.9466	0.8812	0.9311
Stacked_CNN	0.9792	0.9815	0.6347	**0.9977**	0.9907	0.9186	0.6457	0.6316	0.8475
Parallel_CNN	**1.0**	**0.9977**	0.6075	0.9536	**1.0**	**1.0**	0.9513	0.9282	0.9298
CNN-RNN	**1.0**	0.9861	0.6626	**0.9977**	0.9699	0.9907	0.7949	0.7018	0.8880
RoBERTa	0.6448	0.6404	0.6407	0.6448	0.6490	0.7185	0.6404	0.6404	0.6524
RoBERTa-tuned	0.9730	0.9881	**0.9792**	0.8894	0.9921	0.9850	**0.9796**	**0.9753**	**0.9702**
GROVER-DETECT	0.7753	0.7319	0.6976	0.8135	0.6929	0.7536	0.7761	0.7616	0.7503
AVG	0.9186	0.9128	**0.6574**	0.8696	0.8682	0.8777	0.8236	0.7547	

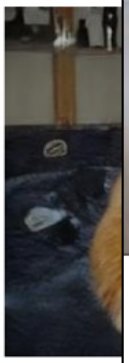
[Uchendu et al., EMNLP 2020]

http://thispersondoesnotexist.com

These Cats Do Not Exist

Learn More: Genera





ENTIRE GUEST SUITE

Ponta Mer Media Aireoper

Athens

Alice

7 guests · 3 bedrooms · 4 beds · 2 baths

My place is 2 mins frou shopping malls and restaurants open 360 HM have a double cable television. We will offer all the necessary noise and make your stay enjoyable with high speed internet available in the kitchen but also will be fully resolved to a comfortable room while day. Guests have access to the balc Tons of green pubs and all the restaurants are 3 trams ride away. Steeped in
bathroom, WC, French balco
front and back porch to ensure guests to sit around the corner for entry or dinner. There is

https://thisrentaldoesnotexist.com/

C2. Neural Fakes: Video

C2. Neural Fakes: Video

Text-based Editing of Talking-head Video

Ohad Fried*, Ayush Tewari^, Michael Zollhöfer*, Adam Finkelstein†, Eli Shechtman‡,
Dan B Goldman, Kyle Genova†, Zeyu Jin‡, Christian Theobalt^, Maneesh Agrawala*

* Stanford University
^ Max Planck Institute for Informatics
† Princeton University
‡ Adobe

Source: SIGGRAPH 2019

Potential Deepfake Scenario

single image

Eg, Samsung AI

Simple Animation

1-min audio

Eg, Lyrebird AI

text transcript

Synthesized Audio

hour-long video

text transcript

Eg, Stanford / UW / Albany AI methods

Synthesized Video

Commodity Technology

FaceApp - AI Face Editor 9+

Photo & Video Editor

FaceApp Technology Limited

#4 in Photo & Video

4.7 • 1.1M Ratings

Free · Offers In-App Purchases

Screenshots iPhone iPad

Facebook DFDC

Final Thought

- Detecting (traditional) false information
  - Extensive research since 2016
    - Several solutions with >90% accuracy
  - A few remaining challenges:
    - Early detection, explainability, few shot learning

- New challenges and opportunities
  - Detecting neural false information
  - Attributing authors of false information
  - Attacking and defending fake news detectors

➔ NO good computational solutions yet !

Final Thought

- Documentation is no longer evidence
- “Implied false effect”
- “Reality apathy” – Oyadya, 2019
- “Liar’s dividend” – Chesney and Citron