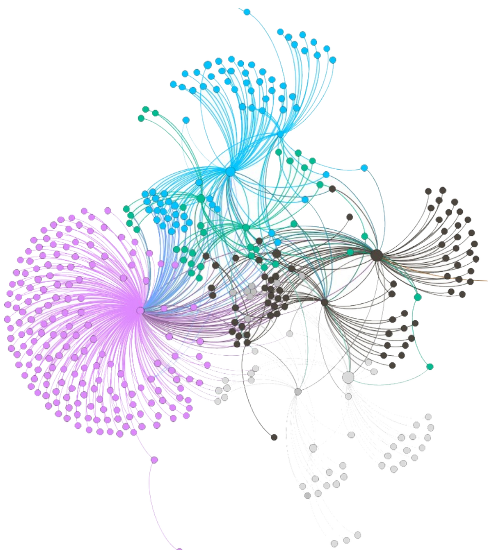
Behavioral Data Science

UTS Data Science

UTS
UNIVERSITY OF TECHNOLOGY SYDNEY

NATURAM PRIMUM COGNOSCERE RERUM



Profiling information warfare on social media: The anatomy of a scare disinformation campaign

Marian-Andrei Rizoiu

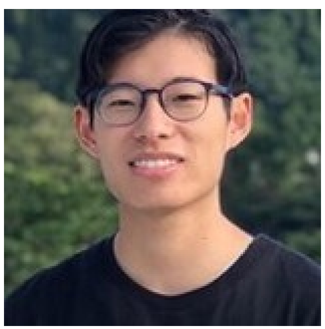
The research group

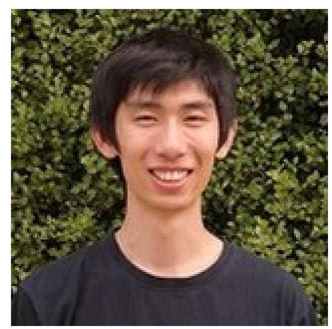
Behavioral Data Science

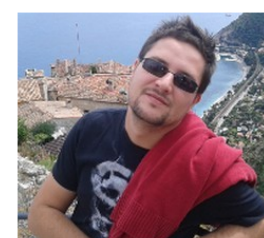
5 PhD students, 4 Honors students, 1 lecturer





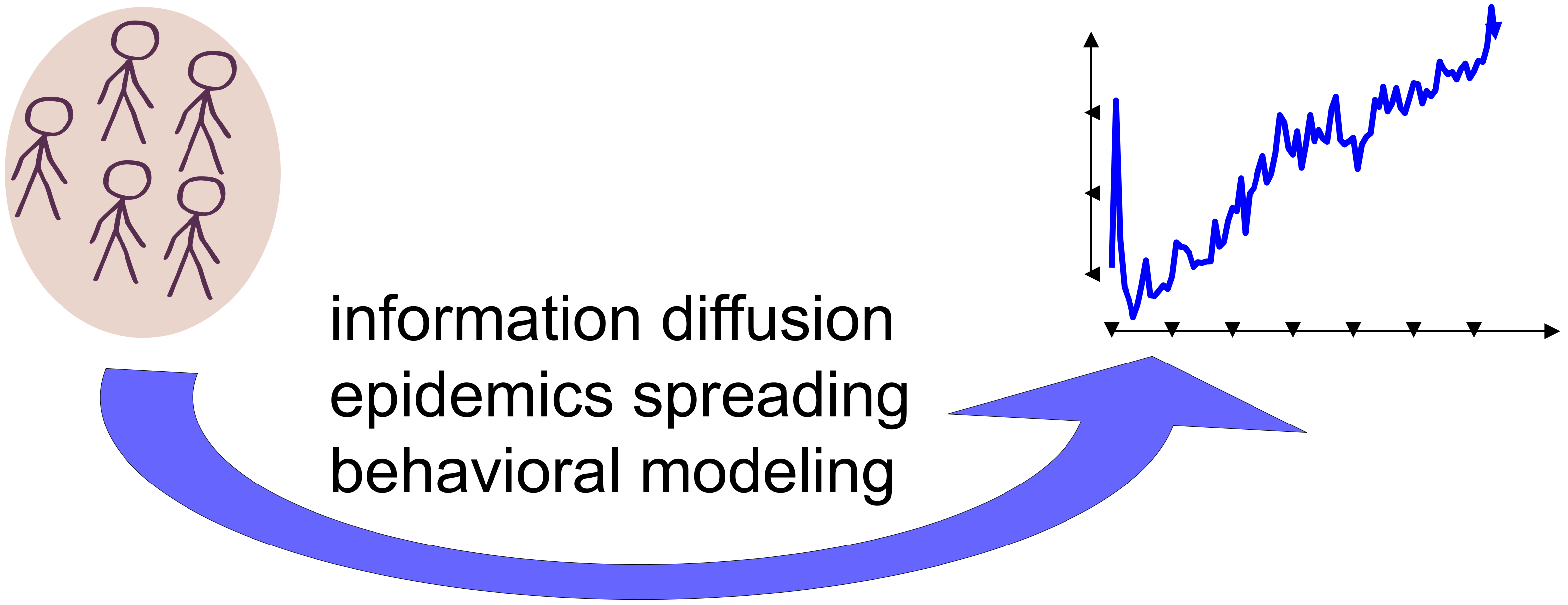







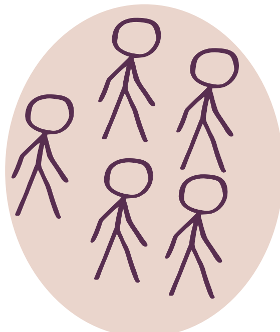
Research objectives

1.

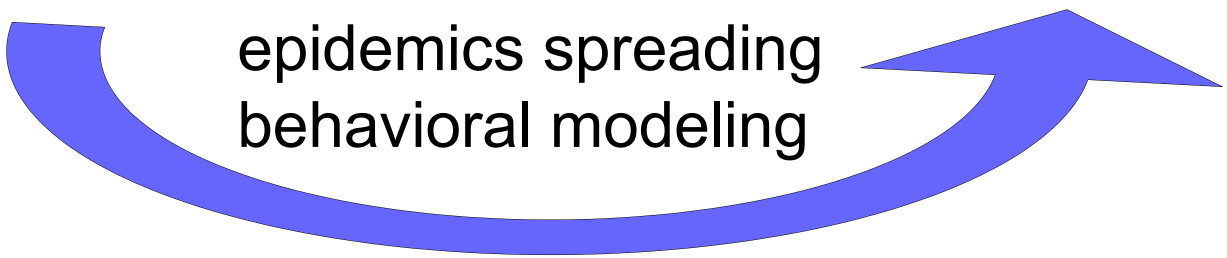


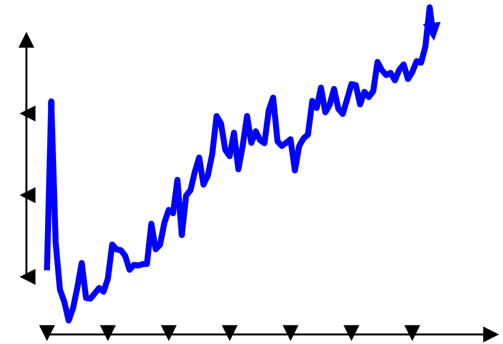
Research objectives

1.

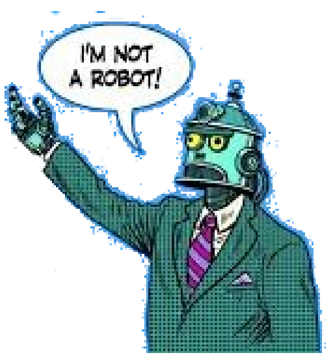


information diffusion
epidemics spreading
behavioral modeling





2.

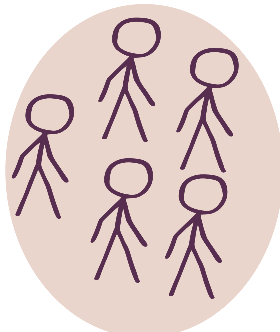


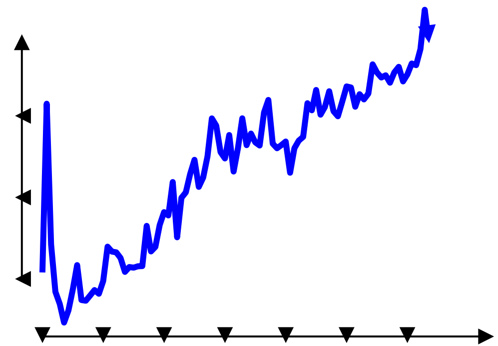
[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

Research objectives

1.



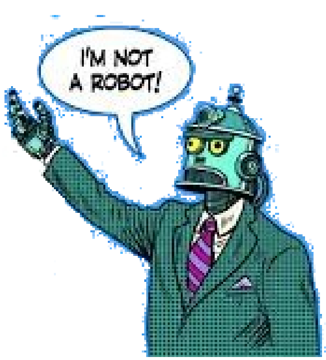


information diffusion
epidemics spreading
behavioral modeling

3.

[Rizoiu et al IJCAI'20]

2.



[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

Prior expertise

Australian National University
CRAWFORD SCHOOL OF PUBLIC POLICY

Tracking Disinformation Campaigns

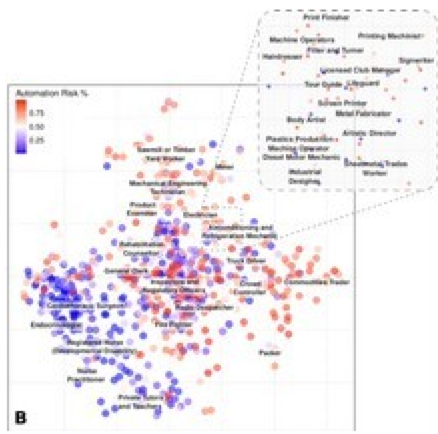
Australian Government
Department of Defence
Defence Science and Technology Group

Opinion manipulation and information warfare

Hate Speech propagation on Social Media

NSW GOVERNMENT

Expert roundtable for Defamation law reform



Occupation transition recommender systems

WIKIMEDIA FOUNDATION

WIKIPEDIA
The Free Encyclopedia

Detecting and quantifying privacy loss over time

The team

Behavioral Data Science

Thomas Willingham
Honors student – Computer Science & Engineering, ANU
ASD-ANU co-Lab scholarship

Kriti Tripathi
Honors student – Computer Science & Engineering, ANU
ASD-ANU co-Lab scholarship

Jennifer Hunt
Lecturer – Security Studies at Macquarie University

Marian-Andrei Rizoiu
Lecturer – UTS Data Science Institute, UTS

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

ASD AUSTRALIAN SIGNALS DIRECTORATE

MACQUARIE University

Australian National University
CRAWFORD SCHOOL OF PUBLIC POLICY

Fallacy #1:
There is no foreign intervention in Australian politics.

Fallacy #2:
The Australian democratic system is immune to the spread of fake news.

The 2019 Australian elections?

death tax: Labour intends to institute a tax on inheritance

To vote correctly put a number 1 next to Liberal

7NEWS Melbourne - St Kilda gang rampage | Faceb...

St. Kilda is terrorized by African gangs

Overview

1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
5. Future Work

The *#DeathTax* incident

Michael Tiyce
@MichaelTiyce

Shorten adopting the Trumpist "fake news" is utter garbage. Check the source of the death tax issue Labor is running from. In their own words #auspol #DEATHTAX

BK32 🦈🐯🏆🏆🇦🇺
@BK6785

Work hard all your working life and upon your passing Your loved ones are then hit with a Labor Union approved Inheritance #DeathTax Bill Taking almost half of what you've left to them Only Labor would tax your passing #auspol #AusVotes19 #7news #9news #qldpol #nswpol #springst

Google search trends for **death tax**

THE HON JOSH FRYDENBERG MP
Treasurer

MEDIA RELEASE

24 January 2019

DEATH TAXES – YOU DON'T SAY, BILL!

Facing growing pressure over Labor's disastrous housing and retirees taxes, Bill Shorten today sought to deflect attention by flippantly remarking that the next thing they say will be "that Labor wants to introduce death taxes."

The *#DeathTax* incident

Michael Tiyce
@MichaelTiyce

Shorten adopting the Trumpist “fake news” is utter garbage. Check the source of the death tax issue Labor is running from. In their own words #auspol #DEATHTAX

BK32
@BK6785

Work hard all your working life and upon your passing Your loved ones are then hit with a Labor Union approved Inheritance #DeathTax Bill Taking almost half of what you’ve left to them Only Labor would tax your passing #auspol #AusVotes19 #7news #9news #qldpol #nswpol #springst

Google search trends for **death tax**

THE HON JOSH FRYDENBERG MP
Treasurer

MEDIA RELEASE

24 January 2019

DEATH TAXES – YOU DON’T SAY, BILL!

Facing growing pressure over Labor’s disastrous housing and retirees taxes, Bill Shorten today sought to deflect attention by flippantly remarking that the next thing they say will be “that Labor wants to introduce death taxes.”

Issues:

- Social media is weaponized
- Fake news spreads rapidly
- Misinformation has flown into the traditional media

Dataset

- Twitter dataset:
  2019 Election Period
- Crawled using hashtag
  #auspol
- Big dataset:
  17 M+ tweets
- Content and user info:
  author, time stamp, title, etc.

Four step approach

1. Instances of election misinformation on Twitter
2. Forensic analysis:
   a) The social network of the discussions
   b) Themes and messages
   c) Characterize authors and opinion leaders
3. Building content-based classifiers
4. Spill of misinformation into traditional media

Overview

1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
5. Future Work

Step 2: Forensic analysis

1. Map relationship between all authors

2. Analyze the cluster narratives and identifying different types of clusters

3. Characteristics of clusters of users

E.g. Is a particular cluster more strongly connected than the others? (network science measures)

Step 2: Forensic analysis

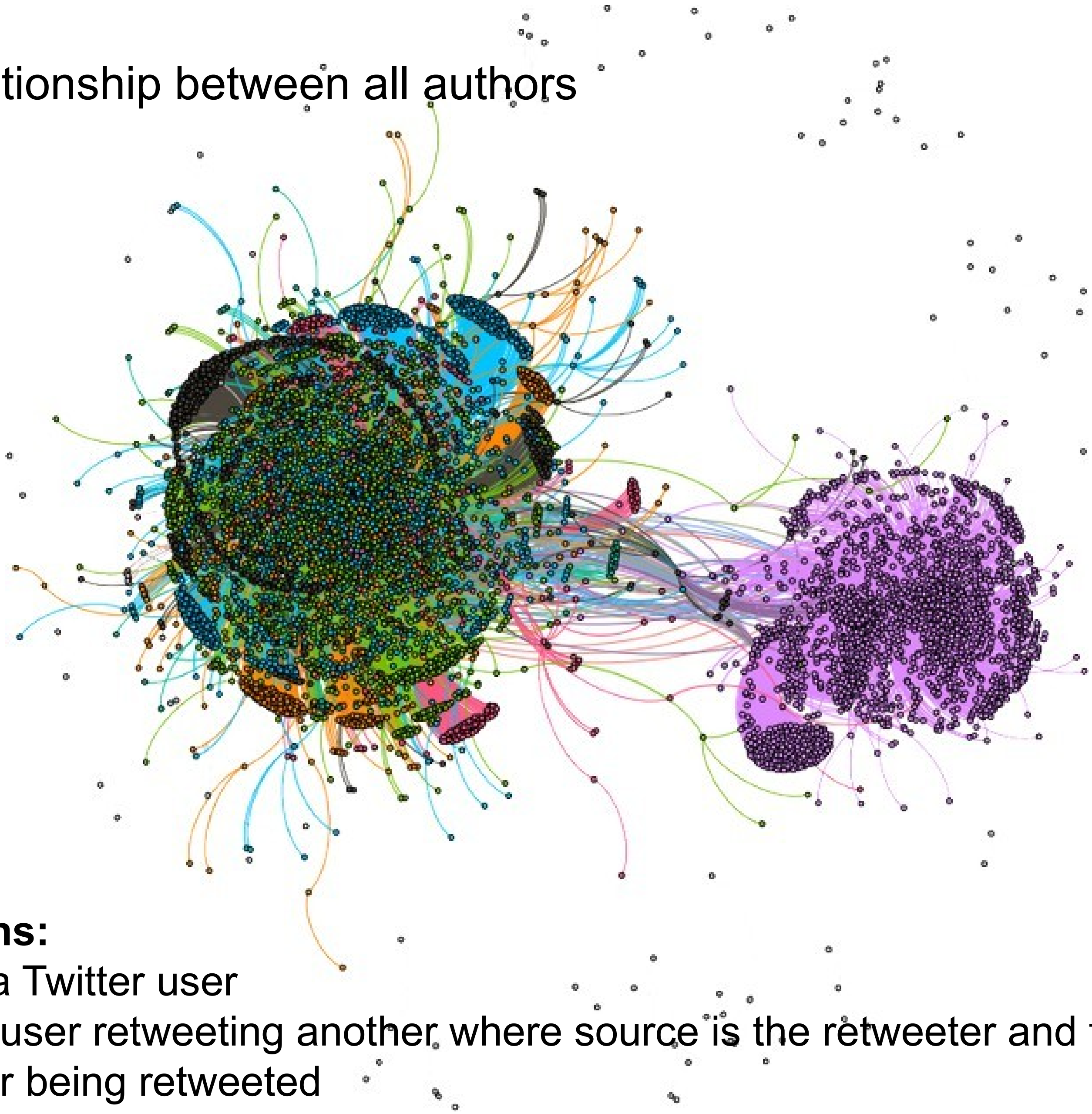
4. Analyze the potential exposure gained by the clusters over time

5. Identify the opinion leaders of the clusters

6. Explore their characteristics
   - Number of followers and friends
   - Geotags/location
   - Verified status

The retweet network of #DeathTax

1. Map relationship between all authors



Observations:

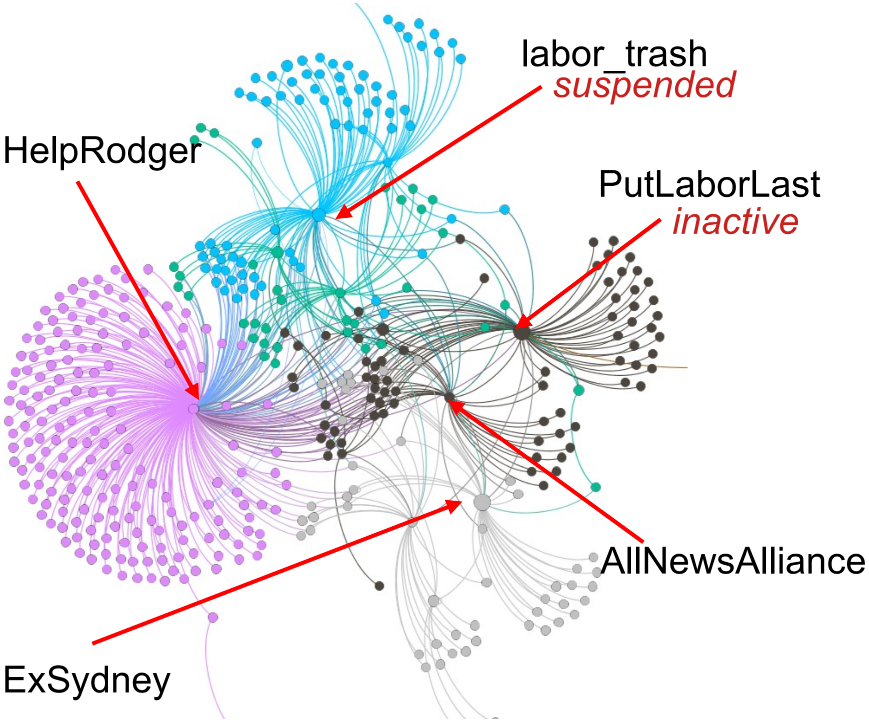
- Nodes – a Twitter user
- Edge – a user retweeting another where source is the retweeter and target is the user being retweeted

Two clusters emerge

Observations:

- Two clusters – one misinformation and one debunking the misinformation
- Misinformation cluster (left) is strongly connected compared to debunking cluster (right)

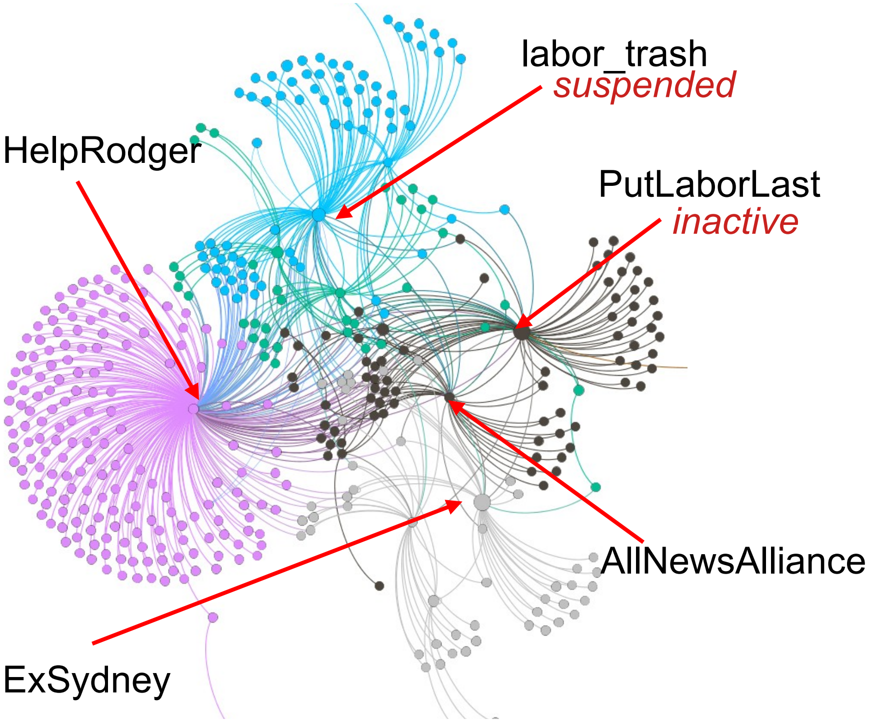
Two clusters emerge

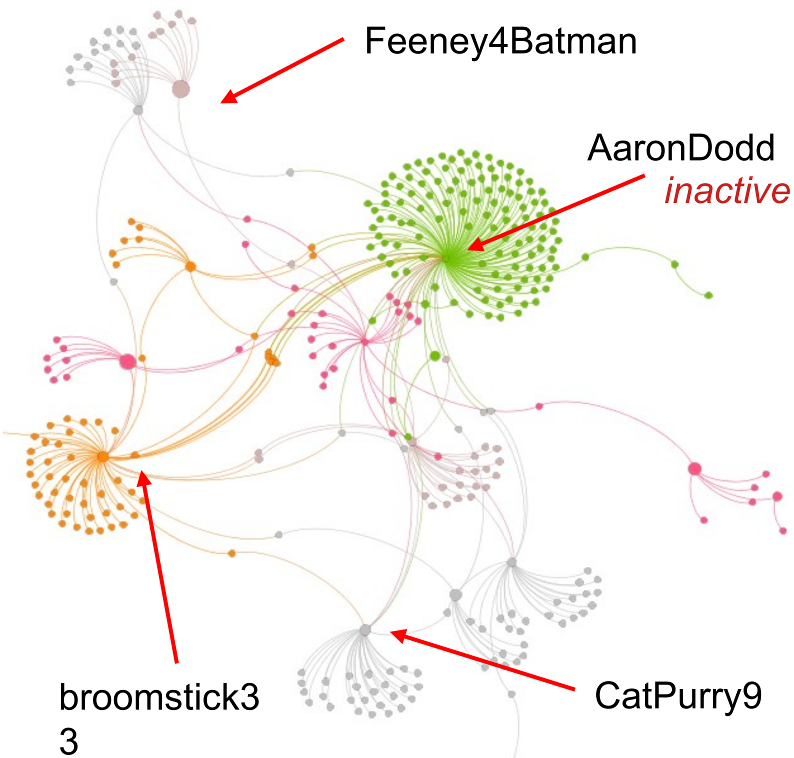


Observations:

- Two clusters – one misinformation and one debunking the misinformation
- Misinformation cluster (left) is strongly connected compared to debunking cluster (right)

Two clusters emerge





Observations:

- Two clusters – one misinformation and one debunking the misinformation
- Misinformation cluster (left) is strongly connected compared to debunking cluster (right)

Analyzing narratives (1)

Misinformation cluster

☻ Ramjet ☻
@HelpRodger

#BREAKING
Tax expert has verified that a Shorten Labor party #DEATHTAX means all valuable items such as gold teeth , sentimental jewellery, anything of value will be taxed 40% on death.
I cant believe Labor would do this to families. #auspol #qanda #insiders #9Today #sunrise

☻ Ramjet ☻
@HelpRodger

#BREAKING
Australians are finally aware of Bill Shortens #DeathTax
There is nothing more abhorrent than a Labor govt robbing dead ppls graves.
I'm sick to the stomach thinking about this.#auspol #qanda #insiders #730Report #9Today #sunrise #TheProjectTV #MKR #livingroom #60mins

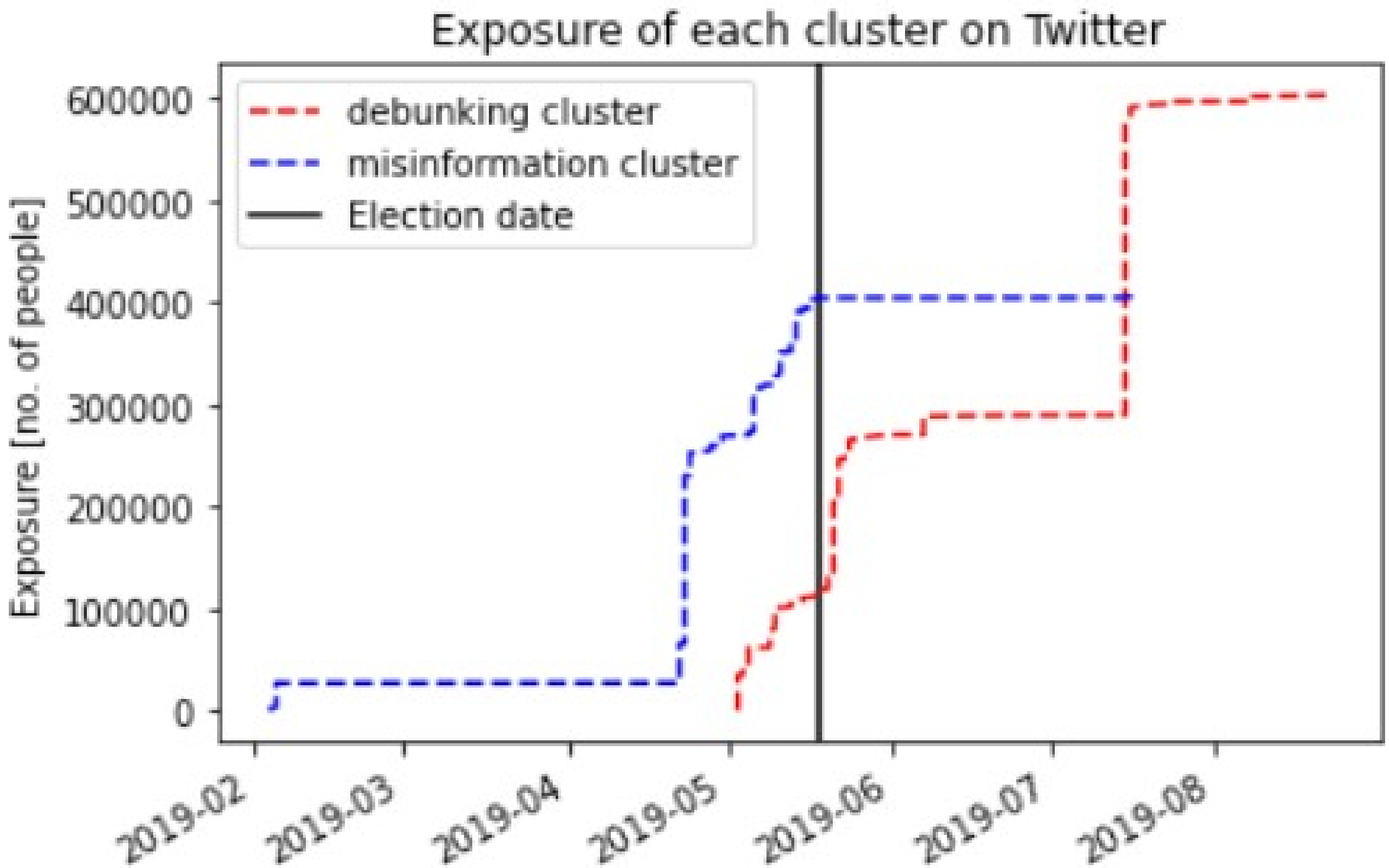
- Perpetuating the misinformation
- Nazi references – "sentimental jewellery", "gold teeth", etc.
- Spiteful phrases – "lose parents", "worth dead" and "truly screwed"
- Confirmative language – "experts verified"

Analyzing narratives (2)

Debunking cluster

- Based around debunking the #DEATHTAX myth
- Phrases like “fabricated”, “completely false”, “scare campaign”

Exposure analysis



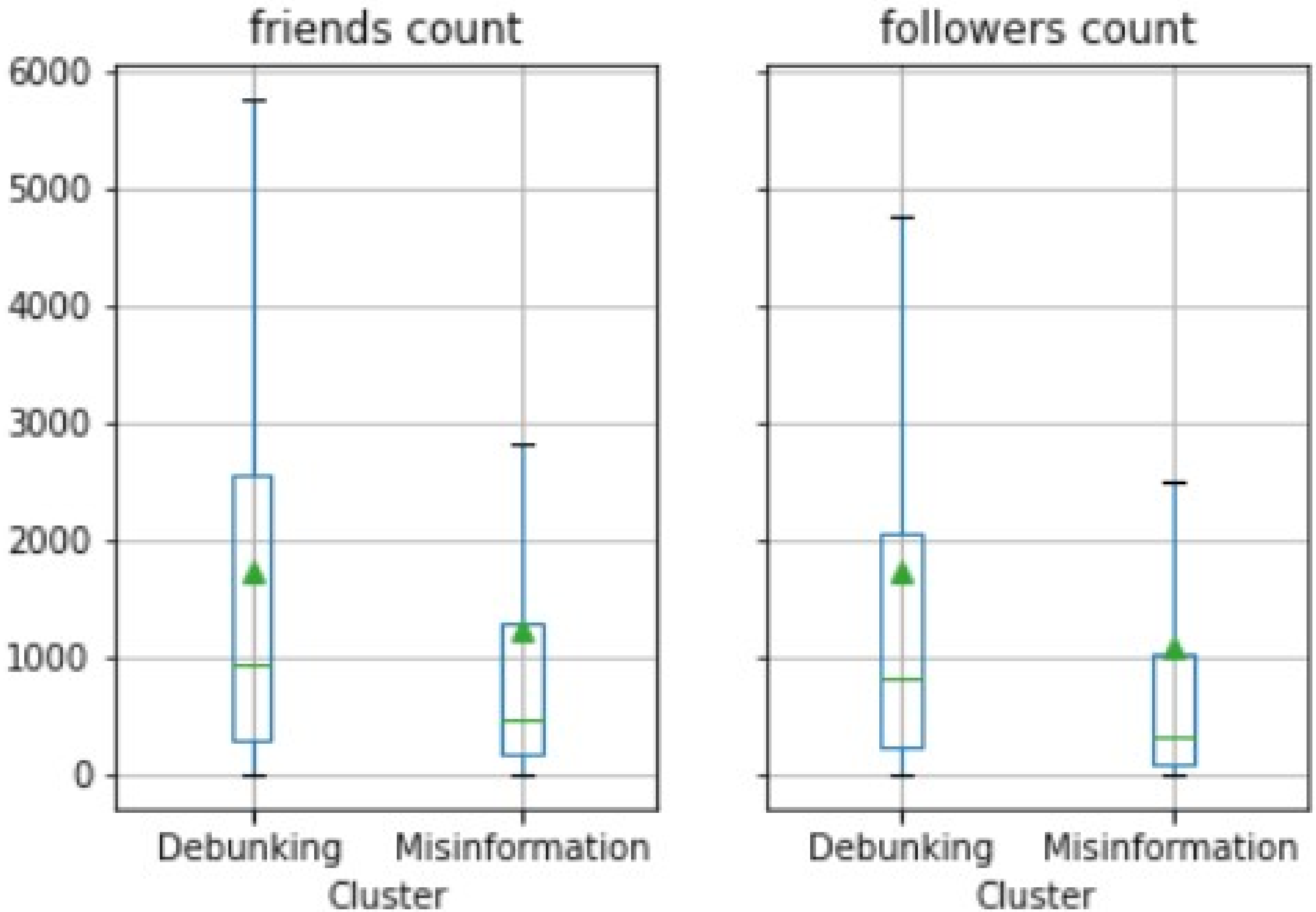
- Early spread of the misinformation campaign
- Exposure of misinformation cluster flattens subsequent to election date
- Debunking cluster gains traction after the election

Account-level analysis

	Misinformation Cluster (%)	Debunking Cluster (%)
Suspended Accounts	15 %	2.7 %
Deactivated Accounts	6.2 %	3.1 %
Total	**21.2 %**	**5.8 %**

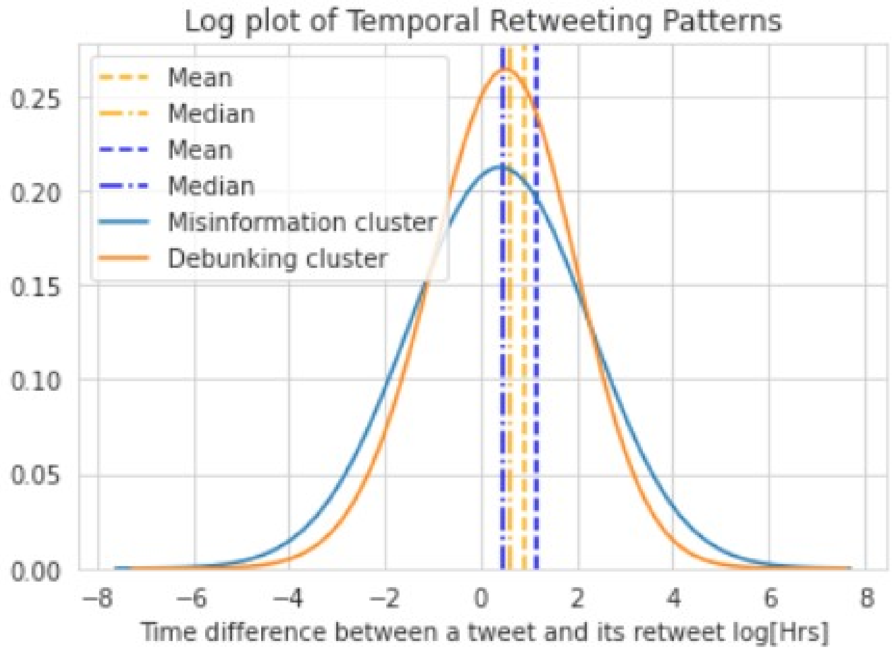
Nearly ¼ of the users in the misinformation cluster are currently not active (suspended or deactivated).

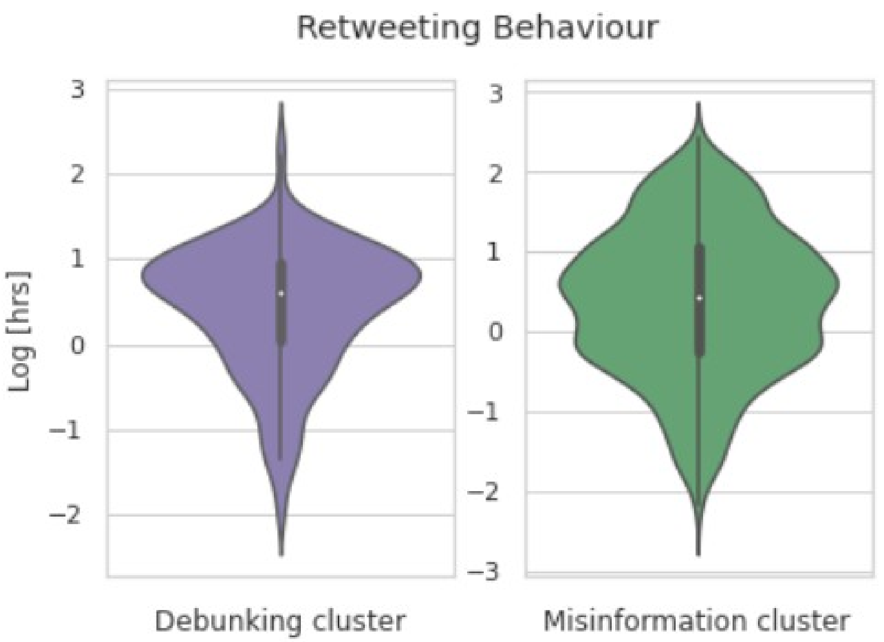
Account-level analysis



Debunking cluster – higher average number of friends and followers than the misinformation cluster.

Account-level analysis





Misinformation cluster have higher probability of retweeting very quickly (seconds) and very late (days) compared to the debunking cluster.

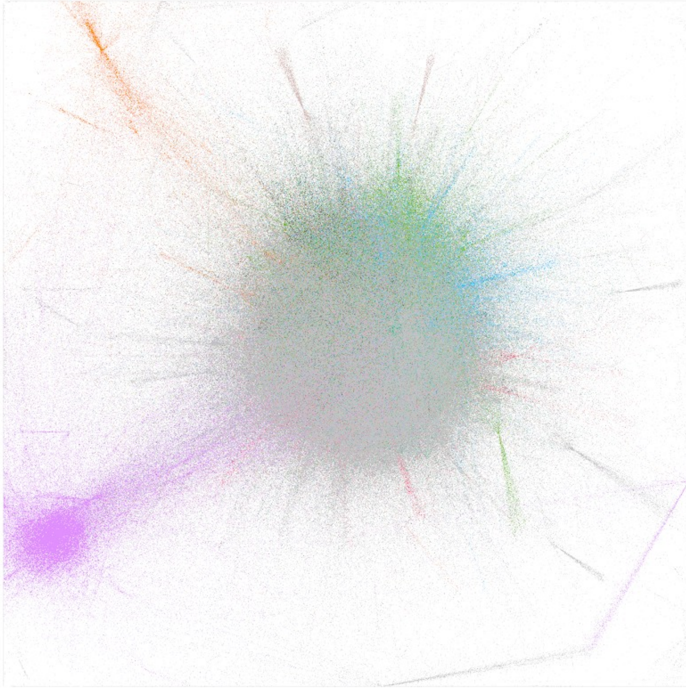
SUMMARY OF RESULTS

	Misinformation	Debunking
Cluster Connectedness	Strongly Connected	Weakly Connected
Narrative	Nazi references, emotionally charged	Focused around debunking the myth
Exposure	Peaks right before elections and plateaus after	Gains traction a few weeks after the elections
Account Status	~ 22% of the accounts suspended or deleted	~ 6 % inactive users
Account characteristics	Lower number of avg. friends and followers	Higher number of avg. friends and followers
Temporal Retweeting Patterns	Two types of behaviors – pushing messages back on timelines and quick retweeting	More natural retweeting behaviour

Overview

1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
5. Future Work

Background



- Linguistic patterns in tweet content
  - Emotionally charged tweets believed to spread more readily
  - Linguistic affordances when people knowingly lie

- Supervised learning based on user modalities from forensic work
  - Can split users post-hoc on a graph by who they retweet

Goals

- Describe *what* users talk about
- Profile *how* users talk, and the ways in which this is different between misinformation and debunking
- Ultimately detect problematic tweets in real-time for further fact-checking

Data

- Everything is from #auspol
- #deathtax
  - ALP rumored to introduce a significant inheritance tax
- #stkilda
  - “African Gangs” said to be roaming St Kilda
- More being collected

Overview

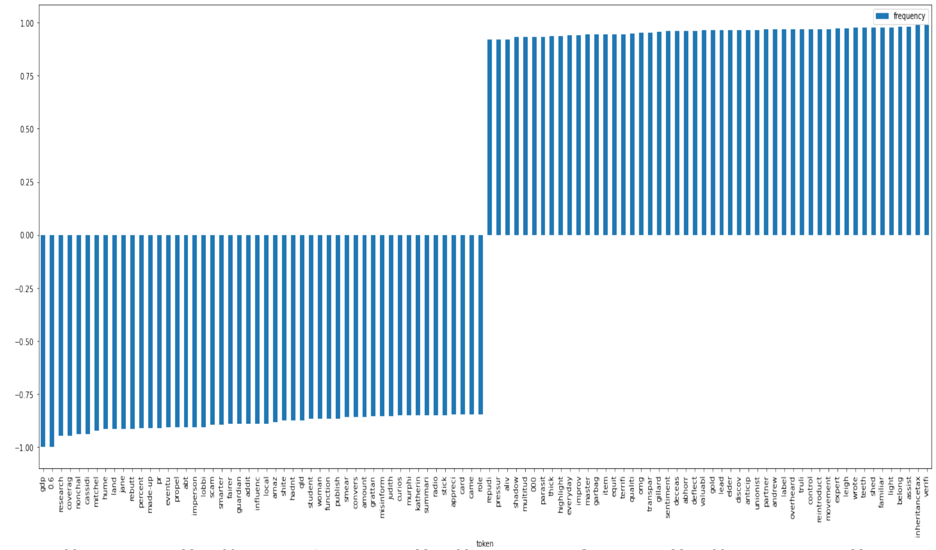
1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
    - *What* are misinformation spreaders talking about?
    - *How* are the words misinformation spreaders use different?
    - *How* is the overall writing style of misinformation spreaders different?
5. Future Work

Token Analysis

- Simple statistical analysis of token usage
- Differentiates *what* groups talk about
- Histogram of Relative Token Usage
  - Shows some tokens are used mostly in one cluster

Token Analysis

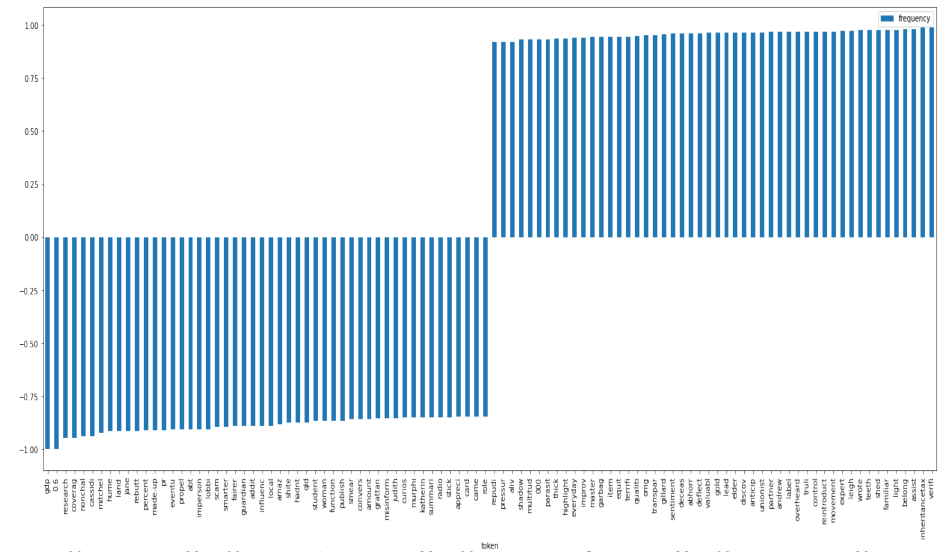
- Simple statistical analysis of token usage
- Differentiates *what* groups talk about
- Histogram of Relative Token Usage
  - Shows some tokens are used mostly in one cluster



“scam”, “made-up”, “misinform”, “smear” all very debunking correlated – they appear to call out misinformation

Token Analysis

- Simple statistical analysis of token usage
- Differentiates *what* groups talk about
- Histogram of Relative Token Usage
  - Shows some tokens are used mostly in one cluster
- Appears to be something in the content initially



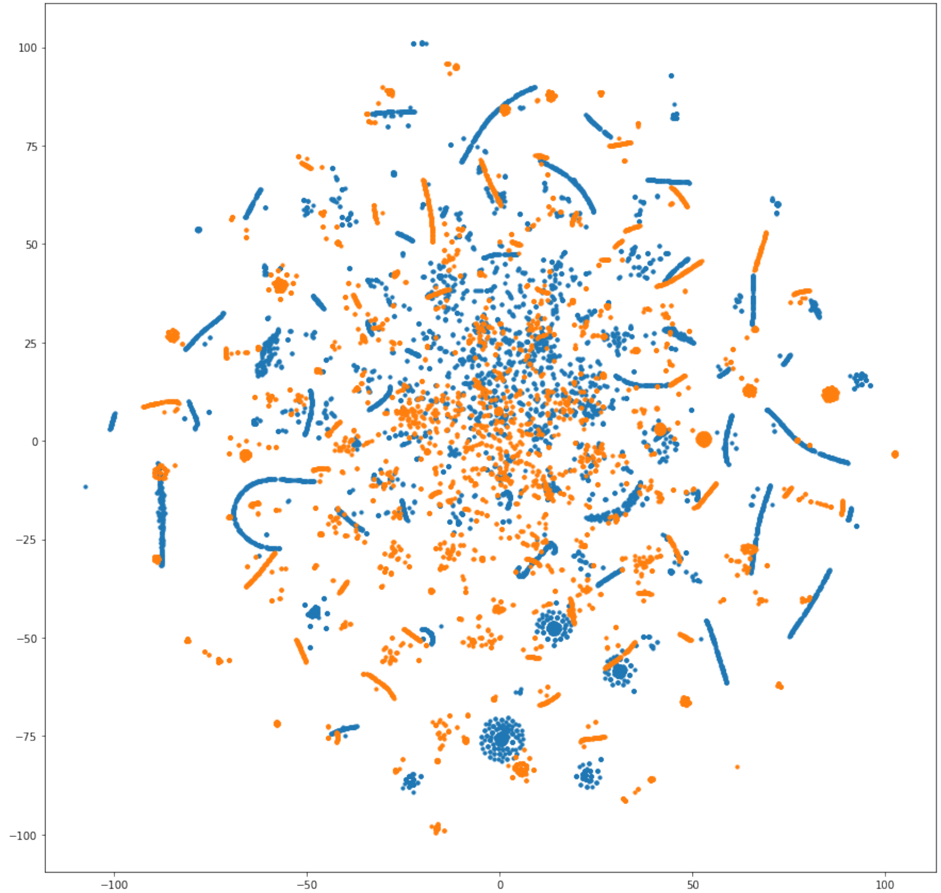
“scam”, “made-up”, “misinform”, “smear” all very debunking correlated – they appear to call out misinformation

Topic Detection

- Latent Dirichlet Analysis
  - Vectors describing probability a text is talking about topic *n* in position *n*
  - Put through a dimension reduction tool to get visual split
- Differentiates *what* groups talk about

Topic Detection

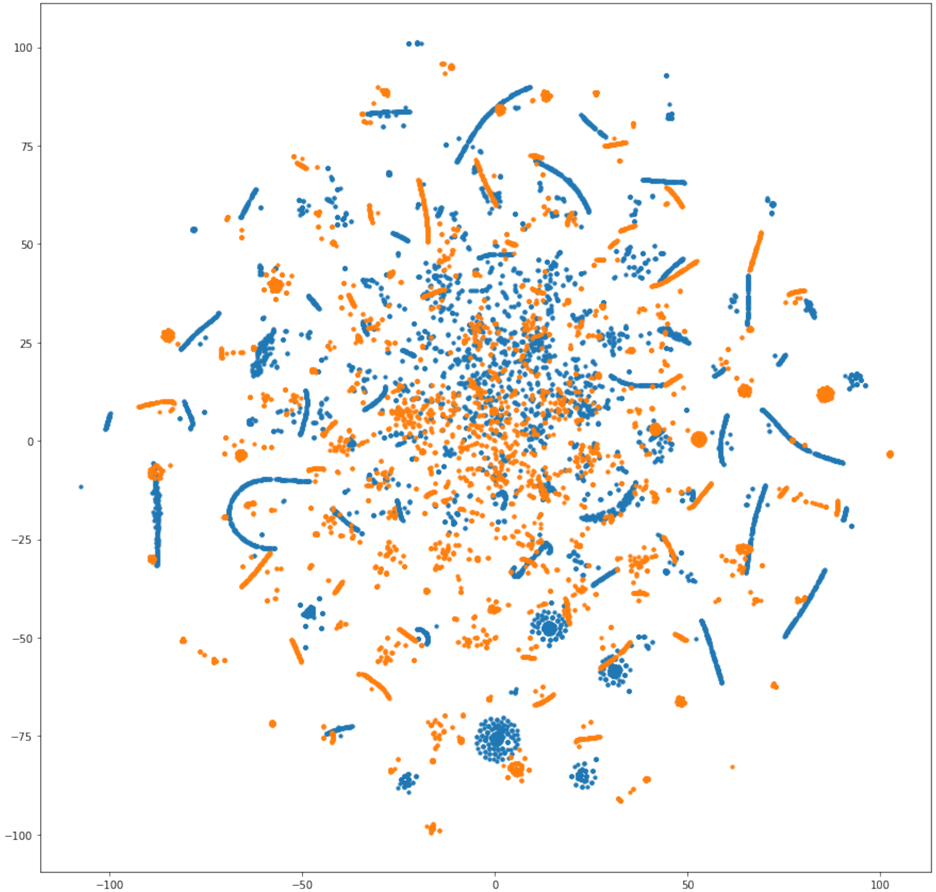
- Latent Dirichlet Analysis
  - Vectors describing probability a text is talking about topic *n* in position *n*
  - Put through a dimension reduction tool to get visual split
- Differentiates *what* groups talk about



F1 Score: 0.700

Topic Detection

- Latent Dirichlet Analysis
  - Vectors describing probability a text is talking about topic *n* in position *n*
  - Put through a dimension reduction tool to get visual split
- Differentiates *what* groups talk about
- Problem: Very context-dependent
  - Generalises very poorly



F1 Score: 0.700

Overview

1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
   - *What* are misinformation spreaders talking about?
   - *How* are the words misinformation spreaders use different?
   - *How* is the overall writing style of misinformation spreaders different?
5. Future Work

Semantic Embeddings – Word2Vec

- Each token gets an embedding
- Sum and normalize embeddings from a text for overall score
- Generalise the model by hiding parts of speech
- First attempt at differentiating *how* groups talk

Semantic Embeddings – Word2Vec

- Each token gets an embedding
- Sum and normalize embeddings from a text for overall score
- Generalise the model by hiding parts of speech
- First attempt at differentiating *how* groups talk

F1 Scores:

Trained on #deathtax:

- #deathtax texts: 0.833
- #stkilda texts: 0.250
- #deathtax users: 0.544
- Random texts: 0.130

Semantic Embeddings – Word2Vec

- Each token gets an embedding
- Sum and normalize embeddings from a text for overall score
- Generalise the model by hiding parts of speech
- First attempt at differentiating *how* groups talk
- Problem: Performs poorly when texts get large

F1 Scores:

Trained on #deathtax:

- #deathtax texts: 0.833
- #stkilda texts: 0.250
- #deathtax users: 0.544
- Random texts: 0.130

Overview

1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
   - *What* are misinformation spreaders talking about?
   - *How* are the words misinformation spreaders use different?
   - *How* is the overall writing style of misinformation spreaders different?
5. Future Work

Semantic Embeddings - BERT

- User-level embeddings
  - Concatenate all text generated by a user on a topic with a "."
- Classification Head
  - Use network modality information as labels

Semantic Embeddings - BERT

- User-level embeddings
  - Concatenate all text generated by a user on a topic with a "."
- Classification Head
  - Use network modality information as labels

F1 Scores

Trained on #deathtax

- #deathtax texts: 0.784
- #stkilda texts: 0.272
- #deathtax users: 0.830
- #stkilda users: 0.248
- Random texts: 0.130

Semantic Embeddings - BERT

- User-level embeddings
  - Concatenate all text generated by a user on a topic with a "."
- Classification Head
  - Use network modality information as labels
- Generalises better than random, but not well

F1 Scores

Trained on #deathtax

- #deathtax texts: 0.784
- #stkilda texts: 0.272
- #deathtax users: 0.830
- #stkilda users: 0.248
- Random texts: 0.130

Summary

- Misinformation spreaders talk about different topics within a scandal to normal users
- We can do better than random at identifying users based on the words they use
- Best results come from looking at sentence embeddings for the total output of a user

Overview

1. Motivation & Background
2. Data and Method
3. Forensic analysis of a campaign
4. Content-based disinformation classification
5. Future Work

FUTURE WORK

- Further author profiling
- Sentiment analysis on the narratives
- Investigate the spread of the #DEATHTAX in traditional media
- Document approach and findings – journal article in *Digital Communications and Networks*

- Fine-tune BERT's language model to our data
  - DINO method for doing this in a supervised way
- Train multiple classification heads
  - Generalisation
  - Ensemble methods for combining outputs

REFERENCES

Allcott, H. and Gentzkow, M., 2017. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), pp.211-236.

Bovet, A., Morone, F. and Makse, H., 2018. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. Scientific Reports, 8(1).

Gupta, A., Lamba, H. and Kumaraguru, P., 2013. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. 2013 APWG eCrime Researchers Summit,.

McSwiney, J., 2020. Social networks and digital organisation: far right parties at the 2019 Australian federal election. Information, Communication & Society, pp.1-18.

Starbird, K., Maddock, J., Orand, M., Achterman, P. and Mason, R., 2014. Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. iConference 2014 Proceedings,.

SUPP: EXISTING LITERATURE

Study	Content Analysis	Account Analysis	Network Propagation	Fact-Checking	Temporal Tweeting Pattern	Dataset – Australian Based
(Gupta et al., 2013)	Y	Y	N	N	Y	N
(Starbird et al., 2014)	Y	N	N	N	N	N
(Allcot and Gentzkow, 2017)	N	N	N	Y	N	N
(Bovet and Makse, 2018)	N	N	Y	N	Y	N
(McSwiney, 2020)	Y	N	Y	N	N	Y
This project	Y	Y	Y	Y	Y	Y

SUPP: USE OF METHODS

Study	Content Analysis	Account Analysis	Network Propagation	Fact-Checking	Temporal Tweeting Pattern
Definition	Explore specific themes/concepts in data	Analysis of social media accounts	How information flows through a social network	Verifying information in texts to assess its validity	Temporal analysis of tweeting and retweeting behaviour
Utilisation in project	Through analysis of the narratives – e.g. word clouds of themes of tweets	Through analysis of users' Twitter accounts – e.g. friends count, verified status, geotags	How the narrative / content of the tweets propagates through the network	By investigating users who that migrated from the misinformation cluster to debunking cluster	By mapping the time interval between tweets and their retweets for each cluster

SUPP: RESULTS

Measure	Misinformation Cluster (Average)	Debunking Cluster (Average)
Group Betweenness centrality $c_B(C) = \sum_{s,t \in V-C;\, s<t} \frac{\sigma(s,t \mid C)}{\sigma(s,t)}$	0.01	0
Group Closeness centrality $c_{close}(S) = \frac{\lvert V-S \rvert}{\sum_{v \in V-S} d_{S,v}}$ $d_{S,v} = min_{u \in S}(d_{u,v})$	~ 1.0	~ 2.0

SUPP: RESULTS

Opinion leader		Created at	Status Count	Verified
Misinformation Cluster	HelpRodger	2017-06-21	13626	False
	labor_trash	Suspended	N/A	N/A
	PutLaborLast	Inactive	N/A	N/A
	AllNewsAlliance	2011-08-31	30559	False
Debunking Cluster	AaronDodd	Inactive	N/A	N/A
	broomstick33	2012-08-10	399401	False
	LeipzigSyd	2013-11-03	103102	False
	greensinspa	2014-03-21	145568	False
	CatPurry9	2015-11-29	2678	False
	Feeney4Batman	2011-11-09	12738	True
	YOKYOKbeers	2011-02-25	54875	False
	matt_warren__	2012-06-08	4597	False

SUPP: RESULTS

Opinion leader		Wayback Frequency	Total visits
Misinformation Cluster	HelpRodger	1.5	3
	labor_trash	1	1
	PutLaborLast	1	2
	AllNewsAlliance	N/A	N/A
Debunking Cluster	AaronDodd	3.14	22
	broomstick33	3.71	26
	LeipzigSyd	2	10
	greensinspa	2.5	5
	CatPurry9	2.33	7
	Feeney4Batman	5.42	38
	YOKYOKbeers	1	2
	matt_warren__	1	2