Trustworthy data science for social good

Xiuzhen Jenny Zhang (xiuzhen.zhang@rmit.edu.au)

RMIT University, Australia

RMIT UNIVERSITY

In this talk …

- The what and why of trustworthy data science
- Transparency: fighting misinformation with explanation
- Fairness: unbiased opinion summarization
- Responsibility: responsible information recommendation
- Discussions and conclusion

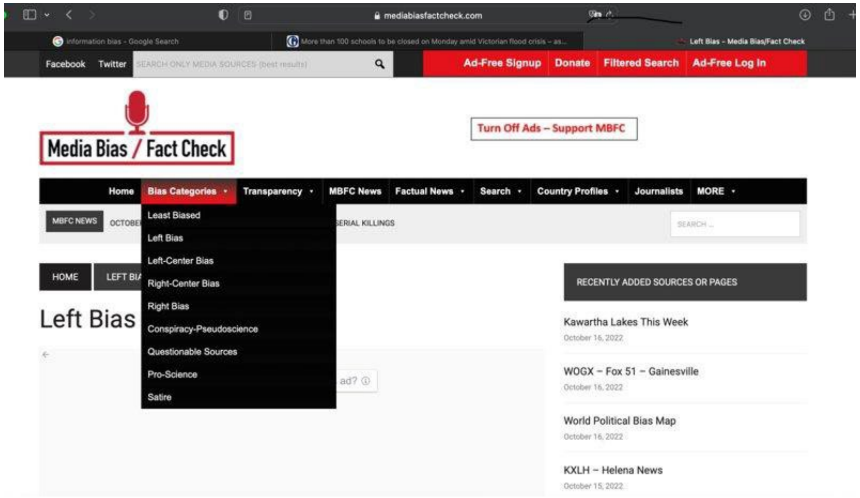
The what and why for trustworthy data science

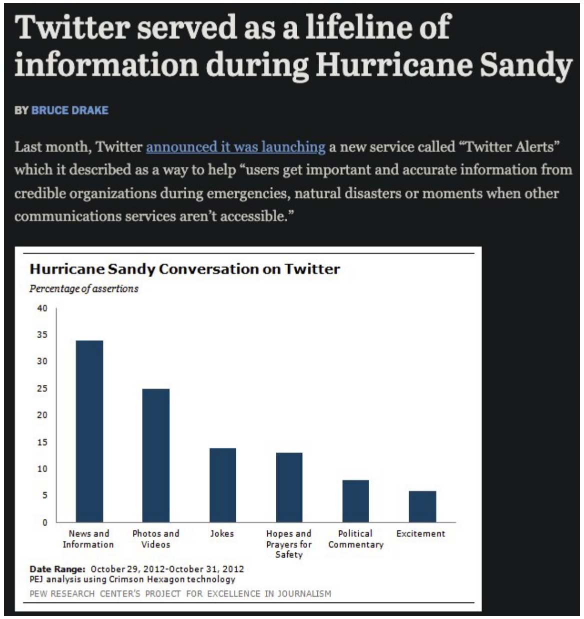
What is trustworthy data science?

- User trust is the ultimate testimony for successful data science and AI.
- What are the qualities of a trustworthy AI system?

Source: Requirements of Trustworthy AI | Futurium | European commission

Why trustworthy data science?
-- data, data, biased data

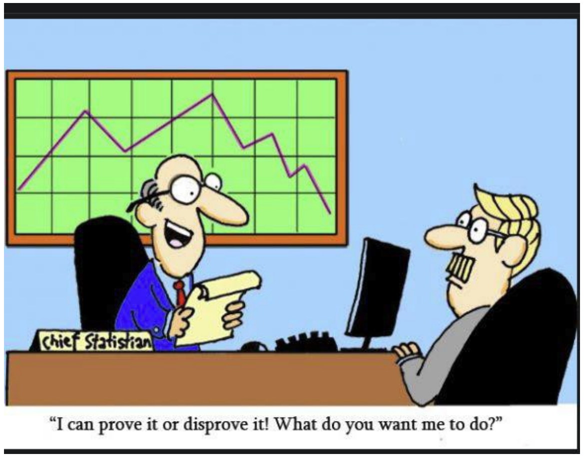




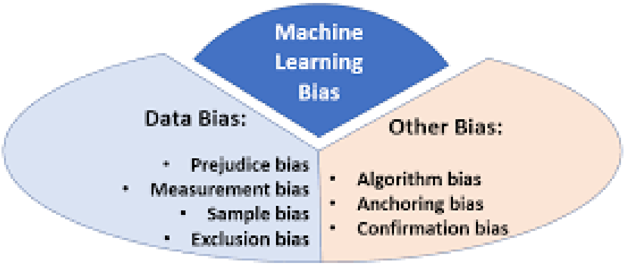
Twitter served as a lifeline of information during Hurricane Sandy

BY BRUCE DRAKE

Last month, Twitter announced it was launching a new service called "Twitter Alerts" which it described as a way to help "users get important and accurate information from credible organizations during emergencies, natural disasters or moments when other communications services aren't accessible."



"I can prove it or disprove it! What do you want me to do?"



Why trustworthy data science?
-- data, data, unlimited data

THE AGE

Apple News

reddit

YouTube

Google News

Why trustworthy data science?
- data, data, misinformation data

NYC EMS Website
@NYCEMSwebsite

Follow

NYPD going under water at East 8th St and Ave C in NYC due to Hurricane #Sandy

Reply Retweet Favorite More

RETWEETS 283 FAVORITES 18

12:57 PM - 30 Oct 2012

Lynn Brittney
@LynnBrittney2

Bill Gates is on record saying that his vaccine investments have given the best return - 20 to 1 and allowed him to buy a 66,000 sq.ft mansion, private jet, 242,000 acres of farmland, investments in fossil fuel dependent industries such as airlines. But he insists WE go green.

,230 Likes

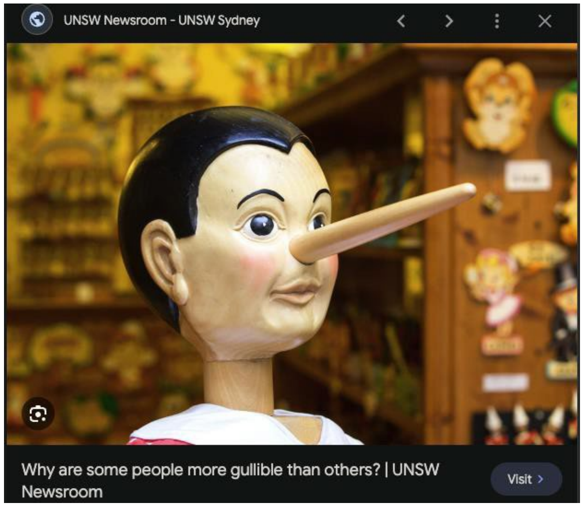
Viral 'Momo challenge' is a malicious hoax, say charities

Groups say no evidence yet of self-harm from craze, but resulting hysteria poses a risk

The Momo messages are said to come from a profile with this distorted image of a woman, but experts say the story is no more than a 'moral panic' among adults. Photograph: PSNI

It is the most talked about viral scare story of the year so far, blamed for child suicides and violent attacks - but experts and charities have warned that the "Momo challenge" is nothing but a "moral panic" spread by adults.

Why trustworthy data science?
- but, users, users, credulous users




Contents lists available at ScienceDirect

Computers in Human Behavior

Full length article

On the credibility perception of news on Twitter: Readers, topics an
features

Shafiza Mohd Shariff, Xiuzhen Zhang, Mark Sanderson

ARTICLE INFO

ABSTRACT

So, what can we work on NOW?

Trustworthy data science technologies are --

Transparent: automatic prediction with explanation,

Fair: generate information free of bias, and

Responsible: ensure positive social impact and responsibility.

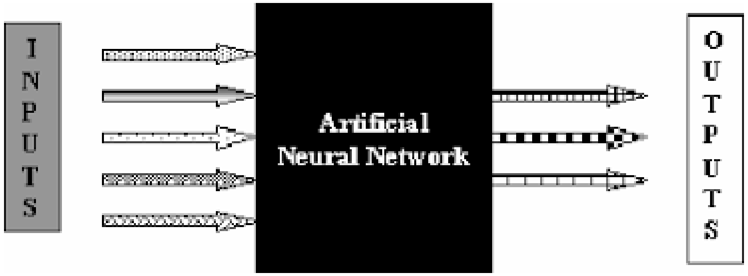
xiuzhen.zhang@rmit.edu.au

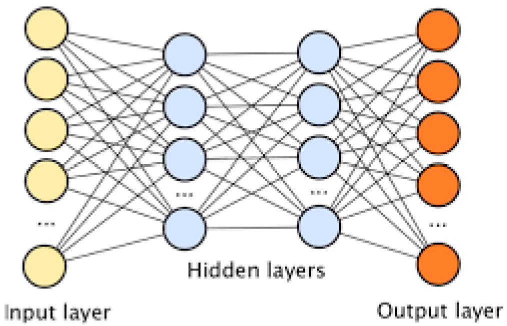
Combating misinformation with explanation

- Tian, L., Zhang, X. and Lau, J.H., 2023. CMA-R: Causal Mediation Analysis for explaining rumour detection. 2023. In submission.
- Tian, L., Zhang, X.J. and Lau, J.H., 2022, July. DUCK: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4939-4949).
- Tian, L., Zhang, X. and Lau, J.H., 2021. Rumour detection via zero-shot cross-lingual transfer learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21* (pp. 603-618). Springer International Publishing.
- Tian, L., Zhang, X., Wang, Y. and Liu, H., 2020. Early detection of rumours on twitter via stance transfer learning. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42* (pp. 575-588). Springer International Publishing.

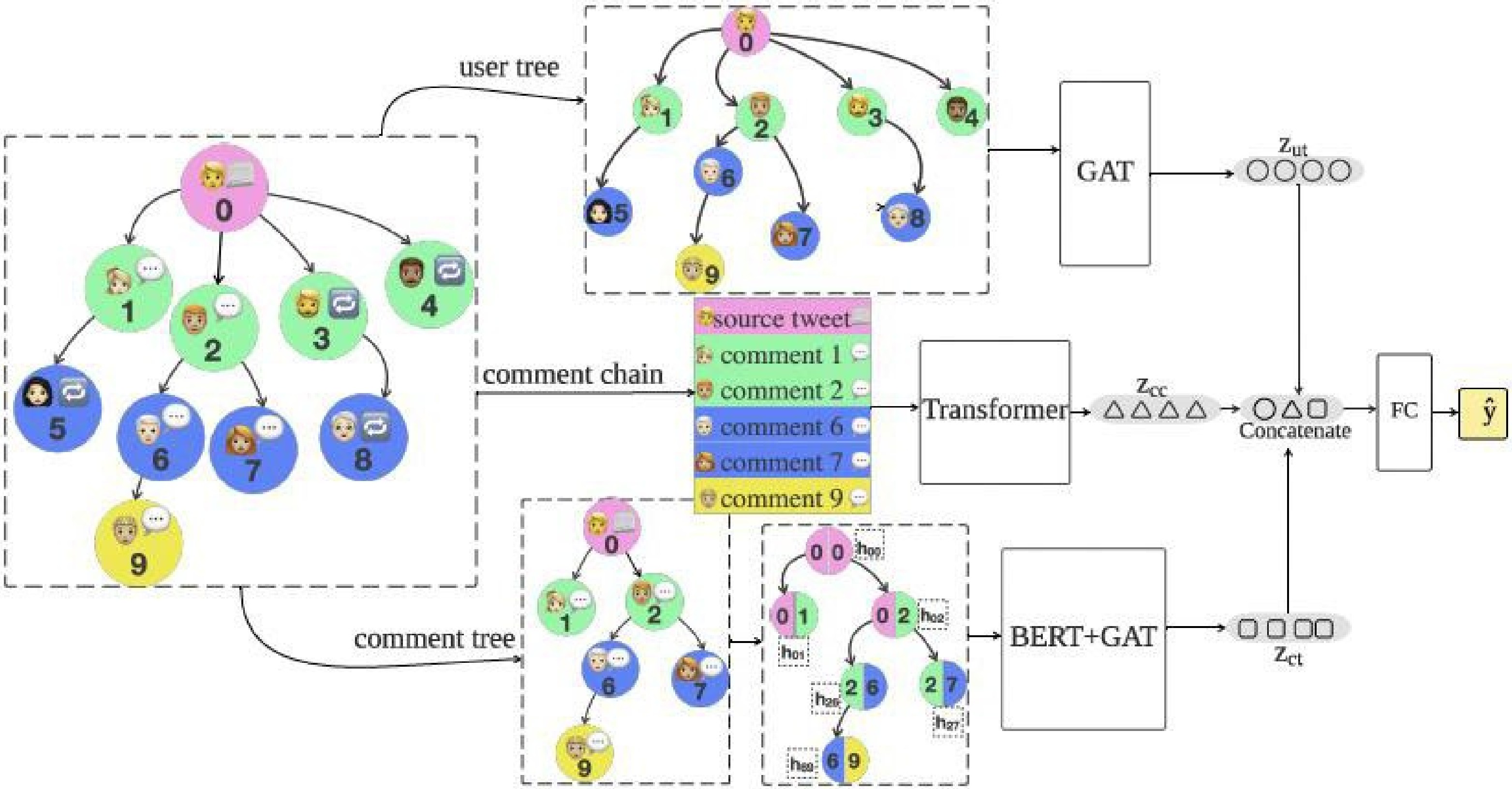
Automatic rumour detection on social media

The explainability issue





DUCK: rumour detection on social media by modelling user and comment propagation networks

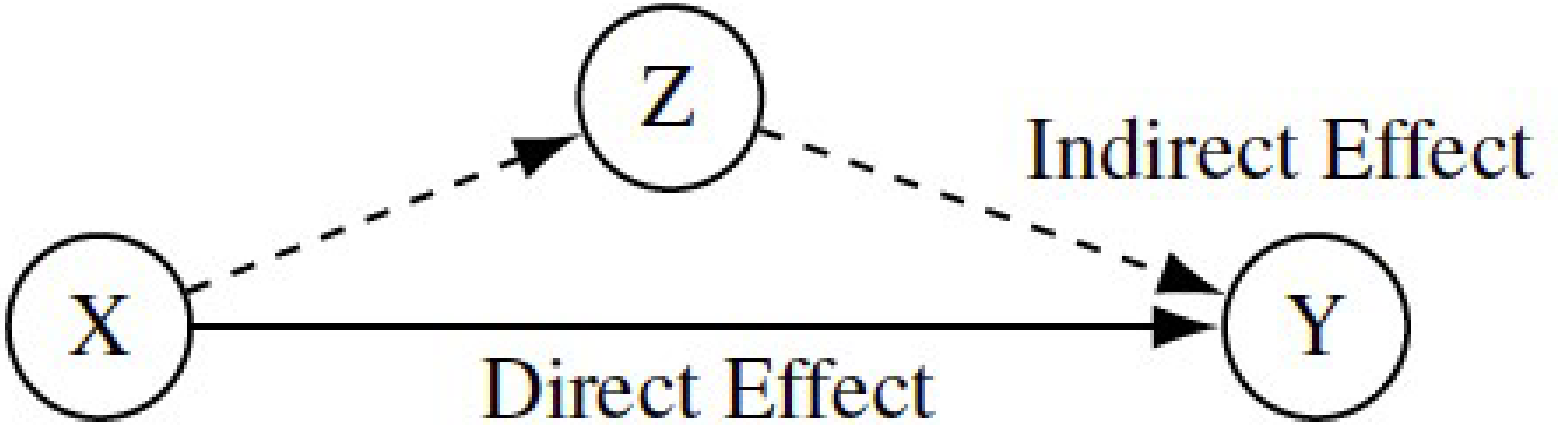


Results

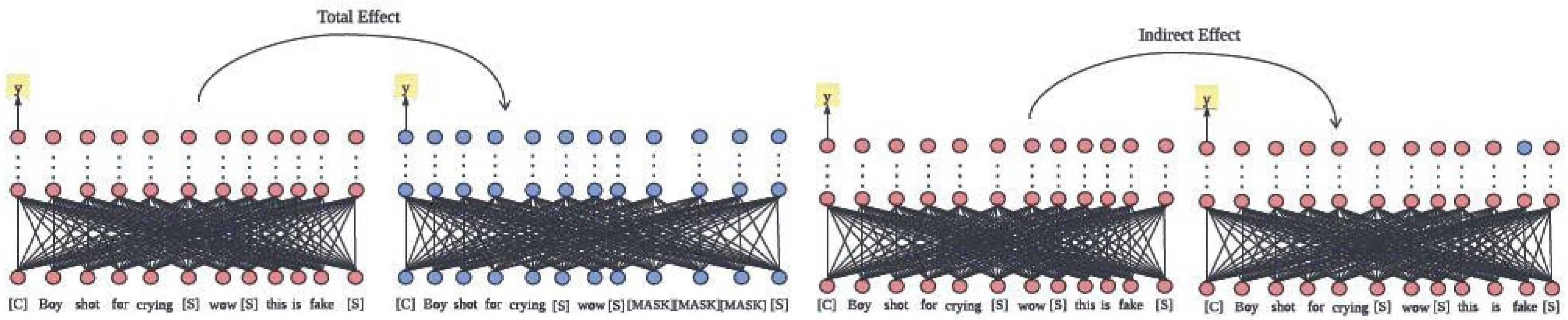
	Twitter15					Twitter16					CoAID			WEIBO		
Model	F1	FR	TR	NR	UR	F1	FR	TR	NR	UR	F1	T	F	F1	NR	R
RvNN	0.72	0.76	0.82	0.68	0.65	0.74	0.74	0.84	0.66	0.71	0.78	0.98	0.57	0.91	0.91	0.91
RNN+CNN	0.53	0.51	0.30	0.36	0.64	0.56	0.54	0.40	0.59	0.67	–	–	–	0.92	0.91	0.92
stance-BERT	0.82	0.82	0.85	0.87	0.71	0.83	0.82	0.88	0.83	0.77	0.90	**0.99**	0.81	–	–	–
Bi-GCN	0.86	0.85	0.91	0.84	0.82	0.86	0.86	**0.93**	0.79	0.86	0.83	**0.99**	0.68	0.96	0.96	0.96
GCAN	0.69	0.75	0.75	0.63	0.68	0.72	0.73	0.78	0.67	0.72	–	–	–	0.92	0.92	0.92
DUCK$_{\neg CT}$	0.82	0.72	0.91	0.82	0.85	0.84	0.88	0.81	0.88	0.79	0.91	**0.99**	0.82	0.93	0.93	0.93
DUCK$_{\neg CC}$	0.85	0.91	0.86	0.81	0.82	0.85	0.84	0.91	0.78	0.87	0.87	0.98	0.75	0.94	0.94	0.94
DUCK$_{\neg UT}$	0.88	**0.92**	0.84	**0.91**	0.85	0.89	**0.91**	0.91	0.87	0.88	0.91	**0.99**	0.83	0.97	0.97	0.97
DUCK	**0.90**	0.91	**0.93**	0.88	**0.88**	**0.91**	0.89	**0.93**	**0.93**	**0.91**	**0.92**	**0.99**	**0.85**	**0.98**	**0.98**	**0.98**

How to open the DUCK black box?

Causal mediation analysis



Causal mediation analysis of DUCK for explanation



Results

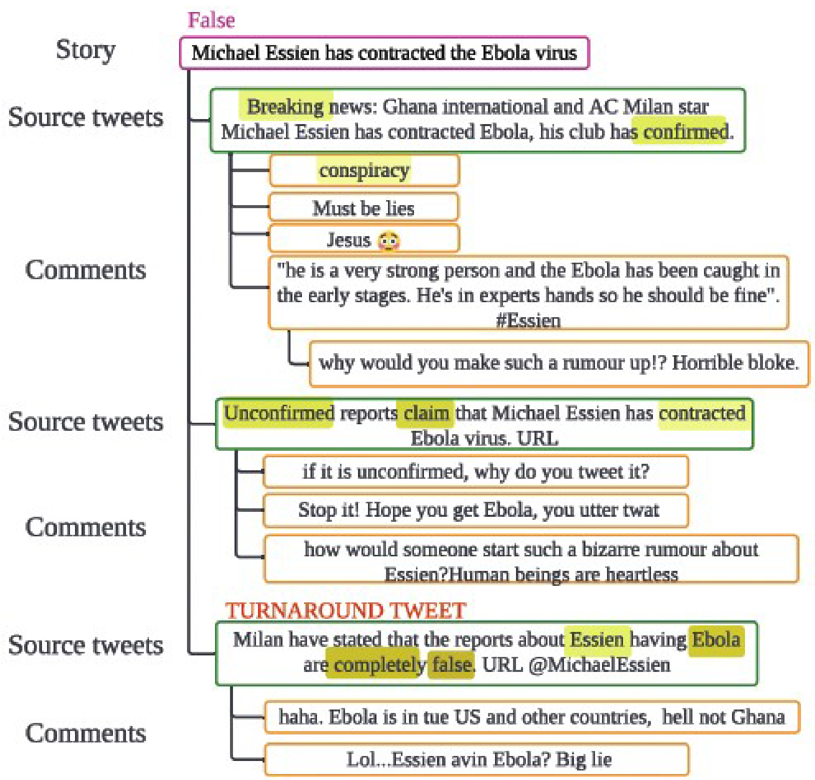


Figure 4: A labelled story in PHEME. Additional stories can be found in Appendix D

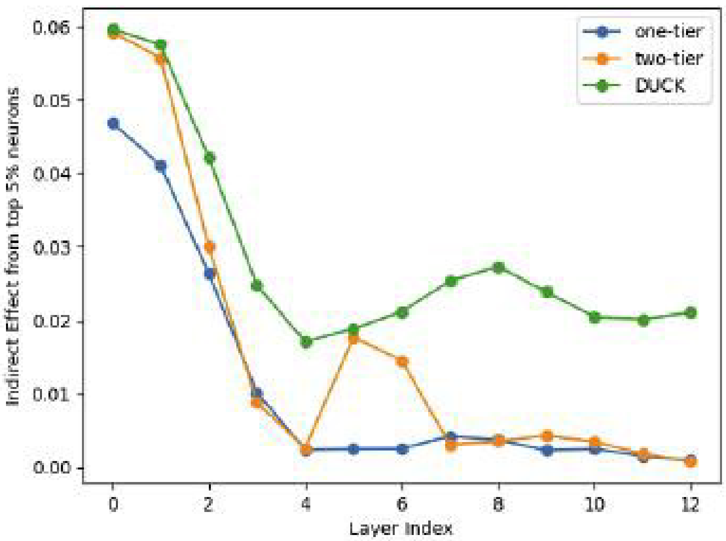


Figure 5: Indirect effects over different layers

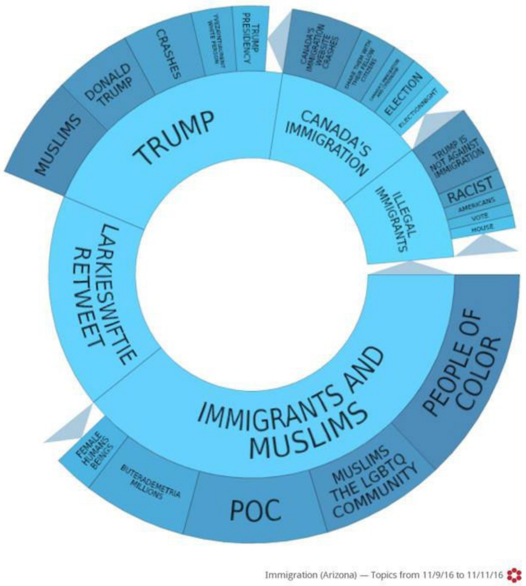
Summary

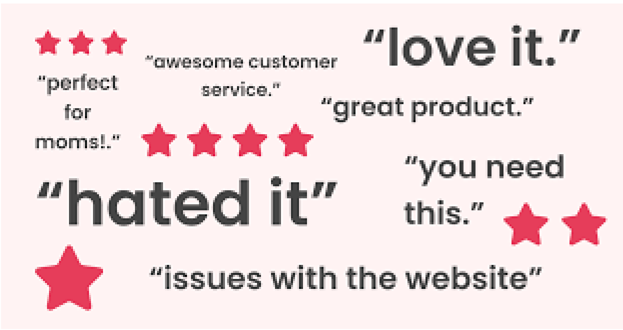
- Neural networks modelling of tweets and their conversation structure is effective for automatic rumour detection.
- Causal mediation analysis can open the blackbox of neural networks to identify critical tweets and tokens to explain the model predictions.

Reducing bias for fair opinion summarization

- Huang, N., Fayek, H. and Zhang, X., 2023. The bias in opinion summarization from pre-training to adaptation: a case study in political bias. In submission.
- Tang, A., Dinh, M., and Zhang, X., 2023. Aspect-based key point analysis for quantitative summarization of reviews. In submission.
- Huang, N. and Zhang, X., Evaluation of Review Summaries via Question-Answering. *ALTA 2021*.

Opinion summarization





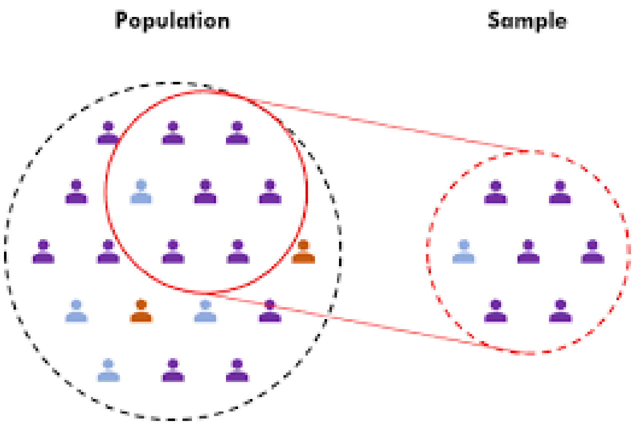

Beyond Opinion Mining:
Summarizing Opinions in Customer Reviews

Reinald Kim Amplayo, Arthur Bražinskas, Yoshi Suhara, Xiaolan Wang, Bing Liu

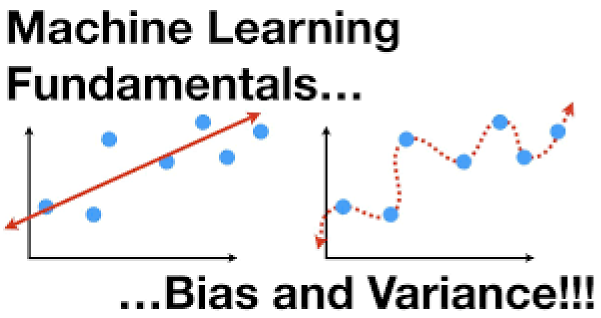
Google Research grammarly Megagon Labs UIC

The challenges

The input data bias



The algorithmic bias in models



Our goal: fair summarization of opinions

microsoft/**coresets-fair-diverse**

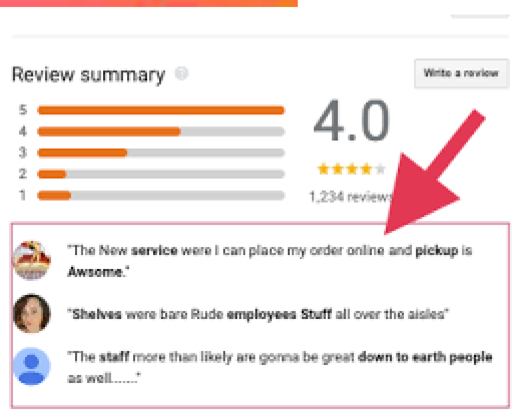
A repository for the paper Core-sets for Fair and Diverse Data Summarization, accepted for NeurIPS2023.

3 Contributors · 0 Issues · 1 Star · 1 Fork

guyfe/**Tweetsumm**

A dataset focused on summarization of dialogs, which represents the rich domain of Twitter customer care conversations

2 Contributors · 3 Issues · 20 Stars · 10 Forks



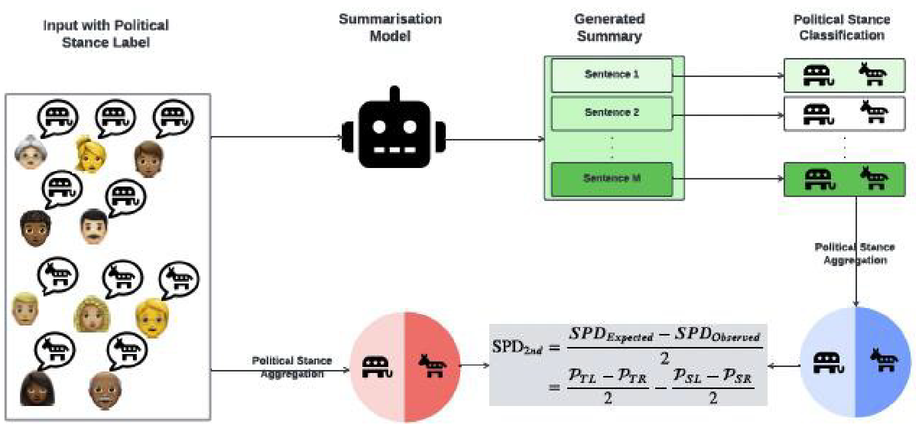
Beyond Opinion Mining: Summarizing Opinions in Customer Reviews

Reinald Kim Amplayo, Arthur Bražinskas, Yoshi Suhara, Xiaolan Wang, Bing Liu

Google Research · grammarly · Megagon Labs · UIC

Measuring the political bias in Twitter summarization

Figure 1: The process of measuring fairness in our study. For the input documents, each tweet has a label indicating the tweet is expressing a left or right-leaning stance. After feeding the input documents to the summarisation models, we split and classify each sentence in the summary to capture its left or right-leaning stance. We aggregate both the source documents and summary sentences on political stance, calculate the Second-order SPD (more detail in Section 4.2), and use it as the fairness measurement.



Results

Table 1: Intrinsic bias in different models under zero-shot setting for summary generation. The Second-order SPD (SPD_{2nd}) is reported for measuring the fairness of models using different input proportions (equal, more left-leaning, and more right-leaning). Model performance can be found in Table 4 in Section A.1.

Model	SPD_{2nd}-Equal	SPD_{2nd}-Left	SPD_{2nd}-Right
BART Base	-0.0262	0.1219	-0.2285
BART Large	-0.0240	0.0708	-0.2279
Distil GPT-2	-0.1154	0.0321	-0.3520
GPT-2	-0.0345	-0.0115	-0.2839
GPT-2 Medium	-0.0162	-0.0160	-0.2619
GPT-2 Large	0.0012	-0.0345	-0.2913
T5 Small	-0.0415	0.0424	-0.1957
T5 Base	-0.1385	-0.0390	-0.2479
T5 Large	-0.0160	0.1205	-0.2698

Table 2: Results of model performance and fairness evaluation. We highlight the adaptation methods apart from standard fine-tuning with the highest ROUGE score using *. We report Second-order SPD (SPD_{2nd}) with different input proportions (equal, more left-leaning, and more right-leaning), the lowest absolute values are bolded and the ranking compared between adaptation methods is provided inside the brackets.

Model	Adaptation Methods	ROUGE-1	ROUGE-2	ROUGE-L	SPD_{2nd}-Equal	SPD_{2nd}-Left	SPD_{2nd}-Right
BART Base	**Standard**	32.02	12.02	22.73	-0.2582 (4)	-0.1111 (3)	-0.4617 (4)
	Adapter	31.88*	12.21*	22.80*	-0.0530 (3)	-0.0090 (2)	-0.1106 (2)
	Prefix	29.37	9.89	20.00	0.0502 (2)	0.1666 (4)	**-0.1083 (1)**
	Last Layer	29.82	10.39	20.56	**-0.0470 (1)**	**0.0247 (1)**	-0.2370 (3)
BART Large	**Standard**	31.20	11.63	22.06	-0.2895 (4)	-0.1582 (3)	-0.4664 (4)
	Adapter	31.95*	12.22*	22.73*	**-0.0520 (1)**	**0.0518 (1)**	**-0.1869 (1)**
	Prefix	26.87	9.01	16.80	-0.0835 (2)	0.1735 (4)	-0.2004 (2)
	Last Layer	29.98	10.00	20.33	-0.1648 (3)	-0.0816 (2)	-0.3906 (3)
Distil GPT-2	**Standard**	21.76	5.78	16.44	-0.2788 (3)	-0.0829 (3)	-0.4766 (3)
	Adapter	21.12*	4.95*	14.95*	**-0.1568 (1)**	**0.0347 (1)**	**-0.3307 (1)**
	Prefix	10.39	3.02	8.21	-0.3532 (4)	-0.1368 (4)	-0.5357 (4)
	Last Layer	12.83	2.86	9.63	-0.2110 (2)	-0.0673 (2)	-0.3812 (2)
GPT-2	**Standard**	22.74	5.93	16.05	-0.2264 (4)	-0.0883 (3)	-0.4768 (4)
	Adapter	21.34*	4.97*	14.84*	-0.1331 (2)	**0.0272 (1)**	-0.3889 (3)
	Prefix	10.13	2.61	7.99	**-0.0833 (1)**	0.1136 (4)	**-0.3611 (1)**
	Last Layer	19.23	4.22	13.87	-0.1549 (3)	-0.0569 (2)	-0.3634 (2)
GPT-2 Medium	**Standard**	23.39	6.43	16.94	-0.2262 (4)	**0.0077 (1)**	-0.4227 (3)
	Adapter	22.46*	6.12*	16.20*	**-0.1421 (1)**	0.0291 (3)	-0.3844 (2)
	Prefix	16.78	5.78	12.80	-0.1525 (2)	0.0638 (4)	**-0.3711 (1)**
	Last Layer	19.37	3.61	13.50	-0.1835 (3)	-0.0165 (2)	-0.4478 (4)
GPT-2 Large	**Standard**	24.58	8.13	18.45	-0.2030 (4)	-0.0225 (3)	-0.3490 (4)
	Adapter	23.52*	6.62*	16.30*	-0.1715 (3)	-0.0172 (2)	**-0.2951 (1)**
	Prefix	12.54	4.25	9.42	**-0.0670 (1)**	**0.0166 (1)**	-0.3038 (2)
	Last Layer	19.26	5.18	14.24	-0.1403 (2)	-0.0554 (4)	-0.3425 (3)
T5 Small	**Standard**	27.75	9.74	19.52	-0.1891 (2)	-0.0672 (2)	**-0.3129 (1)**
	Adapter	24.89	9.05	17.42	-0.3464 (4)	-0.1681 (4)	-0.5191 (4)
	Prefix	28.10*	9.56*	19.03	-0.2494 (3)	-0.0983 (3)	-0.4784 (3)
	Last Layer	27.86	9.31	19.28*	**-0.1831 (1)**	**-0.0485 (1)**	-0.3791 (2)
T5 Base	**Standard**	29.86	9.82	20.49	-0.1297 (3)	**0.0338 (1)**	-0.2512 (2)
	Adapter	27.94*	10.17*	20.19*	**0.0284 (1)**	0.1397 (4)	**-0.1263 (1)**
	Prefix	25.40	9.03	18.11	-0.2150 (4)	-0.0593 (3)	-0.3530 (4)
	Last Layer	26.49	7.85	17.38	-0.1293 (2)	0.0430 (2)	-0.2913 (3)
T5 Large	**Standard**	31.08	11.52	22.20	-0.1211 (3)	**0.0072 (1)**	-0.2951 (2)
	Adapter	30.34*	11.30*	21.85*	-0.1207 (2)	0.0133 (2)	**-0.2069 (1)**
	Prefix	26.44	9.66	18.91	-0.4917 (4)	-0.2427 (4)	-0.7376 (4)
	Last Layer	22.80	7.58	16.25	**-0.0808 (1)**	0.0570 (3)	-0.3522 (3)

Review summarization: the issue

Food was good. Service was quick and friendly. The only reason I would give this place 5 stars is because it's not as good as it gets. Food was mediocre at best. Would not recommend.

Our goal: Key point-based quantitative summarization of reviews

(a) **The input comments.** Each box represents a review containing several comments

Review	Comments (review sentences)
1	**1.1:** The service is great and the staff is friendly and engaging. **1.2:** The food is excellent but the portion is quite small and quite expensive.
2	**2.1:** The food has great taste but very small portion and the service is slow.
3	**3.1:** The service was good and the food was delicious. **3.2:** Staff is friendly and attentive.
4	**4.1:** Food was excellent and delicious. **4.2:** Service and staff are excellent.
...	...

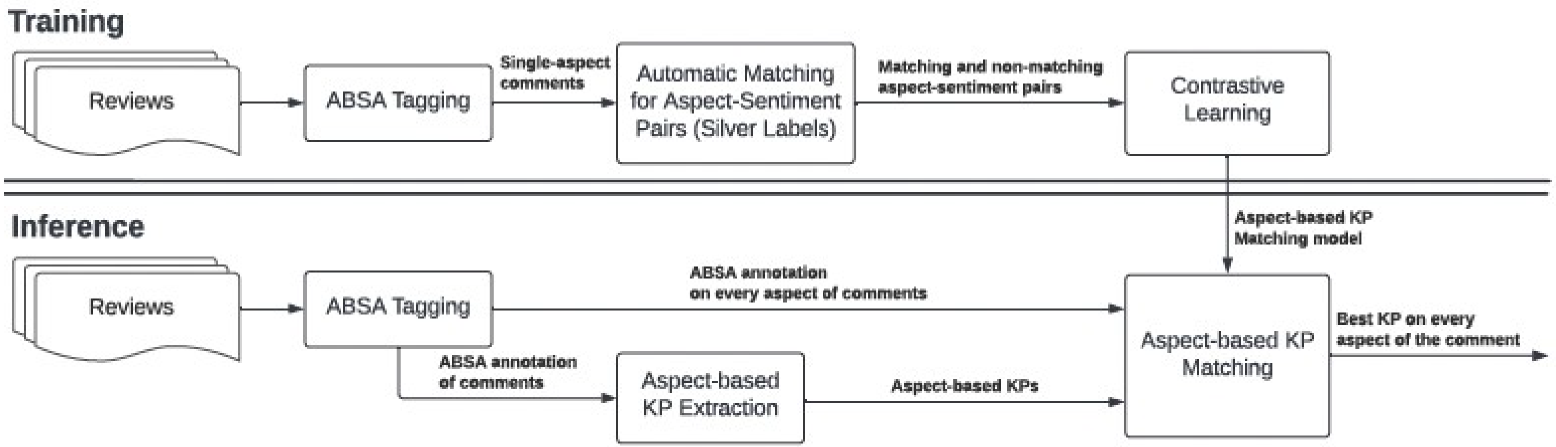
(b) Sentence-based KPs and their salience score (Bar-Haim et al., 2021, 2020a) output. Note that a commment can only be matched with one KP on of highest confidence.

Key points	Matched Comments	Salience score
KP1: Service and staff are excellent.	1.1	1
KP2: Service was prompt and friendly. (*redundant*)	3.1	1
...	...	...
KP3: Small and overpriced portion.	1.2	1
KP4: Small food portion and slow service. (*redundant*)	2.1	1
...	...	...

(c) **ABKPA KPs and their salience score. ABKPA ensures retrieving single-aspect key points with better opinion quantification specific to every comment's aspect**

Key points	Matched Comments	Salience score
KP1: Food was excellent and delicious.	1.2; 2.1; 3.1	3
KP2: Service was prompt and friendly.	1.1; 3.1	2
KP3: Staff is friendly and attentive.	1.1	1
...	...	...
KP4: Small and overpriced portion.	1.2; 2.1	2
KP5: Service was poor and slow	2.1	1
...	...	...

ABKPA:



xiuzhen.zhang@rmit.edu.au

Results on Yelp reviews

Table 7: Top 6 positive-sentiment key points ranked by their predicted prevalence on "Restaurants" datasets. While ABKPA generates distinct KPs on single aspects, baseline models generate KPs with overlapping aspects and opinions. KPs that overlap with higher-ranked ones (i.e., KPs with higher prevalence) are noted with a (*redundant*) postfix

ABKPA	SMatch	RKPA+	RKPA	ABKPA$_{\neg C}$
Staff was courteous and accommodating.	Staff was courteous and accommodating.	Staff was courteous and accommodating.	Employees are friendly and attentive.	Staff was courteous and accommodating.
Generous sized portions.	Prices are fair and reasonable.	The service here was exceptional.	The service here was exceptional.	Fresh food , using local produce.
Service was prompt and friendly.	Fresh food , using local produce.	Fresh food , using local produce.	Ambiance is casual and comfortable.	Customer service is excellent.
Fantastic drink selection.	The service here was exceptional.	The food is consistently excellent!	Fresh food , using local produce.	The service here was exceptional. (*redundant*)
Prices are fair and reasonable.	Generous sized portions.	Customer service is excellent. (*redundant*)	Really delicious food , well balanced!	Lots of outdoor seating.
Delicious and expertly prepared food.	Service was prompt and friendly. (*redundant*)	Prices are fair and reasonable.	Staff was courteous and accommodating. (*redundant*)	Amazing authentic flavor!

Table 3: AP score of KP Matching models. The best result of each experiment is in bold.

Dataset	All comments				Multiple-opinion comments			
	ABKPA	SMatch	comm-Match	RKPA	ABKPA	SMatch	comm-Match	RKPA
Arts	**0.99**	0.98	0.94	0.79	**0.99**	0.88	0.83	0.90
Auto	**0.77**	0.75	0.43	0.54	**0.80**	0.70	0.42	0.71
Beauty	**0.98**	0.97	0.84	0.62	**0.94**	0.88	0.81	0.62
Hotels	**0.99**	0.98	0.98	0.81	**0.93**	0.89	0.93	0.81
Restaurants	**0.87**	0.85	0.73	0.50	**0.83**	0.75	0.73	0.56
Average	**0.92**	0.91	0.78	0.65	**0.90**	0.82	0.74	0.72

Summary

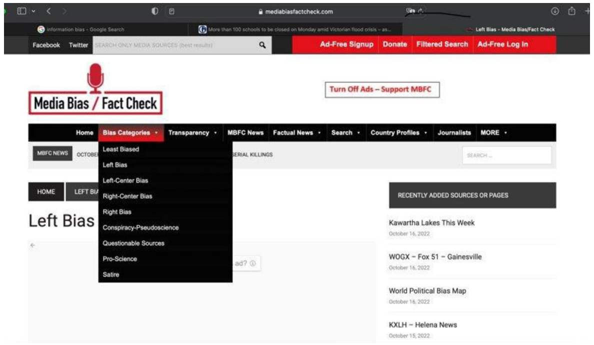
- Large language models have inherent bias and can propagate into summarization of social media opinions.
- Lighter fine-tuning strategies imply less distortion of the political stance in source input.
- Quantitative summarization is effective for including diverse opinions for review summarization.

Information recommendation

Responsible information recommendation

- Wang, S., Zhang, X., Wang, Y., Liu, H., and Ricci, F., 2023. *Trustworthy Recommender Systems*. ACM Transactions on Intelligent Systems and Technology. To appear.
- Wang, S., Liu, N., Zhang, X., Wang, Y., Ricci, F. and Mobasher, B., 2022. *Data Science and Artificial Intelligence for Responsible Recommendations*. KDD 2022.
- Wang, S., Xu, X., Zhang, X., Wang, Y. and Song, W., 2022, April. Veracity-aware and Event-driven Personalized News Recommendation for Fake News Mitigation In *Proceedings of the Web Conference.*

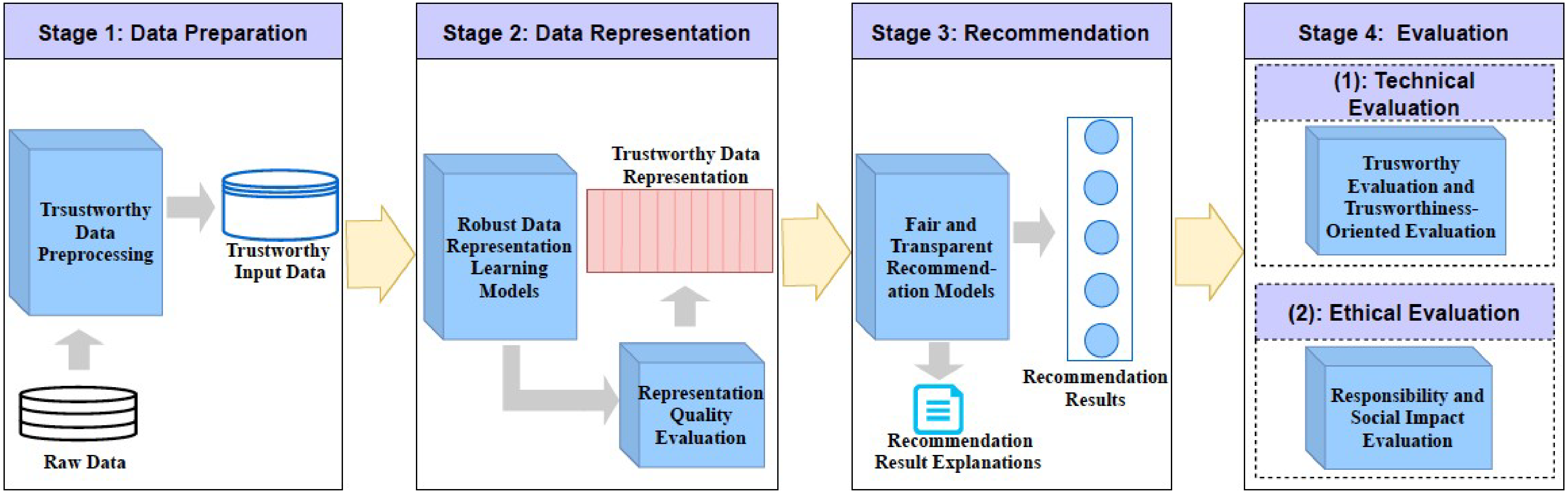
Challenges: misinformation, bias and moral value



Our goal: Personalized, responsible recommendation of information items

Responsible recommender systems have the objective of promoting moral value as well as personal value for users.

Building trustworthy recommender systems



Research questions

- How to learn veracity-aware item representation?
- How to recommend relevant news?
- How to only recommend true news when the veracity of candidate news is unknown?
- How to model the transition over latent events while avoiding the interference from veracity-related information (e.g., news content style)?

Rec4Mit: Veracity-aware news recommendation*

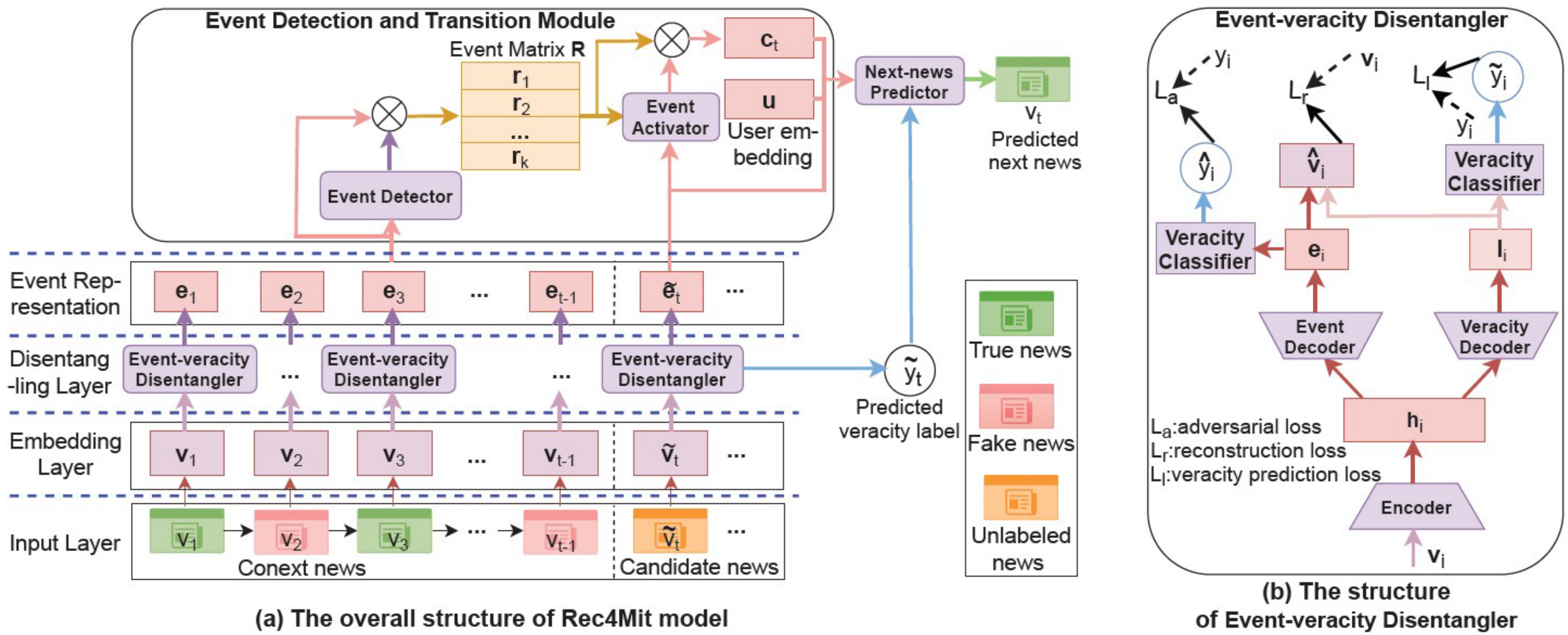


Figure 1: (a) Rec4Mit is built on three main components: Event-veracity Disentangler, Event Detection and Transition Module, and Next-news Predictor; (b) The Event-veracity Disentangler is built on the Encoder, Event Decoder, Veracity Decoder and Veracity Classifier.

Results: recommendation accuracy + ratio of true news

Table 2: Comparison of prediction accuracy with baselines on two datasets, *the improvement is significant at $p < 0.05$.

	PolitiFact						GossipCop					
	REC@5	REC@20	MRR@5	MRR@20	NDCG@5	NDCG@20	REC@5	REC@20	MRR@5	MRR@20	NDCG@5	NDCG@20
SKNN	0.2176	0.6088	0.1171	0.1553	0.1414	0.2524	0.1697	0.6252	0.0394	0.0911	0.0703	0.2074
CSRM	0.3752	0.6773	0.2629	0.2923	0.2906	0.3763	0.4764	0.6387	0.3496	0.3661	0.3813	0.4281
SR-GNN	0.3678	0.6741	0.2562	0.2865	0.2837	0.3711	0.4920	0.6239	0.3933	0.4067	0.4180	0.4560
SASRec	0.2962	0.6608	0.1582	0.1933	0.1924	0.2954	0.2419	0.4655	0.1009	0.1244	0.1358	0.2010
DAN	0.1874	0.7405	0.0784	0.1338	0.1049	0.2637	0.3174	0.4541	0.3157	0.3257	0.3161	0.3512
NRMS	0.4752	0.8260	0.3103	0.3449	0.3511	0.4511	0.6354	0.8239	0.4505	0.4702	0.4966	0.5516
LSTUR	0.4827	0.8111	0.3166	0.3491	0.3577	0.4515	0.6950	0.8817	0.4955	0.5156	0.5454	0.6005
FedNewsRec	0.3584	0.7949	0.1940	0.2377	0.2344	0.3596	0.2267	0.4892	0.1248	0.1498	0.1499	0.2237
FIM	0.3711	0.7042	0.1930	0.2250	0.2371	0.3311	0.3521	0.5911	0.2340	0.2570	0.2631	0.3312
Rec4Mit	**0.5561***	**0.8868***	**0.3462***	**0.3808***	**0.3979***	**0.4944***	**0.7552***	**0.9543***	**0.4984***	**0.5205***	**0.5625***	**0.6220***
Improvement[1]	15.21%	7.36%	9.35%	9.08%	11.24%	9.50%	8.66%	8.23%	0.59%	0.95%	3.14%	3.58%

[1] The improvement over the best-performing baseline methods whose performance is underlined.

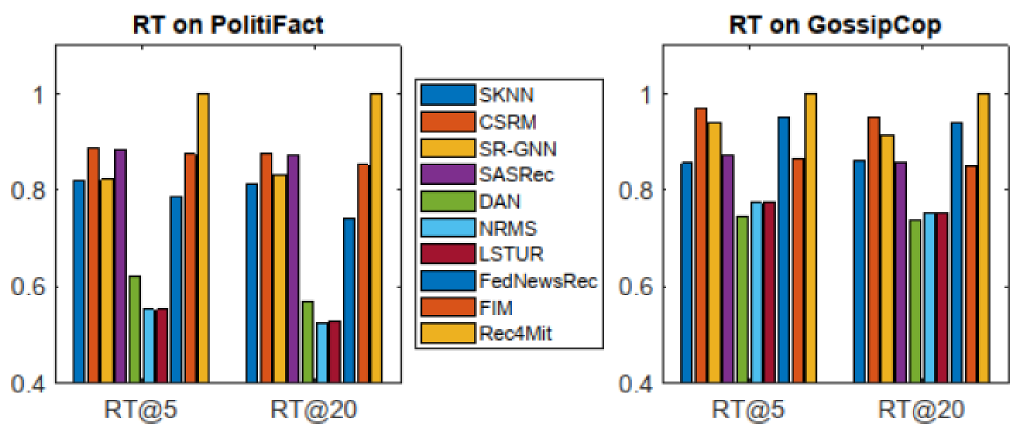


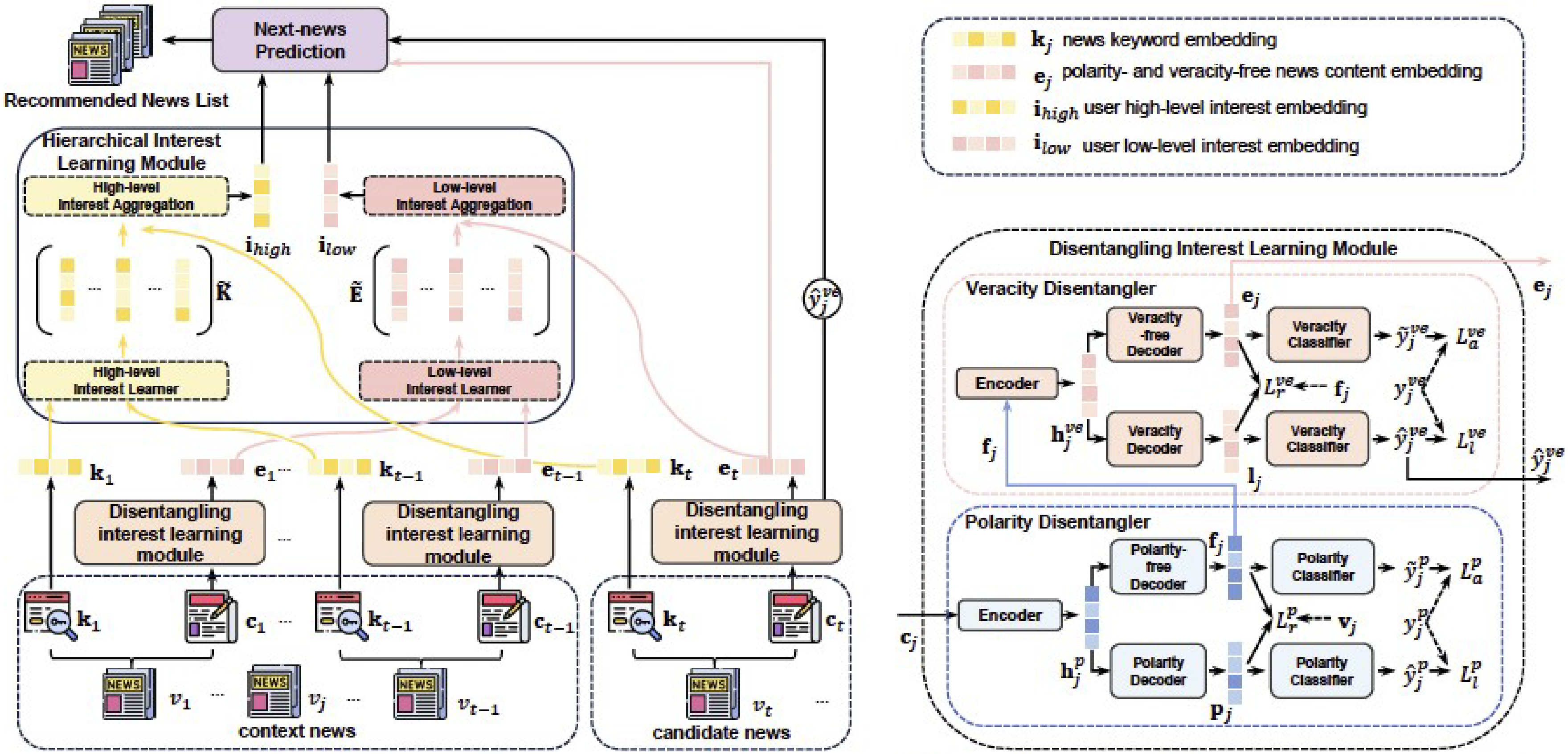
Figure 2: The ratio of true news (RT) in recommendation lists.

Case studies: the generated recommendation results

Table 4: **Recommendation Lists for 5 Users Sampled from GossipCop Dataset.**

User$_1$	**Context news (CN)**	CN$_1$:[1]jennifer lawrence says u mother u led to darren split	CN$_2$:[2]where is travis scott why kylie jenner s boyfriend avoids the spotlight	CN$_3$:jennifer lawrence says u mother u led to darren split	CN$_4$:chris pratt[3]files for divorce from anna faris	
	Recommended news (RN)	RN$_1$:tori kelly is engaged to basketball player boyfriend u e	RN$_2$:steven innovative co creator of u nypd blue u u hill street blues u dies at	RN$_3$**(ground truth)**:[4] chris pratt and anna faris finalize divorce one year after separating reports	RN$_4$:rita ora kisses cardi b in the new video for controversial track u girls u	RN$_5$:harvey weinstein timeline how the scandal unfolded
User$_2$	**Context news (CN)**	CN$_1$: selena gomez brings a and a bikini to australia u but not justin bieber	CN$_2$:justin bieber selena gomez their time apart is driving him crazy	CN$_3$:justin bieber and selena gomez may have broken up for good this time	CN$_4$:justin bieber s ex baskin champion wows in a bikini amid his engagement to hailey baldwin	
	Recommended news (RN)	RN$_1$**(ground truth)**:selena gomez u s mom responds to justin bieber relationship rumors	RN$_2$:taylor swift s stalker sentenced to year probation and gps monitoring	RN$_3$:celebrities with tattooed eyebrows including helen mirren rooney michelle	RN$_4$:prince harry and harry styles reunite	RN$_5$:kristen bell hosts sag awards in series of gowns see the stunning looks
User$_3$	**Context news (CN)**	CN$_1$: brad pitt he had a blast playing with kids during secret cambodian family reunion	CN$_2$:kim kardashian responds to claims she was attacked in los angeles such weird rumors	CN$_3$:pepsi pulls controversial kendall jenner ad after outcry	CN$_4$:girls cast spoofs golden girls on jimmy kimmel live	
	Recommended news (RN)	RN$_1$**(ground truth)**:brad pitt u s red carpet surprise at u lost city of z u premiere	RN$_2$:the fast food guide	RN$_3$:kesha s mother drops against dr luke	RN$_4$:jason aldean and wife brittany kerr revealed the gender of their baby in the cutest way	RN$_5$:video justin timberlake announces opening act for man of the woods tour u z
User$_4$	**Context news (CN)**	CN$_1$:selena gomez demi lovato bond over boys possible duet more during epic reunion	CN$_2$:real reason behind justin bieber and selena gomez u s breakup has finally been revealed	CN$_3$:poor joe jonas is trying desperately to look like ex gigi hadid s new boyfriend zayn malik	CN$_4$:katie holmes pushing jamie foxx to go more public with their relationship u why he u s u hesitant u	
	Recommended news (RN)	RN$_1$:first look at ryan murphy s new fox series	RN$_2$:video justin timberlake announces opening act for man of the woods tour u z	RN$_3$:jennifer aniston	RN$_4$:best royal wedding gowns of all time	RN$_5$**(ground truth)**: justin s wife, his character may have relationship issues

HDInt: unbiased and true news recommendation



(a) The overall structure of HDInt

(b) The detailed structure of disentangling interest learning module

Results: recommendation accuracy, fairness and true news ratio

Table 2: Comparison of prediction performances with baselines on two datasets, *the improvement is significant at $p < 0.05$.

Dataset	Metric	SKNN	SR-GNN	DAN	NRMS	LSTUR	FedRec	FIM	ESM	Rec4Mit	SentiRec	ProFairRec	HDInt	Imp.(%)
PolitiFact	REC@5	0.2183	0.3729	0.3612	0.4712	0.4725	0.3731	0.3606	0.4608	0.4663	0.4156	0.5312	**0.5781***	8.83
	REC@20	0.6086	0.6703	0.6935	0.8234	0.8056	0.7715	0.7002	0.7964	0.7883	0.7350	0.8454	**0.8594***	1.66
	NDCG@5	0.1417	0.2956	0.2977	0.3386	0.3454	0.2547	0.2279	0.3232	0.3323	0.3211	0.3598	**0.3675***	2.14
	NDCG@20	0.2524	0.3971	0.2699	0.4392	0.4403	0.3528	0.3221	0.4185	0.4261	0.4237	0.4436	**0.4459***	0.52
	PE@5	0.8438	0.7476	0.6330	0.5526	0.5818	0.7060	0.6294	0.7082	0.6508	0.6054	0.7534	**0.8546***	1.28
	PE@20	0.8320	0.6956	0.6222	0.6440	0.6502	0.6982	0.6830	0.7116	0.7192	0.7834	0.8450	**0.8656***	2.44
	FS@5	-0.1469	0.2157	0.3113	0.3449	0.3347	0.2569	0.3261	0.2392	-0.3061	-0.3625	-0.2126	**-0.1273***	15.40
	FS@20	-0.1491	0.2589	0.3231	0.2892	0.2860	0.2591	0.2723	0.2384	-0.2441	-0.1853	-0.1320	**-0.1211***	9.00
	F1@5	0.2427	0.4237	0.4050	0.4199	0.4335	0.3743	0.3346	0.4438	0.4400	0.4196	0.4870	**0.5140***	5.54
	F1@20	0.3873	0.5056	0.3765	0.5222	0.5250	0.4687	0.4378	0.5270	0.5351	0.5500	0.5818	**0.5886***	1.17
GossipCop	REC@5	0.1698	0.4889	0.3212	0.6297	0.66	0.2174	0.3661	0.6745	0.6931	0.4062	0.663	**0.7188**	3.71
	REC@20	0.6249	0.6253	0.4635	0.8158	0.8545	0.4664	0.6135	0.8922	0.9341	0.7344	0.9148	**0.9531***	2.03
	NDCG@5	0.0703	0.4215	0.3177	0.4953	0.5293	0.1449	0.2665	0.4902	0.4957	0.3076	0.4845	**0.5406***	2.13
	NDCG@20	0.2073	0.5602	0.3559	0.5583	0.5560	0.2067	0.3337	0.5894	0.5675	0.3995	0.5585	**0.6103***	3.55
	PE@5	0.5508	0.6256	0.6120	0.5752	0.5522	0.6138	0.6316	0.6282	0.5470	0.6720	0.6566	**0.6968***	3.69
	PE@20	0.5484	0.6142	0.6754	0.6286	0.6246	0.6444	0.6468	0.6386	0.6194	0.6784	0.7324	**0.7358***	0.46
	FS@5	-0.3980	-0.3287	-0.3454	-0.3480	-0.3690	-0.3115	-0.2965	-0.2968	-0.4032	-0.2901	-0.2986	**-0.2231***	30.03
	FS@20	-0.4029	-0.3390	-0.2626	-0.2982	-0.3022	-0.2851	-0.2830	-0.2879	-0.3335	-0.2837	-0.2290	**-0.2263***	1.19
	F1@5	0.1247	0.5037	0.4183	0.5323	0.5405	0.2345	0.3748	0.5507	0.5201	0.4220	0.5576	**0.6088***	9.18
	F1@20	0.3009	0.5860	0.4662	0.5914	0.5883	0.3130	0.4403	0.6130	0.5923	0.5029	0.6337	**0.6672***	5.29

[1] The improvement over the best-performing baseline methods whose performance is underlined. For FS@K, negative and positive values indicate left biased and right biased respectively, and the smaller its absolute value, the more fair the recommendations.

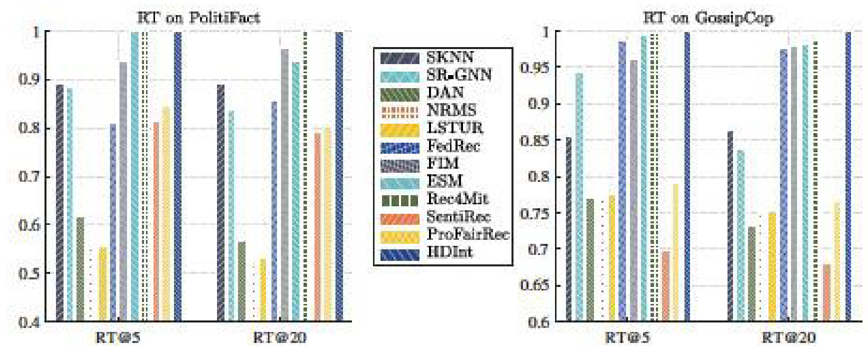


Figure 2: The ratio of true news (RT) in recommendation lists.

Summary

- We have proposed end-to-end frameworks for unbiased, truth-driven personalized news recommendation.
- Experiments on political news and entertainment news on Twitter show their performance in terms of recommendation accuracy, fairness score and true news ratio.

DISCUSSIONS AND CONCLUSION

Discussion: Generative AI can be misused

INNOVATION

AI-Generated Reviews Threaten Business Reputations

Cassio Goldschmidt Former Forbes Councils Member
Forbes Technology Council
COUNCIL POST | Membership (Fee-Based)

POST WRITTEN BY
Cassio Goldschmidt

Apr 4, 2019, 07:00am EDT | 1,459 views

Monday, Volume 21, Number 11 - 7 November 2016

first monday
PEER-REVIEWED JOURNAL ON THE INTERNET

Social bots distort the 2016 U.S. Presidential election online discussion

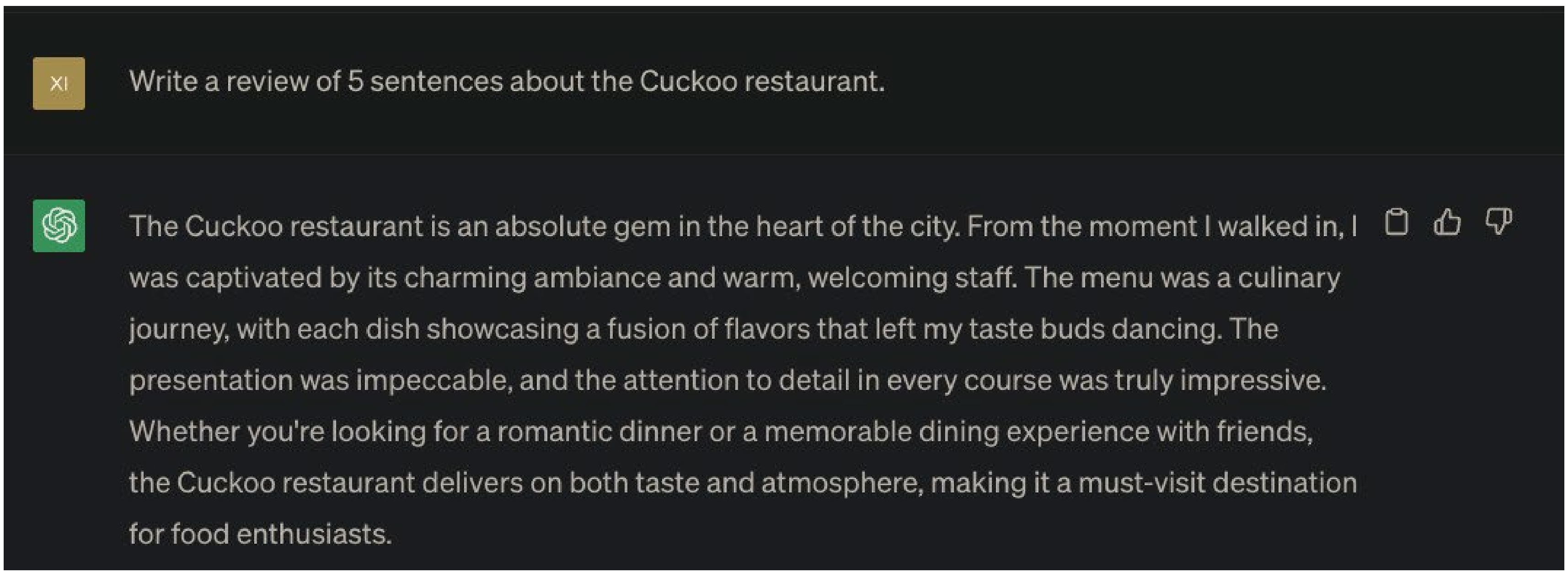
by Alessandro Bessi and Emilio Ferrara

Abstract

Social media have been extensively praised for increasing democratic discussion on social issues related to policy and politics. However, what happens when this powerful communication tools are exploited to manipulate online discussion, to change the public perception of political entities, or even to try affecting the outcome of political elections? In this study we investigated how the presence of social media bots, algorithmically driven entities that on the surface appear as legitimate users, affect political discussion around the 2016 U.S. Presidential election. By leveraging state-of-the-art social bot detection algorithms, we uncovered a large fraction of user population that may not be human, accounting for a significant portion of generated content (about one-fifth of the entire conversation). We inferred political partisanships from hashtag adoption, for both humans and bots, and studied spatio-temporal communication, political support dynamics, and influence mechanisms by discovering the level of network embeddedness of the bots. Our findings suggest that the presence of social media bots can indeed negatively affect democratic political discussion rather than improving it, which in turn can potentially alter public opinion and endanger the integrity of the Presidential election.

How to detect AI-generated contents?

Machine-generated contents are difficult to identify, even for humans.

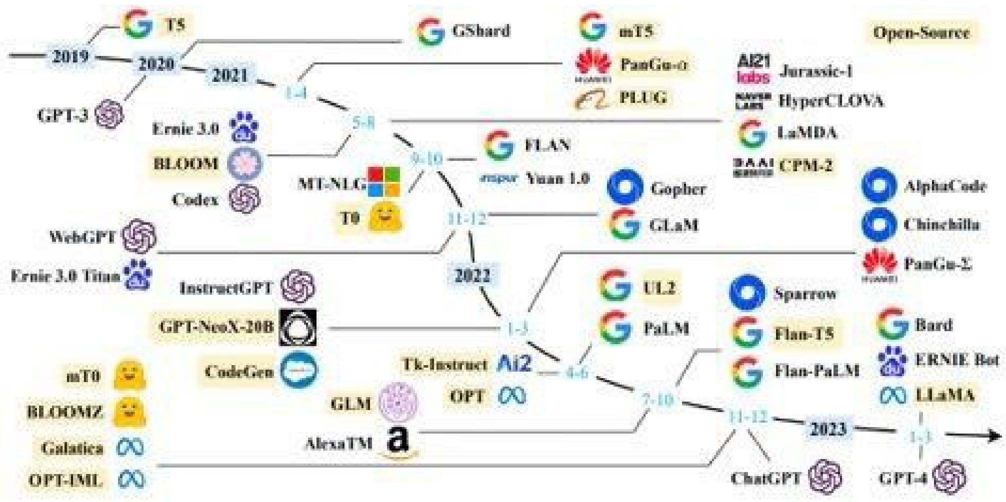


Some preliminary research on detecting AI-generated contents

- Gagiano, R. and Tian, L., 2023. A prompt in the right direction: prompt-based classification of machine-generated text detection. ALTA'2023. To appear.
- Gagiano, R., Fayek, H., Kim, M.M.H., Biggs, J. and Zhang, X., 2023. Automated text identification shared task – Team OD-21. IberLEF 2023. Jaen, Spain.
- Gagiano, R., Kim, M.M.H., Zhang, X.J. and Biggs, J., 2021, December. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association* (pp. 119-127)

A challenge

With the rapid development of generative AI, how to train systems to cope with the novel, generated contents that are out-of-distribution.



Discussion: Can responsible recommendation positively change user information behaviour?

- Xu X., Zhang, X., and Deng, K., 2022. Mirage: An ad-hoc social network for research on responsible information recommendation. https://joinmirage.online/

Vulnerable users

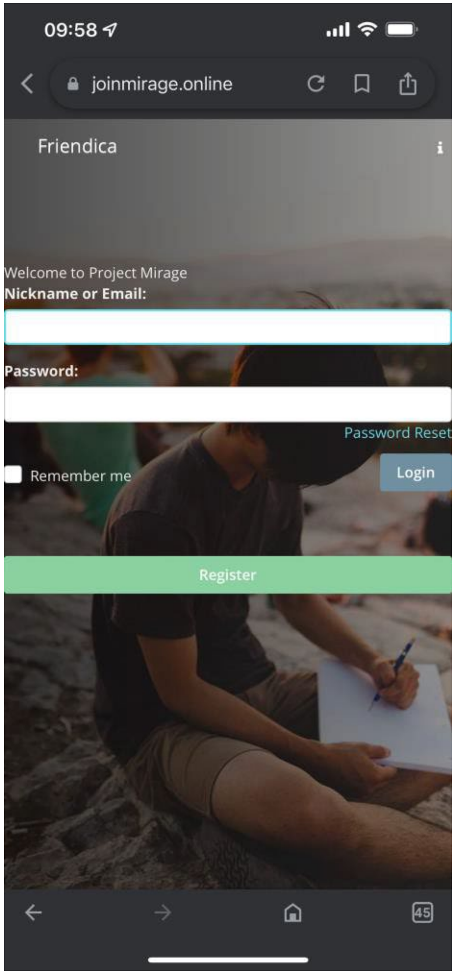
The Social Dilemma: The technology that connects us also controls us

By Kelly Mutizira

The Social Dilemma is a Netflix documentary-drama hybrid that shows how dominant and largely unregulated social media companies manipulate users by harvesting personal data, while using algorithms to push information and ads that can lead to social media addiction, dangerous anti-social behaviour and hate crime. By Kelly Mutizira

A real online information environment

Mirage: https://joinmirage.online/



Conclusions

Trustworthy data science is not a choice but a necessity.

Towards trustworthy data science, research has focused on model transparency and explanability, algorithmic fairness for making automatic decisions, as well as the social impact and responsibility for end users.

The rapid development of generative AI presents unprecedented challenges to data science and requires significant research efforts.

Acknowledgements

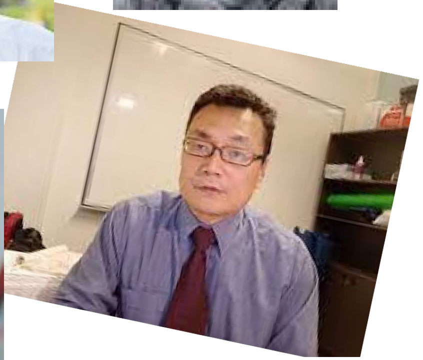












Australian Government
Australian Research Council

Australian Government
Department of Defence
Defence Science and Technology Group

Contact: xiuzhen.zhang@rmit.edu.au

http://www.xiuzhenzhang.org/