

ALTA 2022 tutorial

## Combating misinformation on social media: from detection to mitigation

Xiuzhen (Jenny) Zhang ([xiuzhen.zhang@rmit.edu.au](mailto:xiuzhen.zhang@rmit.edu.au)), RMIT University

Jey Han Lau ([jeyhan.lau@unimelb.edu.au](mailto:jeyhan.lau@unimelb.edu.au)), The University of Melbourne



# Overview

- Part I: Misinformation and human perception (Jenny)
  - Social media is a double-edged sword
  - Information credibility
  - User perception of information perception
- Part II: Misinformation detection (Jey Han, online)
  - Automated Fact Checking
  - Rumour Detection
  - Challenges
- Part III: Misinformation mitigation. (Jenny)
  - Information propagation models
  - Network-level mitigation
  - Personalised mitigation

# Part I: Misinformation and human perception

Xiuzhen Jenny Zhang ([xiuzhen.zhang@rmit.edu.au](mailto:xiuzhen.zhang@rmit.edu.au))

RMIT University

# Contents

- Social media is a double-edged sword
- Information credibility
- User perception

# Social media is more than text-based media

- User generated contents: text posts and comments, photos and videos.
- Social network: connecting users or groups.



# Social media is a double-edged sword

## What is Twitter, a Social Network or a News Media?

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon

Department of Computer Science, KAIST  
335 Gwahangno, Yuseong-gu, Daejeon, Korea  
{haewoon, chlee, hosung}@an.kaist.ac.kr, sbmoon@kaist.edu

**Science** Home News Journals Topics Careers

Institution: RMIT UNIVERSITY | Log in | My account | Contact Us

**Become a member** Renew my subscription | Sign up for newsletters

**REPORT**

**The spread of true and false news online**

Soroush Vosoughi<sup>1</sup>, Deb Roy<sup>1</sup>, Sinan Aral<sup>2,\*</sup>

<sup>1</sup>Massachusetts Institute of Technology (MIT), the Media Lab, E14-526, 75 Amherst Street, Cambridge, MA 02142, USA.  
<sup>2</sup>MIT, E62-364, 100 Main Street, Cambridge, MA 02142, USA.  
\*Corresponding author. Email: sinan@mit.edu

0 - Hide authors and affiliations

Science 09 Mar 2018:  
Vol. 359, Issue 6380, pp. 1146-115  
DOI: 10.1126/science.aap9559

**World Health Organization**

**Science** Vol 359, Issue 6380 09 March 2018  
Table of Contents  
Print Table

**POLICY FORUM** SOCIAL SCIENCE

**The science of fake news**

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Fili

+ See all authors and affiliations

Science 09 Mar 2018:  
Vol. 359, Issue 6380, pp. 1094-1096  
DOI: 10.1126/science.aao2998

**Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation**

Home / News / Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation

العربية 中文 Français Русский Español



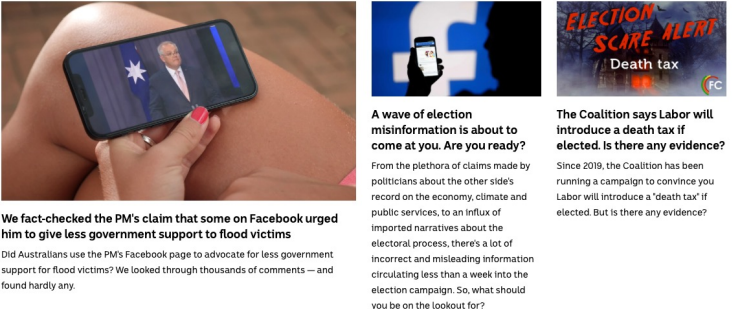
# What is misinformation?

- “Misinformation is incorrect or misleading information” – Merriam-Webster Dictionary.
- Types of misinformation: false claims, rumours, fake news, and disinformation.

**RMIT ABC Fact Check**

**AUSTRALIA VOTES** Keep up with the latest from the campaign trail in our federal election live blog < 1/2 >

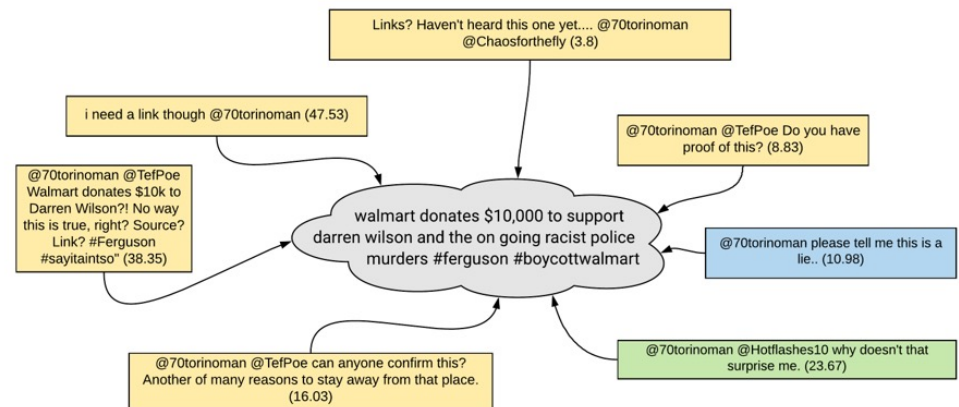
**Fact Check** About RMIT Archive Contact



**We fact-checked the PM's claim that some on Facebook urged him to give less government support to flood victims**  
Did Australians use the PM's Facebook page to advocate for less government support for flood victims? We looked through thousands of comments — and found hardly any.

**A wave of election misinformation is about to come at you. Are you ready?**  
From the plethora of claims made by politicians about the other side's record on the economy, climate and public services, to an influx of imported narratives about the electoral process, there's a lot of incorrect and misleading information circulating less than a week into the election campaign. So, what should you be on the lookout for?

**The Coalition says Labor will introduce a death tax if elected. Is there any evidence?**  
Since 2019, the Coalition has been running a campaign to convince you Labor will introduce a 'death tax' if elected. But is there any evidence?



# Understand user perception of social media information credibility

- “Credibility is **the quality of being believed** or accepted as true, real, or honest, whether it regards the information or the source”. (Tseng & Fogg, 1999).
- Judgement of information credibility is a type of information behaviour.



11 users have different credibility ratings:

- Very credible - 6
- Somewhat credible - 1
- Not credible - 3
- Cannot decide - 1



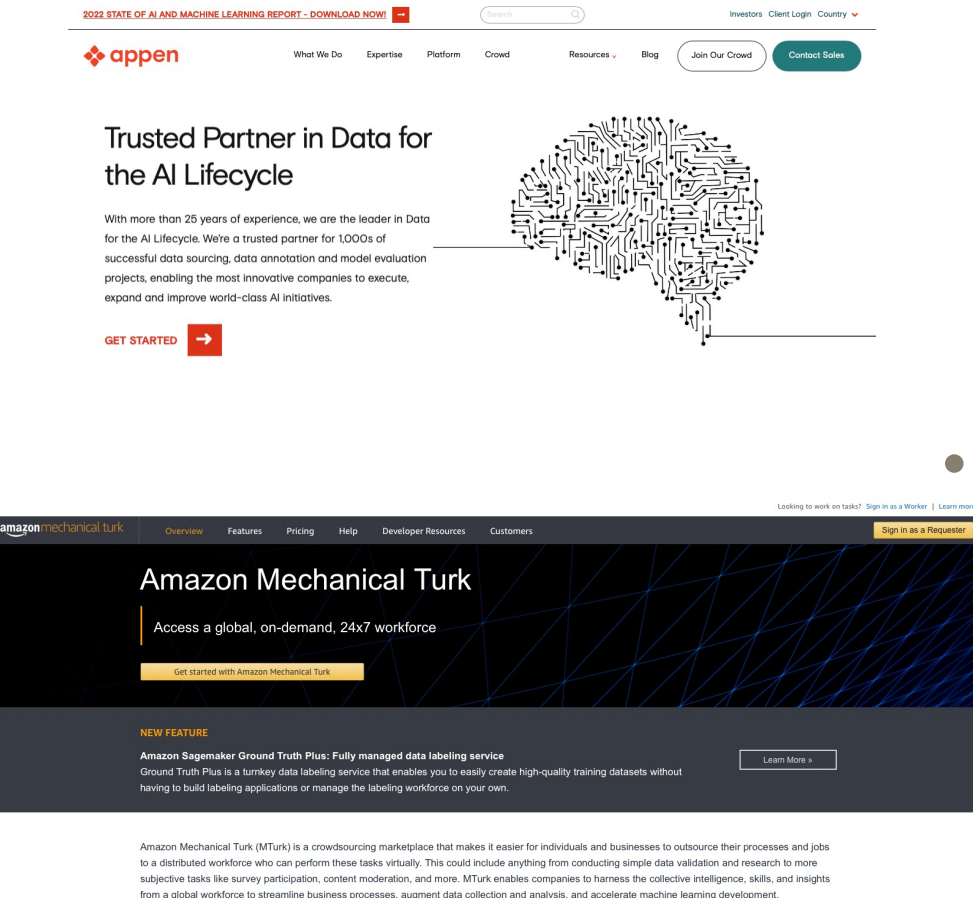
# Research questions

- Do users' demographic attributes significantly correlate with their judgement of information credibility?
- What features users use to judge social media information credibility?
- Are there confounding factors for information credibility judgement?
- Are humans credulous (more than machines) about social media information?

# Information credibility for news on Twitter

- Twitter news data: keyword search on Twitter
- User demographic data: questionnaires on crowd sourcing platform
- User credibility ratings: crowd source user study
- Data analysis: multi-way chi-square analysis and association analysis

# User study on crowdsourcing platforms



- Ensuring data quality:
  - Specify qualification level of crowd workers.
  - Gold questions were used to ensure data quality.
- Filter crowd workers. 40 (2\*20) Gold questions are randomly inserted as gold questions.

# User demographics: gender, age, education, country.

754 users(crowd workers) from 76 countries. Demographic attributes grouped into binary and categorical levels.

- Age: binary (young, old); categorical (Boomer, Gen X, Gen Y, Gen Z).
- Education: binary (below and above university); categorical (school, non-degree tertiary, undergrad and postgraduate)
- Location: binary (Eastern and Western); categorical (Asia-pacific, America, Europe, Africa)

**Table 1**  
Demographic profiles distribution.

Demographic	Value	#	%
Gender	Male	521	69.2
	Female	233	30.8
Age	16-19 years old	58	7.7
	20-29 years old	327	43.4
	30-39 years old	243	32.2
	40-49 years old	89	11.8
	50 years and older	37	4.9
Education	High school	127	16.8
	Technical training	58	7.7
	Diploma	81	10.7
	Bachelor's degree	287	38.1
	Master's degree	137	18.2
	Doctorate degree	14	1.9
Location	Professional certification	50	6.6
	Asia	275	36.5
	Europe	247	32.8
	South America	130	17.2
	North America	65	8.6
	Africa	37	4.9

# Twitter posts

- 1510 news tweets for 20 news topics from major online newswires.
  - BBC, Reuters, CNN, Guardian and New York Times.
  - Topic keywords are used to search for relevant Twitter posts. Examples: “US government shutdown”, “Navy Yard shooting”, “Earthquake in Pakistan”.
- Three types of news -- breaking news, political news and natural disaster news -- in 4 years.
- Trending and non-trending news topics.
- Two writing styles: opinion/emotion-bearing and factual.

**Table 2**  
Tweets news attributes distribution.

News attribute	Value	#	%
News type	Breaking news	509	33.8
	Natural disaster	500	33.2
	Politic	499	33.0
Year	2011	374	24.8
	2012	375	24.9
	2013	377	25.0
	2014	382	25.3
Trending	Trending	781	51.8
	Not trending	727	48.2

# Credibility ratings

- Four credibility rating levels:
  - Definitely credible
  - Somewhat credible
  - Definitely Not credible
  - Cannot decide
- Exclude the middle level “maybe credible” option to avoid “lazy” judgements.
- Free text data to collect features.

### Distinguishing Credibility Level Of A Tweet Finished

**Instructions**

In this task you will be given a set of relevant tweet messages from Twitter regarding a newsworthy topic in the area of politics, news and disaster/emergency events. You will need to indicate a level of credibility of the tweet messages.

- Credibility definition: “offering reasonable grounds for accepting truth or false”.
- The criteria of a credible newsworthy tweet would meet the following requirements:
  - o affirm a fact or something that really happened o informative and interesting to the public, not only for friends o not based purely on personal / subjective opinions o not a conversation among friends
- Credibility level: “Definitely credible”, “Seems credible”, “Definitely not credible” and “Can’t decide”

---

**Topic:** US government shutdown

**Topic Definition:** US Government Heads Toward a Shutdown

**Date Tweeted:** 01/10/13

**Twitter ID:** 4Falloy

**Tweet:** US #govtshutdown impasse continues: Protesters clash with police theguardian.com/world/video/2013/oct/14/us-gove...  
“We are not Democrats or Republican, we are Americans

**Choose only one:**

- ☐ Definitely credible
- ☐ Seems credible
- ☐ Definitely not credible
- ☐ Can't decide

Please explain reasons for your choice. For example: the tweet source (twitter ID), URL links (reliable external source), hashtags, retweet messages, keywords (please define them), message structure or others (please specify). We need this to validate your judgement.



---

**Topic:** Mexico storm disaster

**Topic Definition:** Mexico storms death toll rises, crop lands damaged

**Date Tweeted:** 12/10/2013

**Twitter ID:** ABC News

**Tweet:** At least five dead, 550,000 evacuated as Cyclone #Phailin crashes into India's eastern coast <http://ab.co/1hLKKXQo>

**Choose only one:**

- ☐ Definitely credible
- ☐ Seems credible
- ☐ Definitely not credible
- ☐ Can't decide

Please explain reasons for your choice. For example: the tweet source (twitter ID), URL links (reliable external source), hashtags, retweet messages, keywords (please define them), message structure or others (please specify). We need this to validate your judgement.



# User reported features for credibility judgements

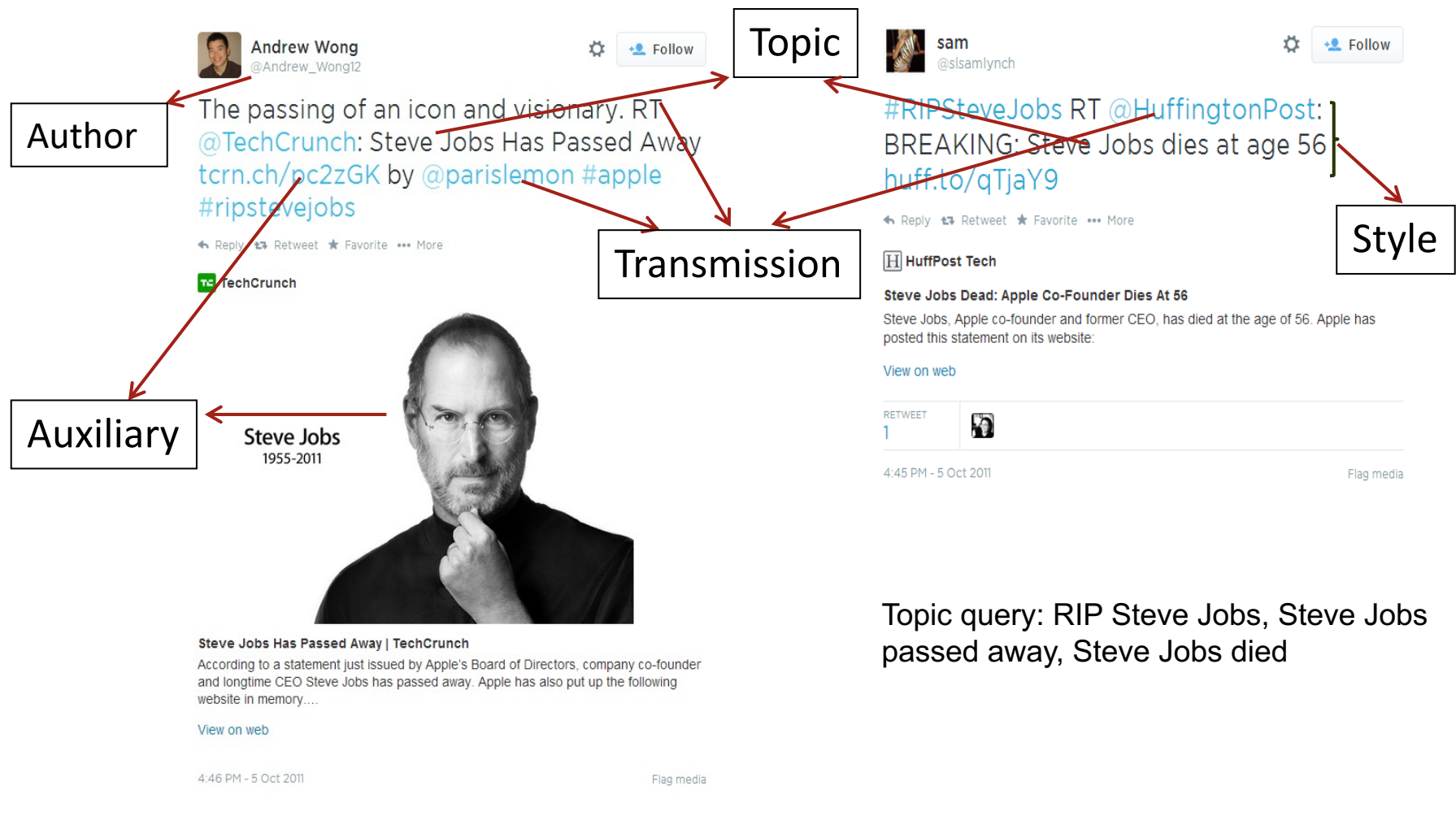
## Summative content analysis of textual data (Hsieh, 2005)

**Table 3**

Features reported by readers to judge credibility for news tweets.

Category	Feature	Description
Author	Tweet author	Twitter ID or display name e.g. Sydneynewsnow
Transmission	User mention	Other Twitter user's Twitter ID mentioned in the tweet starting with the @ symbol e.g. @thestormreports
	Hashtag	The # symbol used to categorise keywords in a tweet e.g. #Pray4Boston
	Retweet	Contain the letters RT (retweet) in the tweet and the retweet count
Auxiliary	Link	Link to outside source - URLs, URL shortener
	Media	Picture or video from other sources embedded within the tweet
Topic	Alert phrase	Phrase that indicates new or information update regarding a news topic - e.g. Update
	Topic keyword	The search keyword regarding a news topic e.g. Hurricane Sandy
Style	Language	The language construction of the tweet (formal or informal English)
	Author's opinion	Tweet that conveys the author's emotion or feeling towards the news topic
	Fact	Factual information on the tweet regarding the news topic

# A user study on features for users to judge credibility of news on Twitter



# Correlation analysis of user demographics and credibility ratings

- Correlation analysis (chi-square) of user demographics at original, binary and categorical levels.
- Only location is significantly correlated with credibility ratings at all levels. The African-user –cannot-decide combination gives the highest dependence.
- Education is strongly correlated with credibility judgement. Further analysis (correlation rule, Brin et al, 1997) reveals that strong dependence originates from the group of users with a professional certification giving “not credible” rating.

**Table 5**  
Demographic profiles and credibility perception chi-square results.

Demographic	Data setting	Credibility	
		$\chi^2$	<i>p-value</i>
Gender	Original	1.51	0.68
	Binary	1.51	0.68
	Categorical	1.51	0.68
Age	Original	14.87	0.25
	Binary	4.68	0.20
	Categorical	9.84	0.13
Education	Original	49.43	9.20E-5
	Binary	4.78	0.19
	Categorical	12.29	0.20
Location	Original	80.79	2.92E-12
	Binary	39.62	1.29E-8
	Categorical	80.33	1.39E-13

# Correlation analysis of news attributes and credibility

- News-type correlates with Credibility, and strong dependence for “breaking news” and “very credible”.
- Trending correlates with Credibility, and strong dependence for “trending” and “very credible”.

**Table 7**

News attribute correlation with reader's credibility perception.

News attribute	Credibility	
	$\chi^2$	<i>p-value</i>
News type	93.75	5.04E-18
Year	61.89	5.78E-10
Trending	8.09	0.04

# Correlation analysis of user demographics and news attributes

- User-age correlates with news-year.
- User-location correlates with news-type.
- User-location correlates with news-year.

**Table 9**

Correlation between combination of reader's demographics and news attributes with credibility perception.

Demographic	News attribute	Credibility	
		$\chi^2$	<i>p</i> -value
Gender	News type	6.94	0.33
	Year	8.43	0.49
	Trending	7.38	0.06
Age	News type	35.53	0.06
	Year	53.06	0.03
	Trending	18.59	0.10
Education	News type	47.81	0.09
	Year	64.56	0.15
	Trending	16.92	0.53
Location	News type	38.35	0.03
	Year	55.16	0.02
	Trending	17.17	0.14

# User perception versus machine prediction for credibility



(a)



(b)

Gupta, Aditi, et al. "Tweetcred: Real-time credibility assessment of content on twitter." *SocInf* 14.



# Human perception vs. machine prediction for credibility

Humans are more credulous than machines on the credibility of news tweets.

*Table 4.4: The agreement matrix between reader's credibility perception and automated credibility prediction*

		TweetCred			
		Very credible	Somewhat credible	Not credible	Total
Readers	Very Credible	256	654	67	977
	Somewhat credible	51	230	50	331
	Not credible	1	4	3	8
	Total	308	888	120	1316

# User demographics and credibility features

- All demographic attributes are somewhat correlated with credibility features.
- Demographic attributes are most correlated with topic and style features while least correlated with transmission features.
- Association analysis shows that Author and Auxiliary and Transmission are frequently used together.

**Table 11**

The chi-square correlation between demographics and features used in credibility perception.

Demographic	Feature categories				
	Author ( $\chi^2$ )	Topic ( $\chi^2$ )	Style ( $\chi^2$ )	Auxiliary ( $\chi^2$ )	Transmission (p)
Gender	0.01	***18.15	***23.27	1.59	0.59 <sup>a</sup>
Age	***16.63	***26.65	***41.99	8.65	1.00 <sup>a</sup>
Education	11.12	***31.87	***50.12	**16.53	*0.03 <sup>a</sup>
Location	***46.87	***83.81	***67.35	***13.60	1.00 <sup>a</sup>

\*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001.

<sup>a</sup> Calculated using Fisher's Exact Test.

# Looking ahead

- Strategies for publishing effective, truthful information to convince end users.
  - Official news media, fact-checking services
- Insights for improving machine learning systems for better information service.
  - Location-based information service.
- What about machine generated misinformation?
  - Can machines generate seem-credible information to deceive social media users?
  - Do the features in machine generated texts affect human perception in the same way?

# Acknowledgements



# References

- Castillo, C., Mendoza, M. and Poblete, B., 2011, March. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675-684).
- Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P., 2014, November. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics* (pp. 228-243). Springer, Cham.
- Morris, M.R., Counts, S., Roseway, A., Hoff, A. and Schwarz, J., 2012, February. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*(pp. 441-450)
- Shariff, S. M., Zhang, X., & Sanderson, M. 2017. On the credibility perception of news on Twitter: Readers, topics and features. *Computers in Human Behavior*, 75, 785-796.
- Shariff, S. M., Sanderson, M., & Zhang, X. 2016. Correlation analysis of reader's demographics and tweet credibility perception. In *Proc. ECIR 2016*.
- Shariff, S. M., Zhang, X., & Sanderson, M. 2014. User perception of information credibility of news on Twitter. In *Proc. ECIR 2014*.



Contact: [xiuzhen.zhang@rmit.edu.au](mailto:xiuzhen.zhang@rmit.edu.au)

<http://www.xiuzhenzhang.org/>