Profiling Information Warfare on Social Media: A Forensic Analysis of the 2019 Australian Elections

Kriti Tripathi

A thesis submitted for the degree of
Bachelor of Advanced Computing (Honours)
at The Australian National University

Supervised by Marian-Andrei Rizoiu
February 2021

Except where otherwise indicated, this thesis is my own original work.

Kriti Tripathi

21 February 2021

I want to dedicate this thesis to my family. Thank you mum, dad and Yatindra for always supporting me.

Acknowledgments

In this journey, I had a lot of support from many people along the way. I want to thank my peers in the data science research group for your countless support. In particular, I would like to thank Rohit and Quyu for their contributions to this research. I would also like to thank Chinmay, Razi and Namisha for supporting me through the last stretch of my thesis writing.

And of course, I would like to thank Jennifer Hunt who gave me invaluable advice along the way. And most of all, I want to thank my supervisor Marian-Andrei Rizoiu. Without Andrei and his guidance, this would not have been possible.

Abstract

In the era of heavy social media reliance, there is a growing concern about the number of potential vulnerabilities that we are being exposed to but the most detrimental of them all are the ones that operate silently without a hint of detection. One such threat is mis-/dis-information – misinformation being the spread of in-factual information online and disinformation being the deliberate or intentional spread of inaccurate information. Current literature, while vast, does little to offer a concrete approach with which to tackle this problem. Additionally, there is a glaring gap in our current understanding of typical characteristics of a mis-/dis-information campaign.

In this thesis, we investigate a widely-debated instance of mis-/dis-information campaign during the 2019 Australian elections. We use the 2019 #auspol Twitter data and analyse it using our novel four-tiered analytical approach that we introduce in this work. This approach constitutes of looking at the data at four different levels: content-level, user-level, network-level and diffusion-level. We use the results obtained from this analysis to report on distinguishing features of our chosen instance of a mis-/dis-information campaign.

At the content level, we find that the result of using sentiment analysis on tweets yields a higher objectivity score and higher neutral sentiment scores for the cluster which perpetuates the mis-/dis-information. Using this, we derive that the users that belong to the cluster which perpetuates the mis-/dis-information tend to have deliver their message more objectively, and with little trace of positive or negative emotion, to adopt a false sense of factuality. At the user level, we find that 63% of all inactive or suspended accounts belong to users that spread mis-/dis-information. We use this result to infer that a higher proportion of suspended or inactive accounts is a distinguishing feature of mis-/dis-information cluster.

At the network level, we find that cluster of users that spread mis-/dis-information have higher centrality measure. From this, we find strongly connected networks with a high level of inter-connectivity to be a defining characteristic of the mis-/dis-information campaign. Lastly, investigation into retweeting behaviour reveals that misinformation cluster users have much faster as well as slower retweeting times compared to other users. We find that this twofold retweeting behaviour is inorganic in nature and is an indication of an organised effort at spreading disinformation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last few decades have observed an exponential rise in popularity of social media consumption ranging from trivial use in ever-day life, such as for communication, to more consequential usage such as employing social media as a platform to debate and spread false news about crucial topics such as politics [Bradshaw and Howard, 2018]. There are many advantages of social media ranging from ease of communication to accessibility. However, less widely considered by the general public, are the vulnerabilities that we are exposed to as we continue to migrate our identities online. Our increased reliance on social media leaves room for exploitation of users and inadvertently provides free real-estate for ill-intentioned actors to operate.

1.1 Motivation

The sheer vastness of social media makes it challenging to monitor and riddle all the content which is being circulated on these platforms and any potential effort is hampered by the ever-evolving techniques that are being used to evade detection by users who intend to misuse this technology. The lack of comprehensive online content-monitoring capabilities allows for ample opportunity for fictitious information to circulate unchecked online [Vraga and Bode, 2018]. This is called misinformation – the spread of information that is not entirely factual – and it can be observed in many forms online such as false rumours or fake news [Chen, 2016]. It is important to note the distinction between misinformation and disinformation – where the latter is false information that is being purposefully spread to deceive and the former is false information that is being spread regardless of the intent [Wardle et al., 2018]. Simply put, disinformation is misinformation with the intent to deceive.

The unsupervised spread of mis-/dis-information online is a particularly perturbing issue as it directly threatens our freedom to make informed decisions and can have serious impacts on us – including our health and well-being [Bode and Vraga, 2018]. Unchecked spread of mis-/dis-information can be particularly harmful to our society as it can downplay the seriousness of a situation, thus causing inaction from people where an immediate response should be taken, and can also exaggerate the criticality

of a situation which can cause widespread panic. The spread of mis-/dis-information on online social media has fast become an issue of growing importance as more and more parts of our lives and identities continue to migrate online. The severity of the issue can only be expressed by emphasizing that countries around the world, including Australia, presently categorise and recognise disinformation as a weapon of destruction and a threat to democracy.

Now, more than ever, the spread of mis-/dis-information is proving to be a real threat and not only to our society as a whole but, to us, individually. Regardless of the incongruity between their rationales, multiple different parties, ranging from governments and their cyber-defence units to social media companies such as Facebook and Twitter, reserve a significant interest in solving this problem [Prier, 2017]. Successful detection and characterisation of mis-/dis-information and its spread would have significant impacts on our online and physical environments – beginning with safer online social media platforms, the use of which wouldn't completely compromise our freedom of thought, all the while reducing our level of vulnerability online. On a more national level, it would also allow detection of potential attacks from an adversary which, if otherwise undetected, could be a matter of national security [Kalpokas, 2017; Thomas, 1996]. Furthermore, the findings from this research can be used to build models that can detect mis-/dis-information in real-time online.

1.2 Research Questions

In the last few years, there has been a growing number of research conducted on misinformation and disinformation on social media. The majority of this research, however, is in the context of other countries such as the United States which begs the question of **whether there are instances of mis-/dis-information in the Australian context** and what they might look like.

Despite the amount of literature on the subject, there is no overt agreement within the research published on the types of analytical tools that should be employed for the detection of mis-/dis-information [Figueira et al., 2018]. While there is some overlap in technique, for e.g. a number of papers use content analysis as a tool to help detect misinformation, the analytical approaches documented for different papers are all but unique to the authors. For example, research by Gupta et al. [2013] and Mihaylov et al. [2015] document using account analysis as the primary approach when looking to categorise mis-/dis-information. On the other hand, Shin et al. [2018] and McSwiney [2020] focus on content analysis as a tool to observe the spread of mis-dis/information. We also have the likes of Forelle et al. [2015] who adopt the information propagation perspective as a method for finding traces of mis-/dis-information. With the disparity in the analytical approaches, the question that we can ask then is, **what is an all-encompassing analytical method that we can use to examine these scenarios?**

There is also a noteworthy lack of any readily available theoretical set of features that can be referred to when looking at identifying any instance as mis-/dis-information. We can ask, **what is a set of characteristics of a mis-/dis-information instance that can be used to categorise it as such?** Furthermore, the current spotlight on social media leaves little research that has been conducted on the **linkage of traditional media to social media and mis-/dis-information**.

We plan to address the above questions by choosing an instance of mis-/dis-information spread and analysing it with a four-tiered approach detailed in Chapter 3 Section 3.2. Using this four-tiered approach, we plan to extract a set of features which can then be generalised and used to identify a mis-/dis-information campaign.

1.3 Thesis Outline

In this thesis, I will present my work over the course of five chapters. In Chapter 2, I will be covering the background and introducing the existing literature on the subject. Chapter 3 will detail the dataset that I use for the purposes of this research as well as the general approach that I take to obtain my results which will be presented in Chapter 4. Lastly, I will summarise my work and discuss potential extensions to this research in Chapter 5.

Chapter 2

Background and Related Work

The literature on the detection of disinformation and misinformation on social media has been steadily growing in recent years. A significant amount of research has been conducted on the topic due to its growing relevance and exponential rise of social media's popularity over the years. Despite the vast literature, there have been underlying themes to the research and techniques which we have observed that can be categorized into: account analysis that we discuss in Section 2.1, content analysis which is covered in Section 2.2 and information propagation analysis that we elaborate on in Section 2.3. We also summarise other relevant works that have taken a deviating approach from the aforementioned three in Section 2.4.

2.1 Account Analysis

Account analysis is a type of analysis performed on account properties of users – such as the number of friends or followers – and is one of the most common approaches to detect disinformation. Gupta et al. [2013] investigated misinformation on Twitter subsequent to the Boston Marathon Bombings. The authors use regression to show that the virality of content can be predicted, with an impact metric R^2 of approximately 0.7, given the impact of users who propagate it through the 'verified' Twitter feature.

Similarly, Mihaylov et al. [2015] utilize account features of users such as the number of comments posted and the number of days the user has been on the platform to find opinion manipulation trolls online. The authors show that they can detect these trolls online with 93% accuracy when using a combination of all of the account characteristics of a given user as features for their model.

Shu et al. [2018] attempt to characterize the differences between users who can accurately identify fake news and naive users who cannot. The authors use explicit (already provided in raw user metadata) and implicit features (not directly provided like age, gender and personality) of a user profile to characterize these differences. The paper finds that older people are more likely to trust fake news and more male

users engage in news consumption on social media. The results also show that users that are extroverts and friendly are more likely to trust real news.

Rizoiu et al. [2018] explore the role social bots play in manipulation public opinion and attempts to characterize the influence of these bots. The paper models the latent diffusion structure of the tweets using user account features and tweet times. The paper finds that social bots on Twitter are not only more active but also more influential and politically engaged than humans.

This work vs. prior literature. The above works all utilise different forms of account analysis such as exclusively either account features such as friends/followers count or number of comments posted by an account.However, our work differs from these in several ways. In this work, we combine a range of loosely-related features such as account activity/inactivity as well as friends/followers count and status counts. Whereas while some of these works attempt to use a combination of account characteristics for their analysis, there is still room for addition of further depth and complexity in the analysis as the combination of features used are all one-dimensional and strongly correlated (in the realm of count of followers or account creation) [Mihaylov et al., 2015]. This is the phenomenon that we attempt to capture in our version of the account analysis.

2.2 Content Analysis

Content analysis is another popular approach taken to detect misinformation and is done through analysing the actual content and sentiment of the news which is being propagated. Shin et al. [2018] explore the diffusion of misinformation on social media through focusing on three main components – one of which is content mutation. The authors monitor 17 political rumours through their life cycle that circulated on Twitter during the 2012 US Presidential elections. The authors determine that misinformation resurges numerous times, accompanied by textual changes in the content, subsequent to its initial creation whereas facts or true rumours do not.

McSwiney [2020] looks at social networks and digital organization of the far-right parties at the 2019 elections. McSwiney approaches this through analysing the content of four far-right Australian parties: Pauline Hanson's One Nation (PHON), Fraser Anning's Conservative National Party (FACNP), Yellow Vest Alliance (YVA) and United Australia Party (UAP). The paper also maps these parties' relationships to the far-right actors on Facebook as well as Twitter and focuses on how these political parties communicate, mobilize and recruit.

This work vs. prior literature. Both the works summarised above use two different forms of content analysis – sentiment and content mutation analysis in the case of former paper, and qualitative content analysis for the latter. The narrow focus of the techniques used in both these works could be improved by making use of a broader

range of content analysis tools in conjunction. In this work, we combine tools such as sentiment analysis and qualitative content analysis through generation of word clouds, for both textual and hashtag components of a tweet, to achieve the desired overlap that these works are otherwise lacking.

2.3 Information Propagation Analysis

There is a considerable amount of research that employs information propagation as a tool to detect misinformation. The scope of information propagation analysis is not just limited to observing how information travels but also how it diffuses which can include consideration of temporal qualities of fake news' journey. Starbird et al. [2014] investigate whether misinformation can be corrected through crowd sourced information flows in the context of the Boston Marathon Bombings. The paper looks, in particular, at 3 rumours that spread on Twitter and the temporal pattern of their tweets. The authors find that for some of the rumours, misinformation continued to propagate long after its correction phased out.

Forelle et al. [2015] explore the role of political bots in Venezuelan politics and manipulation of public opinion. The paper does this through analysing tweeting, following and retweeting patterns of prominent Venezuelan politicians and prominent Venezuelan bots. The paper finds that the bots themselves only create a small proportion of traffic about Venezuelan politics. The more interesting find, as discussed by the paper, is that these bots were utilised to retweet messages from Venezuelan politicians.

Bovet et al. [2018] attempt to characterise networks of information flow as well as influential spreaders of fake news on Twitter during the 2016 US presidential elections. The paper, in particular, documents finding that the influential spreaders that spread center and left leaning news have a big influence on Clinton supporters' activity. The paper also finds that the activity of Trump supporters influences the dynamics of the top fake news spreaders.

This work vs. prior literature. There are a wide set of techniques that are used in these works that possess the potential to meaningfully contribute to information propagation analysis. The advantage our research has is that it uses these tools, such as retweeting behaviour and diffusion cascades, in combination to extract a broader and deeper range of findings from the data.

2.4 Other works

Some other works that are relevant but approach the problem in a unique way are:

Allcott and Gentzkow [2017] look at fake news spread during the 2016 presidential elections in the US. The aim of this paper was to investigate and provide evidence to the theory that fake news was a significant factor in Trump's victory in the presidential elections. The paper investigates how attributes such as education, age and total media consumption are strongly related to more accurate beliefs about whether a news headline is true or false. The authors claim that the users remembered, on average, at least one or more fake news during the election period. The paper also finds that the users were exposed to more pro-Trump news compared to the pro-Clinton articles. Lastly, the authors claim that a carefully socially engineered fake news article could achieve a bigger impact than a TV ad campaign.

Del Vicario et al. [2016] look at how misinformation spreads online through Facebook through studying user consumption of two very distinctive narratives: scientific and conspiracy news. The authors find that the consumption patterns of consumers of scientific and conspiracy news are somewhat similar. However, they also find that the cascade dynamics of the two narratives differ. The paper offers 'selective exposure' to the content as the principal driver of content diffusion and generator of homogeneous clusters which are also called echo chambers. Lastly, the paper shows homogeneity and polarization to be the main determinants for prediction of cascade sizes.

Quattrociocchi et al. [2011] focus on the Italian political campaign in 2008. The goal of the paper was to introduce a socio-computational model of opinion dynamics where agents are informed by different sources of information: agents can access one of the information sources like experts and media or by exchanging information with another agent. The agents also possess the cognitive mechanisms to accept, partially accept or reject the information that they acquired. The paper finds that the agents that were more or less likely to exchange their opinion with another agent are also exposed to information from central media. The paper also claims that when misinformation is spread by central media, peer-to-peer communication can reduce the effect of informational cheating until the broadcast reaches half the population. The paper also finds that when a proportion of population is not reached by media or experts and is only privy to peer-to-peer communication, the opinions of agents do not converge on the narrative/content spread by the media. Finally, the authors claim that the impact of peer-to-peer communication could hinder informational cheating and corruption.

2.5 Summary

Overall, a significant amount of literature exists on misinformation on social media accompanied by a plethora of diverse techniques to detect it. However, some of the most common techniques include analysis on the basis of user accounts, content of the messages and the propagation of information. These techniques, by themselves or in combination with each other, are shaping the direction in which detection of

misinformation online is headed. The difference between the existing literature and our research is that instead of focusing on the data with one type of lens, be it content, account or information propagation, We collectively use each one of these overarching techniques to draw meaningful conclusions from the data.

Chapter 3

Design and Implementation

In this chapter, we will introduce the dataset that we use for our analysis in Section 3.1. We will also discuss the overall approach that is taken and the four types of analysis we perform on the data in Section 3.2. This approach will be discussed in further detail in Chapter 4.

3.1 Dataset

The dataset that we use was collected by a team of researchers at the Australian National University (ANU) during the 2019 Australian elections. Note that the collection of data occurred prior to my commencement on this research and as such, I had no contributions to the #auspol data collection. The data was sourced from Twitter using the streaming API and filtered using the query ''#auspol'. The result of this query was a dataset of tweets that contained 'auspol' in either the hashtags, or inside the text. The tweets were stored in form of a collection of JSON objects – with the dataset consisting of over 17 million tweets. Each tweet's characteristics are stored within a single JSON object – with information such as author, timestamp, hashtags and description included as fields within each entry.


```
{'entities': [{'concept': 'Inheritance_tax',
   'psr': 0.046229343079138,
   'similarity': 0.9502691341025291,
   'support': 365,
   'surface_form': 'inheritance taxes',
   'uri': 'http://dbpedia.org/resource/Inheritance_tax'}],
 'hashtags': ['auspol', 'ausecon'],
 'lang': 'en',
 'retweet_count': 0,
 'retweeted_status': 'document-tweet-1075500111948480512',
 'timestamp': '2018-12-19T22:14:53+00:00',
 'title': 'RT @PaulNAustin: The case for inheritance taxes. By @danielleiwood https://t.co/CJNS0cHNLO #auspol #ausecon',
 'topic': 'AUS01',
 'type': 'tweet',
 'user': 'user-twitter-34752409'}
```


Figure 3.1: JSON tweet example

Ranking	Sample size: Hashtag	n = 2000 Frequency
1	auspol	630
2	qt	204
3	adshame	22
4	nswpol	19
5	livesheepexports	18
6	lnp	15
7	lnpfail	12
8	morrison	12
9	nauru	12
10	idpwd2018	10

Table 3.1: 10 Most frequent hashtags in first 2000 samples of #auspol dataset

An example of a JSON tweet object is as shown in Figure 3.1. We can see that there are multiple attributes of the tweet object that are captured here ranging from hashtags to retweet count – which is a count of the number of times the particular tweet has been retweeted. For tweets that are simply retweets of another tweet, there is a field called the 'retweeted_status' which points to the tweet id of the original tweet. One of the most important fields here is the 'user' field which tells us who has posted the tweet, and this allows us to map interactions between users.

Table 3.1 provides a good snapshot of the ten most frequent hashtags in the #auspol data. Note, this is only calculated for an arbitrary size of 2000 tweets, a decision motivated by computational difficulty encountered when parsing the massive #auspol dataset of 17 million tweets. From Table 3.1, we can see that the three most frequently used hashtags in the first 2000 tweets are 'auspol', 'qt' and 'adshame'.

Alongside the #auspol data, the ANU researchers also crawled Twitter to obtain information about the users that were involved in the #auspol discussion and this user dataset is integral to our research. The user dataset was originally of approximate size 11 million – prior to filtering which is used to find relevant users that were involved in the death tax debate. This dataset contains information pertaining to these users' accounts such as the number of followers and friends, the username and screen name as shown in Figure 3.2.

3.2 Method

To find a complete case study to analyse, discussions of which are also present in the Twitter data we extracted, we first look at five instances of online spread of misinformation/disinformation in the context of the 2019 Australian elections. We apply an

	user id	name	screen name	friends count	followers count
0	user-twitter-563996238	Joey Moore	Jokeymoore79	482	122
1	user-twitter-833254069	O'Dailagh	ODailagh	360	38
2	user-twitter-2234074554	Sims	us4sims	60	10
3	user-twitter-2951810641	nathan rowlands	NathanSingo	261	34
4	user-twitter-1054198447	Robert Bishop	RobertBishop59	378	161
...	...	...	...	...	...
14616	user-twitter-979622426676989952	.....▪▫	wrath_of_roses	134	87
14617	user-twitter-458091718	Jeremy Watson	jwat5on	244	123
14618	user-twitter-2159828310	Louisa Reid	Louisa_ElJustEl	989	737
14619	user-twitter-846591984974278656	mrworldwide20226	gm_skuj	589	190
14620	user-twitter-365458553	Tom Thorp	real_TomThorp	186	77

14621 rows × 5 columns

Figure 3.2: A snapshot of user data

ethnographic approach and for each of the instance, deep-dive into the discussions surrounding them on Twitter as well as traditional media like news articles. Next, we extract the relevant tweets from the #auspol dataset and transform the dataset that we filter into a form described in Section 4.2, such as a graph or a network, which can be functionally utilised to perform analysis on.

The novel contribution of this research is how we conduct analysis on the dataset. We take a four-faceted approach of analysis which allows us to gain a deeper understanding of the data through investigating it with a multi-dimensional lens. We use four types of analysis, what we call the four-cornerstones of analysis of this research, to extract features of our chosen instance. These are:

1. Content Analysis – The analysis we perform on the content of the dataset which begins with identifying the narratives that exist in our data. Other components of this analysis include generating text and hashtag word clouds for each side of the debate and performing sentiment analysis on the tweets. This is covered in more detail in Section 4.3.1.

2. User Analysis – The analysis we perform on the users within our dataset – where we investigate further into each user's account activity and characteristics as well as identify opinion leaders of the chosen case study. Refer to Section 4.3.2

3. Network Analysis – This is a network-level examination into how the network of users is organised and connected through computing centrality measures, such as in-degree, betweenness, closeness and eigenvector, and triads for the

network. See Section 4.3.3.

4. Diffusion Analysis – For this type of analysis, we look at diffusion patterns such as traction gained by the mis-/dis-information campaign, retweeting behaviour and information cascade sizes. The results and findings of this analysis can be found in Section 4.3.4.

We use each of these four perspectives to deep-dive into the chosen instance to get results that help characterise the spread of misinformation in the scenario. We also link the coverage of this instance on traditional media to its commentary on social media and, in particular, Twitter.

3.3 Summary

In this chapter, we introduce the #auspol dataset that we use to perform analysis on in Section 3.1. We present the structure of each tweet object – where it is stored as a JSON object with associated attributes such as user who posted the tweet, the timestamp, description, hashtags, etc. We detail how this Twitter data was obtained by a team of ANU researchers through the use of the streaming API using queries such as 'auspol'. We also introduce another dataset which constitutes of account information of users whose tweets were a part of the result of the streaming API query.

In Section 3.2, we elaborate on the general approach that is taken to conduct our forensic analysis on the data where we begin by choosing an instance of online mis-information spread. We introduce the four-cornerstones of analysis approach that we take to examine the instance which are: content analysis, user analysis, network analysis and diffusion analysis. Lastly, we discuss linking the coverage of the chosen instance on traditional media to that of social media.

Chapter 4

Results

In this chapter, in 4.1 we start by identifying a list of cases of attempted opinion manipulation on social media and deep-diving into one such instance. Next, in 4.2, We will be presenting how we transform the dataset to a form that we can perform analysis on. In 4.3, we will be relying on our four cornerstones of analysis, as outlined in 3, to forensically analyse one the instances – the #deathtax. We will also present and discuss results of performing each types of analysis on the case study. Lastly, we will be linking the coverage of this instance on traditional media to social media in 4.4.

4.1 Instances of online mis-/dis-information

To investigate the spread of misinformation on social media, we begin by constructing a list of instances of opinion manipulation and misinformation campaigns during the 2019 Australian Elections. We take an ethnographic approach and find articles on traditional media that discuss the stories and misinformation that garnered a significant amount of interest from the Australian people.

Each case of misinformation campaign is analysed to observe how the misinformation was spread to Australian residents and format in which it was carried out. We construct a criteria to evaluate which instance to investigate as the case study. The first criterion we consider, when looking at these instances of opinion manipulation, is the amount of traction that each campaign gained. The second criterion we look at is the delivery of the misinformation in each case to its target audience. The last criterion is formulated around whether the discussion of the misinformation was a balanced argument.

We present the following instances of online opinion manipulation we uncover in the following subsections:

- 4.1.1 – **Misleading signs**: Misleading signs in the Chinese language, in the official colours of the Australian Electoral Commission (AEC), were placed close

(a) Picture of one of the misleading signs

(b) Twitter coverage of the misleading signs instance

(c) Traditional media coverage of misleading signs instance

Figure 4.1: Evidence of misleading signs instance

to election booths on the election day by the Liberal National Party.

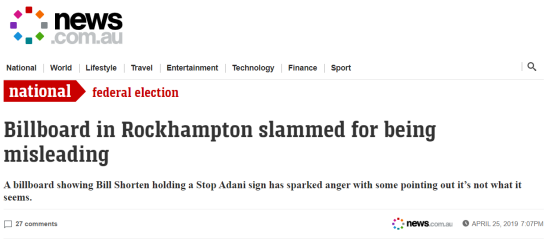
- 4.1.2 – **Stop Adani**: A strategically cropped photo of Bill Shorten holding a #StopAdani sign was advertised publicly on a billboard in Queensland.
- 4.1.3 – **Increased age pension**: Warren Mundine promises an increase in Age Pension as a part of his campaign but this would only be true due to indexation.
- 4.1.4 – **Frydenberg's coup**: GetUp's Oosting spreading rumours about Frydenberg being involved in the coup against Turnbull in GetUp's official magazine guides.
- 4.1.5 – **Death tax**: A misinformation campaign about Australian Labour Party introducing an inheritance tax.

4.1.1 Misleading Election Signs

There were reports of signs in the Chinese language being displayed in Chisholm by the Liberal National Party misleadingly instructing that the "correct way" to vote was to put a 1 next to the Liberal candidate and number the rest in order of preference (4.1(a)). Furthermore, the signs were designed to be white and purple – a similar colour scheme to that of the Australian Electoral Commission (AEC).

There were multiple tweets of pictures of these signs posted on Twitter on the election day as can be seen in 4.1(b). Multiple media articles were published on the topic – an example is an article by The Guardian in 4.1(c). We decide not to proceed with this particular instance of misinformation because while part of the discussion is accessible online, the spread of the misinformation was not through online channels – and

(a) Twitter coverage of #StopAdani instance



(b) Traditional media coverage of #StopAdani instance

Figure 4.2: Evidence of #StopAdani instance

instead this was spread through physical signs in some location – which would not allow for a thorough online analysis due to existence of some off-line components.

4.1.2 Stop Adani Billboard

It was alleged that the Morrison government displayed a billboard in Queensland which strategically shows Bill Shorten holding a '#STOP ADANI' poster. Adani is a mining company that has been looking to mine and burn coal which this has been a subject of a politically-charged debate, particularly in Queensland. It was revealed that the image of Bill Shorten holding that sign was, in fact, doctored and was used to spread misinformation.

A debate around Liberal party's effort to undermine Bill Shorten was carried out on Twitter and an example of a tweet can be seen in 4.2(a). 4.2(b) is an example of an article that was published on this subject. We decide not to proceed with this instance of misinformation for a similar reason to the 'Chinese Signs' – the principal part of the misinformation campaign was carried out off-line and so it would be hard to fully encapsulate this instance online.

(a) Twittter coverage of the Increased Age Pension instance (b) Traditional media coverage of the Increased Age Pension instance

Figure 4.3: Evidence of Increased Age Pension instance

(a) Twitter coverage of Frydenberg's Coup instance (b) Traditional media coverage of Frydenberg's Coup instance

Figure 4.4: Evidence of Frydenberg's Coup instance

4.1.3 Increased Age Pension

Another case of misleading information being spread during the 2019 Australian Elections was Warren Mundine's campaign advertisement on the side of a vehicle alleging that he will increase age pension. This was later corrected by the Prime Minister Scott Morrison who stated that the allegation would be correct, and not because of Mundine, but because of indexation.

As expected, there was a discussion about Mundine's claim on Twitter (see: 4.3(a)). This incident was also captured by an article released by Buzzfeed, refer to 4.3(b), which briefly discussed the misleading information. Similar to the previous two cases, this mis-/dis-information in this instance was spread with the use of off-line tools – disallowing full encapsulation of the instance and its context online.

4.1.4 Frydenberg's Coup

Paul Oosting, who is the boss of GetUp Australia, a self-labelled progressive internet activist group, defended his group's comments implicating Josh Frydenberg in the

(a) Twitter coverage of Death tax instance

(b) Traditional media coverage of Death tax instance

Figure 4.5: Evidence of Death tax instance

coup against Malcolm Turnbull on an interview on a Melbourne radio station. 4.4(a) is an example of a tweet that was posted on Twitter about Oosting's unsubstantiated claims. There were a few articles that also covered this story and an example of one can be seen in 4.4(b). We choose not to pick this instance for our analysis as, despite the fact that in this scenario, the misinformation was spread through media, the debate on Twitter was not a balanced one – with mostly users correcting Oosting's mistake.

4.1.5 Death tax

This misinformation campaign was widely known and circulated during the election period in 2019. The misinformation campaign stated that the Australian Labor Party (ALP) was going to introduce a sizeable tax called the 'death tax' – which is another term for tax the taxable component of any benefits or assets that a beneficiary might inherit when someone, who has them listed as a beneficiary or dependent, passes away.

We find that the entirety of the discussion about #DeathTax took place on online social media platforms such as Twitter, refer to 4.5(a) and Facebook. Furthermore, we find that the #DeathTax discussion was a balanced one – with significantly vocal opposite parties. Additionally, there was a lot of coverage by articles and news programs on this debate – for example in 4.5(b). We arrive at the conclusion that the most suited and interesting misinformation campaign is the #DeathTax. Hence, all of the forthcoming analysis is conducted through the #DeathTax case study and dataset.

4.2 Dataset: from #auspol to #deathtax & the Retweet Network

To make the data #DeathTax-centric, we filter the original #auspol dataset with the hashtags: #deathtax and #inheritancetax to obtain a subset of tweets that were involved in the death tax discussion on Twitter during, and leading up to, the election period. Furthermore, to account for cases where users might not have tagged their tweet with the appropriate hashtag, we also filter through the content of the tweets for keywords such as: 'death tax' and 'inheritance'. Subsequent to filtering, we obtain our #DeathTax dataset of size 30,234 of only relevant tweets that were published by 14,621 users.

We also build the diffusion cascade dataset that is used to analyse diffusion patterns of the #deathtax tweets. The Twitter cascade data we use is the diffusion cascade of online content such as tweets. A single cascade in the data consists of an initial tweet and all of its retweets. The data is structured according to relative times of the retweets to their initial tweet. The largest cascade from this dataset is of size 1429 and is started by the user 'twitter-user-150856088' @Bowenchris. The smallest cascade size of the dataset is 1, also the most common cascade size, and there were 1575 instances of this in the dataset. Accompanying this dataset is another which contains information about the URLs that were linked or referenced in any of the death tax tweets. We find that there were over 200,000 links referenced within the death tax tweets.

From the #deathtax tweets and their authors, we build a retweet graph which is a network representing interaction and endorsements between users as seen in 4.6. To build this network, we use the #deathtax data to establish a list of retweeters of every user in the dataset. Importing this data into Gephi – a graph drawing platform – yields a set of nodes and edges. To obtain 4.6, we use 'Yifan Hu' – one of the many inbuilt layout functionalities that Gephi offers. Yifan Hu is a type of forced-directed algorithm which attempts to find optimal placement of nodes and edges in a network by modelling forces between them which minimises the number of crossing edges. As this was not the main focus of our thesis, we will not discuss this in further detail. However, interested readers can refer to 'Efficient, high-quality force-directed graph drawing' by Hu [2005].

In this graph, we have nodes that represent the users and edges that represent the retweeting interaction. The edges are directed to signify the retweeter-retweetee relationship with the source node of an edge representing the user who is retweeting and the target node representing the user who is being retweeted. Note, in this graph, the node size is representative of a user's follower count. Looking at this graph, we find that the network has three main components. The first component is the massive cluster of users on the left. The second is the cluster of users on the right side of the figure. Lastly, the third component is the nodes or users who do not retweet and

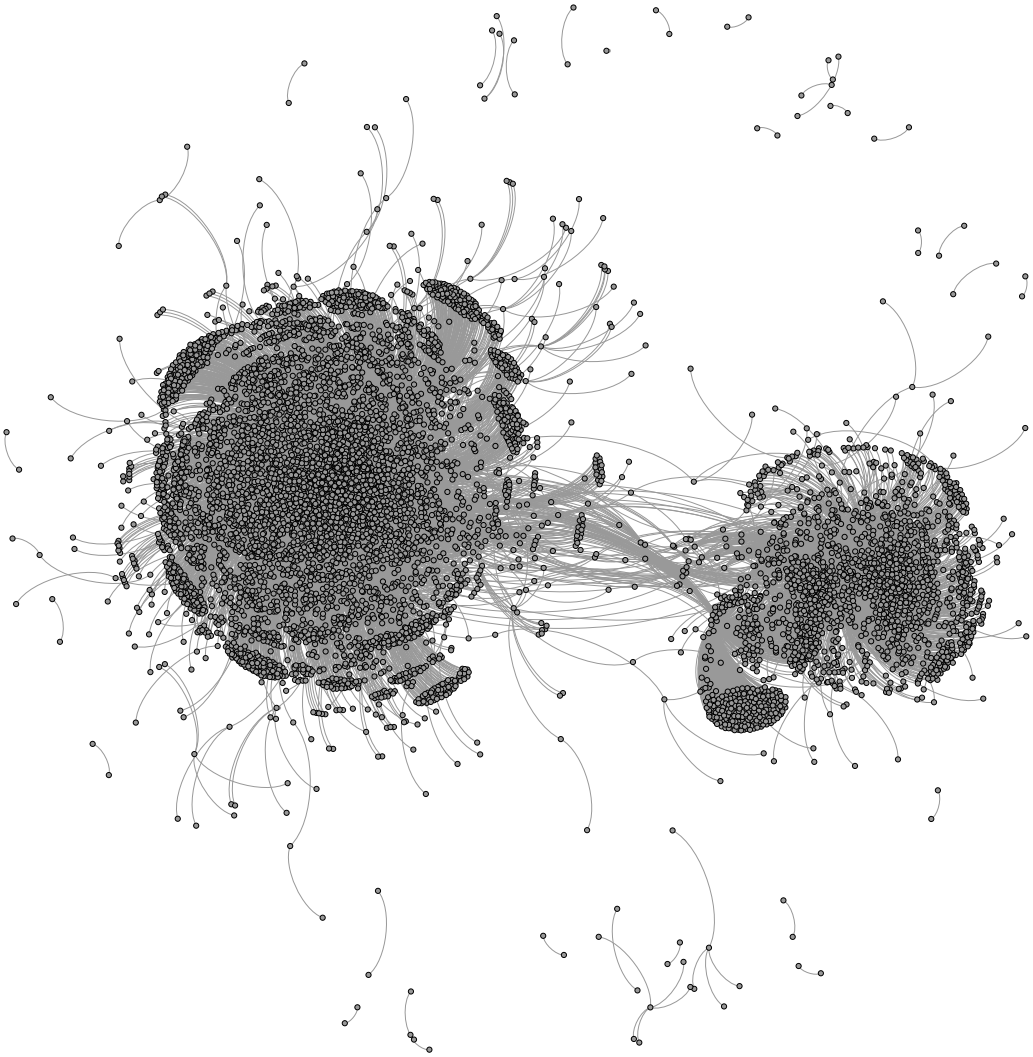


Figure 4.6: Mapping user interaction between Twitter users

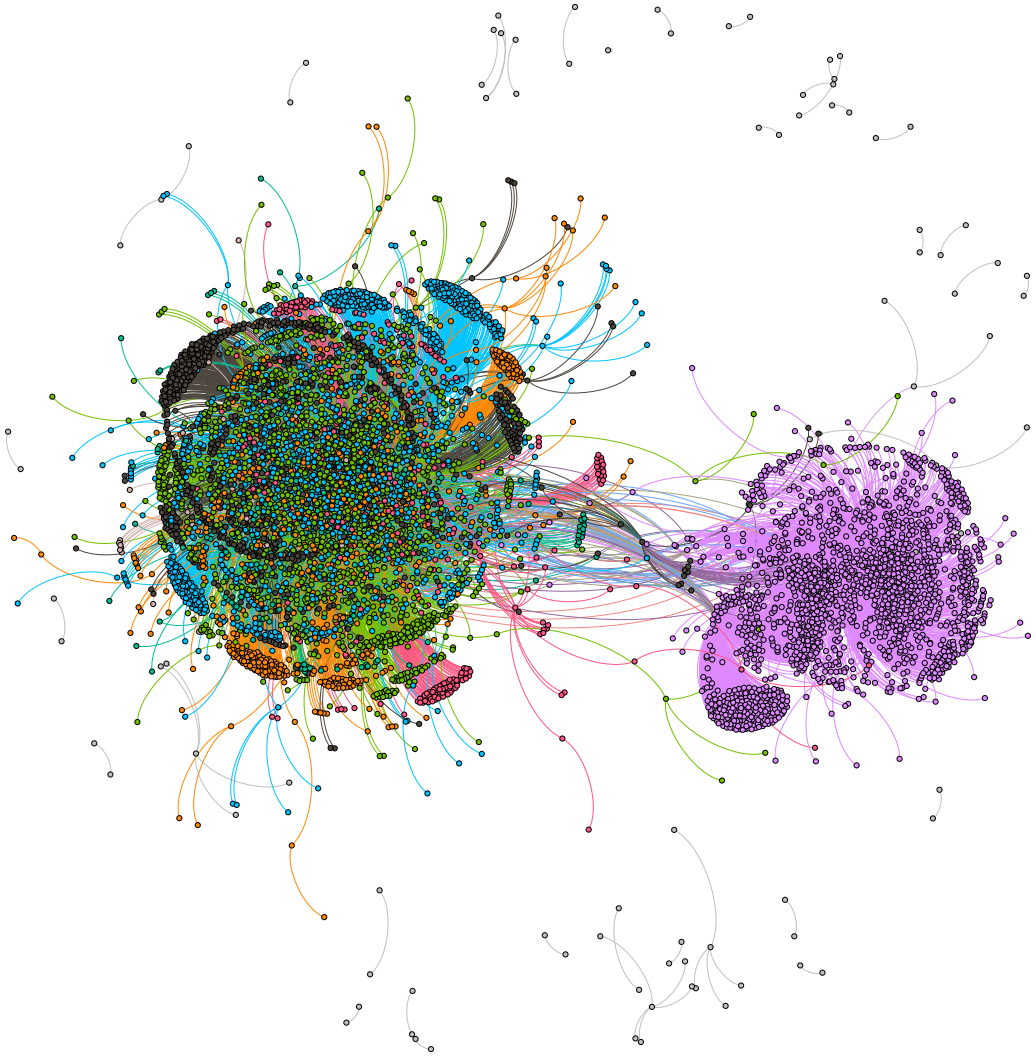


Figure 4.7: Communities of users detected in the network where each colour denotes a different community

are not retweeted by any other user and these nodes can be found in the circle on the outskirts of the two main clusters. To focus on the two main cluster and remove noise, we decide to disregard the disconnected nodes – i.e. disregard users who did not retweet or were not retweeted by anyone.

Running Gephi's clustering algorithm on the network in 4.6 yields 4.7. The clustering algorithm essentially looks for nodes in the network that are more densely connect in comparison to the rest. Each community detected by this algorithm within the network is represented by a unique colour. We can note here the distinction in our usage of the terms cluster and community. In our context, a cluster refers to one of the two main components which we can see in 4.6. A community refers to a group of users indicated by a colour in 4.7. In our instance, one or more communities can exist within a cluster.

The algorithm operates in two stages with the first stage is to detect the communities within the network. In this stage, the algorithm evaluates modularity gain of a node and each of its neighbours that would be achieved if they were grouped in a community together. In this context, the modularity is defined as the value that measures the density of links within the communities as opposed to between communities.

The community for which the gain is maximum is the community which the node would belong to. This evaluation is repeated for each node and its neighbours until no improvement is observed. The second stage of the algorithms involves building a new network where the nodes represent that communities that were detected by the algorithm in the previous stage. The algorithm calculates the link weights between the nodes by computing the weighted sum of the links between the nodes belonging to the two communities which represent the two nodes Blondel et al. [2008].

From 4.7, we can see that out of the two distinct clusters, the one on the left constitutes of many smaller communities, however, the cluster on the right is made up of one big community. We also find that the left cluster is more densely packed with nodes in their respective communities. On the other hand, right cluster has densely packed nodes within the main cluster.

4.3 Analysis

We perform four types of analysis on the interaction network, that we introduced in 4.2, which are: content, user, network and diffusion. The purpose of each of these facets is to be able to explore and characterise the network we have built from a different angles which would aid in knitting a bigger picture that would give us holistic insight into the spread of misinformation in this case study.

(a) (b)

Figure 4.8: Examples of tweets from the left cluster

(a) (b)

Figure 4.9: Examples of tweets from the right cluster

4.3.1 Content analysis

The first type of analysis we perform is content analysis where we examine the core differences in narratives of each of the two clusters that we identified in 4.2. We begin by looking through some of the tweets that have been posted by the users from each cluster.

4.3.1.1 Examples of Tweets

Left Cluster. 4.8(a) and 4.8(b) are examples of tweets from the left cluster in 4.7. We can see that both these tweets are trying to debunk the myth about the #deathtax misinformation campaign.

Right Cluster. On the other hand, 4.9(a) and 4.9(b) are examples of tweets from the right cluster. Both of these demonstrate an overt attempt at perpetuating the #deathtax campaign.

Insights. We can see from above examples that the left cluster is a community of users who are actively trying to debunk the #deathtax myth and the right cluster is a

community of users who are spreading the misinformation. We will call the former cluster the **debunking cluster** and the right cluster, the **misinformation cluster**. So, we find that that our network consists of two clusters of opposing narratives about the issue in question.

4.3.1.2 Text and Hashtag word clouds

To capture the essence of the messages that each of the debunking and misinformation cluster was spreading, we decide to investigate the word frequency use within tweets from each cluster. Note, we only look at original tweets and disregard all retweets that do not have further caption or text added to the original tweet. To visually represent this, we create word clouds of both the textual and hashtag aspects of a tweet. we will first examine the tweet content word clouds for each cluster.

Text (Tweet) Word Clouds

To obtain the word clouds for each cluster, we extract the text from all of the #deathtax tweets from all the users from each cluster. Next, we perform data pre-processing – where we remove punctuation, stop words (commonly used words such as 'and', 'the', 'a', etc.) – as well as ensure that the text case across the tweets is uniform. Then, utilising the WordCloud library in python, we generate a word cloud for each cluster. The results are as depicted in 4.10 and 4.11.

Debunking Cluster. The text word cloud for the debunking cluster reveals that the one of the most frequent phrases in the tweets from users belonging to the debunking cluster is 'death tax' or 'inheritance tax' which is expected. The more interesting finding is that some of the other frequently used phrases/words are 'lie', 'scare campaign' and fake news'.

Misinformation Cluster. Similar to debunking cluster, one of the most frequent phrases used in the tweets of users belonging to the misinformation cluster is 'death tax' and 'inheritance tax'. But again, the more interesting result is that the users within this cluster frequently mention other politically and environmentally driven topics such as 'defence funding' and 'ban coal'.

Insights. Comparing the two text word clouds, we can see that the tone of the words and phrases in the word cloud of the misinformation cluster is coarser than that of the debunking cluster – for example 'death duties'. Additionally, we find that the debunking cluster narrative and usage of words is centered around spreading the facts about the #deathtax myth. On the contrary, the misinformation cluster's word cloud reflects the brutality of the messages they are attempting to spread in their tone.

Figure 4.10: Word cloud representing tweets from the debunking cluster

Figure 4.11: Word cloud representing tweets from the misinformation cluster

Figure 4.12: Word cloud representing hashtags from the debunking cluster

From 4.11, we can also take notice that the misinformation cluster users interleave discussions or mentions of other political and controversial issues within their delivery of their #deathtax narrative. A possible explanation is that the misinformation cluster users attempt to leverage the already existing bias and controversy around these other issues when tweeting about the death tax to present a more convincing argument.

Hashtag Word Clouds

We repeat the same process for the hashtag word clouds that we followed for the text word clouds. The sole difference between the two is that we only consider the hashtags used by each cluster for each tweet. The purpose of analysing the hashtags separately is tied closely to how tweets garner traction on Twitter – it is generally the usage of hashtags that allows for virality to occur. Hence, to fully comprehend how the spread of these narratives differs, it is important to compare the usage of their associated hashtags.

Debunking Cluster. The debunking cluster's hashtag word cloud resembles closely the tone of its text word cloud – with, unsurprisingly, the most frequently used hashtags being #auspol and #ausvotes. However, some of the other hashtags used by the cluster are also #fakenews and #afternoonbriefing.

Misinformation Cluster. The misinformation cluster's hashtag word cloud also resembles closely its text word cloud. Unlike the debunking cluster, the most frequently used hashtags alongside #auspol is #deathtax by the users who belong to the

Figure 4.13: Word cloud representing hashtags from the misinformation cluster

misinformation cluster. Some other hashtags used by the users within this cluster are 'skynews', 'abc' and '9today'.

Insights. It is interesting to observe the difference in how the two clusters are tagging and spreading these messages. Misinformation cluster's excessive use of #deathtax provides a glimpse into the narrative that they are attempting to paint for wider Australia and it is a narrative which is designed to scare the lesser informed Australians into making misinformed political decisions.

Studying 4.13 yields that quite a few of the media outlets are used as hashtags by these users. Hashtags on Twitter, allow for tweets to reach a wider audience who want to follow certain topics [Bruns and Burgess, 2011]. By tagging widely-followed and popular media sources, misinformation cluster users are attempting to reach as wide an audience as possible.

Summary. To summarise, examining the tweet and hashtag word clouds of the two clusters reveals that the misinformation cluster users tend to embed mentions of other politically-charged issues when discussing their narrative. Furthermore, these users also attempt to signpost their tweets with a plethora of news source outlets in order to spread their narrative as widely as possible.

4.3.1.3 Sentiment analysis

Sentiment analysis, a tool which is often used in data science, is a type of analysis which can be used to extract the underlying sentiment of text. Also known as opin-

	Debunking Cluster	Misinformation Cluster
Negative Score	0.235	0.219
Neutral Score	0.660	0.659
Positive Score	0.104	0.122
Compound	-1.0	-1.0

Table 4.1: Distribution of sentiment scores using VADER

ion mining, this tool can employed to examine text-based data and determine if the subliminal tone of the text is negative, neutral or positive. To perform sentiment analysis on the #deathtax dataset, we prepare the data by extracting the textual content of all the tweets from each cluster. To ensure higher accuracy results and filtering of noise, we pre-process the data by removing any punctuation and stop words and create one dataset per cluster – constituting of the original tweets in each cluster in addition to their retweets.

We begin by combining the tweets into one string object for each cluster to get an general understanding of the overall sentiment of each community.

VADER. Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based model which relies on a dictionary that maps words to sentiment scores which are then used to evaluate the sentiment and polarity of some text [Hutto and Gilbert, 2014]. There are multiple python libraries that enable support for sentiment analysis using the VADER model and we explore two in detail for this research: 'NLTK' and 'TextBlob'. The rationale behind choosing these two is that they are more widely-used for National Language Processing (NLP) and provide functionality to assess emotion as well as subjectivity, in case of TextBlob, of a text.

NLTK. There are multiple tools available within the NLTK in python that can be used in this instance, however, the tool which we use from the package is the 'demo_vader_instance' tool within the VADER module. This is useful for determining the sentiment distribution of a text. The results from this tool take form of four component scores: negative, neutral, positive and compound. The scores for negative, neutral and positive components indicate the proportion of text that is categorised as being negative, neutral or positive in sentiment. These three scores should sum up to 1. The compound score is an aggregated score – sum of each word's valence score. A compound score closer to 1 indicates a higher positive sentiment in the text and -1 indicates extremely negative sentiment.

The results of using this tool on the one massive string object for each cluster are documented in 4.1. From this table, we can see that VADER sentiment distribution scores for both the clusters are very similar. VADER recognises the content from the debunking cluster to be more negative, more neutral but less positive than the

	Debunking Cluster	Misinformation Cluster
Polarity Score	-0.0330	-0.0002
Subjectivity Score	0.5061	0.5013

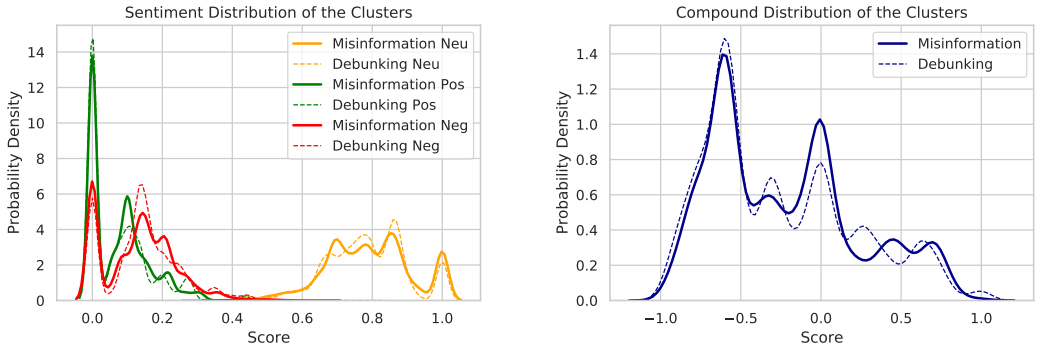
Table 4.2: Polarity and Subjectivity scores using TextBlob

misinformation cluster. The compound scores for both are the exact same which can be interpreted as VADER categorising the content from both clusters as 'extremely negative'.

TextBlob. TextBlob is a module built on the NLTK package. TextBlob analyser is trained using words from product reviews and it operates by considering the 'intensity' of each word in the text and individually assigning them scores. These scores are later aggregated to determine an overall score for the polarity and subjectivity of the text. TextBlob is particularly useful when wanting to determine the general polarity and subjectivity of some text. Polarity is expressed as a fractional number between -1 and 1 – where a polarity of -1 represents only negative sentiment, 0 represents neutral sentiment and 1 represents entirely positive sentiment. Subjectivity of a text is expressed as a fractional number between 0 and 1 – where a subjectivity score of 0 represents complete objectivity and a score of 1 indicates that the text is not factual, and instead, more an opinion.

Similar to 'demo_vader_instance', the data fed into the TextBlob analyser is the one string object that concatenates every tweet from each cluster. The results of this tool can be found in 4.2. We can see that the TextBlob analyser assigns a slightly more negative score to the tweets in the debunking cluster than the misinformation cluster, however, it is worth noting that the two scores are very close in proximity. TextBlob determines the debunking cluster content to be slightly more subjective than the tweets from the misinformation cluster. The higher objectivity score of the misinformation cluster calls attention to the fact that the users belonging to this cluster attempt to present their narrative more objectively – all the whilst masking the very subjective nature of the in-factual information they are perpetuating. Conversely, the debunking cluster users are more organic in their expression of their narrative and hence, are categorised as being more subjective. From these findings, we can gather that the misinformation cluster users are more organised from their careful approach to making their narrative appear more objective. Whereas, the debunking cluster users appear to be more subjective due to the more reactionary, as opposed to planned, nature of their narrative.

While demo_vader_instance and TextBlob give similar overarching results for polarity and sentiment, it is worthwhile to recall that VADER is more suited to be applied in a social media context compared to TextBlob, given the word source of the latter is



(a) Distribution of sentiment scores

(b) Distribution of compound scores

Figure 4.14: Results of performing sentiment analysis using SentimentIntensityAnalyzer on tweets of each cluster

from product reviews. For this reason, we choose to conduct the rest of the sentiment analysis with the NLTK package.

For the next step in performing sentiment analysis on the tweets from the clusters, we use the VADER 'SentimentIntensityAnalyzer' tool to assign the the four component scores to each tweet within a cluster and plot the results for in 4.14(a) and 4.14(b). Both the aforementioned graphs are distribution plots of sentiment scores of all tweets for each cluster. Here, the probability density of the scores are estimated using Kernel Density Estimation (KDE) in 'seaborn' – a python graphing package. The x axis of the graphs represents the score values for the sentiment components returned by SentimentIntensityAnalyzer.

Note, for 4.14(a), the x axis range of 0 and 1 corresponds to the minimum and maximum proportion of text, respectively, that can be classified as being negative, neutral or positive. For 4.14(b), the density curves overshooting the limit of -1 and 1 can be confusing at first glance, however this can be explained due to the density curves being continuous which 'seaborn' automatically extrapolates.

Insights. Looking at 4.14(a), we can see that for there are more tweets in the debunking cluster that contain no positive sentiment than in the misinformation cluster as the green peak for the debunking cluster at a score of 0.0 is higher. Furthermore, the same holds true for small-moderate positive scores of $0.1 - 0.25$ and so we can gather that there are more tweets that are $10 - 25\%$ positive in the misinformation cluster than the debunking cluster. This effect inverses for positive score values around 0.3 indicating that there are more tweets in the debunking cluster that are 30% positive.

Similarly, we find that there are slightly more tweets in the misinformation cluster that have no 'negative' components in comparison to the debunking cluster. We can also observe that there are more tweets with higher proportion of negative sentiment in the debunking cluster. We can see that neither of the clusters have tweets that are

more than $40 - 50\%$ positive or negative. However, both clusters contain tweets that are completely or highly neutral.

From 4.14(b), we can see that the debunking cluster curve is higher than the misinformation cluster's for extreme values. On the other hand, the misinformation cluster curve is noticeably higher for a neutral or moderately positive/negative scores. This means that the misinformation cluster users tend to use neutral language and shy away from extreme expression of either sentiment (positive or negative). Conversely, the debunking cluster users are more passionate in their expression of their opinions – where they use language and tone that is heavy with emotion.

Summary. To summarise, we find that 'VADER' is a tool better suited for performing sentiment analysis on the chosen mis-/dis-information instance as the word source on which the algorithm is trained originates from social media data. Using this tool, we show that the misinformation cluster users are more careful in their expression of emotion in their tweets. The findings from using 'TextBlob' show that these users also attempt to appear more objective in their tone in order to give their narrative the semblance of factuality. This organised nature of the misinformation cluster users is also corroborated by the insights obtained from text and hashtag word clouds of these users where their frivolous use of hashtags to direct more attention to their narrative.

4.3.2 User Analysis

The second type of analysis that we perform is user-level analysis whereby we identify the users within each cluster that shape and drive the #deathtax discussion. We also analyse the intrinsic differences in the account characteristics of these opinion leaders of each cluster as well as all the users that belong to each cluster.

4.3.2.1 Most Retweeted Users

To get a quick snapshot of the type and dynamic of users within the #deathtax network, We identify the five most retweeted users of all users. 4.3 exemplifies that 80% of the most retweeted users from either cluster belong to the debunking cluster whereas only 20% belong to the misinformation cluster. Additionally, the most retweeted user of our network '@Bowenchris' belongs to the debunking cluster and is retweeted 1428 times.

User name	Retweeted Frequency	Cluster Membership
@Bowenchris	1428	D
@labor_trash	1324	M
@vanbadham	1294	D
@slpng_giants_oz	1155	D
@murpharoo	999	D

Table 4.3: Most retweeted users within #deathtax network

4.3.2.2 Opinion Leaders

One of the pivotal aspects of the analysis we perform is identifying and comparing users from both debunking and misinformation clusters that drive the conversation about the death tax. These opinion leaders are the users who were retweeted the most by other users within their cluster. The rationale behind identifying profiles of these users is that it allows us to evaluate the systematic differences between the users who are shape and drive these narratives.

Insights. From 4.4, it is evident that for the misinformation cluster, 60% of the opinion leader have an inactive or a suspended account and only 20% of these opinion leaders have verified accounts. In contrast, all of the debunking cluster's opinion leaders have active accounts and 60% of these accounts are verified accounts.

For the misinformation cluster, we find some cases where the opinion leaders have a either a noticeably small number of friends or a small number of followers. Furthermore, there is a significant disparity in the friends and followers count of the misinformation opinion leaders. On the other hand, the debunking cluster opinion leaders all have a large following and a significant friends count. It is worth noticing that the friend and follower count of the debunking cluster's opinion leaders are all similar in magnitude which is not the case for the misinformation opinion leaders.

We can also see from 4.4 that, on average, the status counts for the opinion leaders of the debunking cluster are significantly larger than those for the misinformation cluster. Furthermore, most of the debunking cluster opinion leaders have established accounts which have been active for decades. On the contrary, the misinformation cluster opinion leaders have relatively new and less-established accounts. The semi-recent account creation dates can be explained by looking at recent research which shows that users that intend to spread mis-/dis-information, do so with accounts that have a recent creation date [Jones, 2019].

Summary. Exploring the account characteristics of the opinion leaders of each cluster yields some defining characteristics that can be used to identify a user as likely

Table 4.4: **Opinion leaders of the debunking and misinformation clusters.**
Due to some accounts being inactive/suspended, we are unable to retrieve the account information.

Opinion Leader		Followers Count	Friends Count	Created-at	Status Count	Verified	Account Status
Debunking Cluster	Bowenchris	86623	2622	June 2010	4339	Yes	Active
	vanbadham	62732	6747	April 2009	246963	Yes	Active
	slpng_giants_oz	16483	743	August 2017	51244	No	Active
	murpharoo	139255	1643	August 2009	49339	Yes	Active
	denisesshrivell	13021	5715	September 2008	230244	No	Active
Misinformation Cluster	labor_trash	3596	2215	-	-	No	Suspended
	HelpRodger	838	2141	June 2017	15847	No	Active
	EddyJokovich	7866	2448	April 2009	39105	No	Active
	AlanJones	23106	18	February 2016	-	Yes	Inactive
	collias_bill	111	241	July 2018	5439	No	Inactive

	Debunking Cluster	Misinformation Cluster
Active Accounts	4585 (92%)	2233 (76.8%)
Inactive Accounts	396 (8%)	674 (23.2%)
Total	4981	2907

Table 4.5: Number of active and inactive accounts for each cluster

belonging to the misinformation cluster. We find that the conversation drivers of a narrative that have a combination of inactive/suspended and unverified accounts are significantly more likely to be members of a community that is spreading mis-/disinformation. Furthermore, if these users have newly-established accounts and/or noticeably low account activity (such as number of status counts), the likelihood of this user being a suspicious actor increases significantly. At first glance, having knowledge of whether a set of users are active might seem inconsequential as inactive/suspended account status implies that appropriate action has already been taken towards that user. However, this can still prove to be useful, when examining a cluster of users where a large proportion are inactive/suspended, in categorising the still-active users belonging to the same cluster as misinformation spreaders.

4.3.2.3 Account Activity Analysis

We begin the investigation into accounts of each cluster's users by examining how many users currently have an active account. We use 'Tweepy' to retrieve active/inactive status of every user. Note, the original module 'Twint' that we use to retrieve account information of the users is made unavailable during the course of this research due to changing API and as a result, we utilise the Tweepy module instead.

Insights. From 4.5, we can see that approximately 24% of the users that belong to the misinformation cluster now have inactive or suspended accounts. This proportion is significantly less for the users within the debunking cluster at 8%. The fact that almost a quarter of the misinformation cluster users now have inactive or suspended accounts indicate that they are more frequently breaking the community rules or inappropriately conducting themselves on Twitter than the debunking cluster users. Furthermore, we can see that the total number of inactive/suspended accounts in the network is 1070. This means that approximately 63% of all of the inactive/suspended accounts in the entire network belong to the misinformation cluster.

Summary. Investigating the number of active accounts of users within each cluster reveals that approximately a quarter of the misinformation cluster users have inactive/suspended accounts. This number is significantly lower for the debunking cluster users at almost 8%. Hence, a user that has an inactive/suspended account

	Debunking Cluster	Misinformation Cluster
Verified Accounts	139 (2.8%)	26 (0.8%)
Non-Verified Accounts	4446 (89.2%)	2207 (76%)
Inactive Accounts	396 (8%)	674 (23.2%)
Total	4981	2907

Table 4.6: Number of users with verified accounts in each cluster

is much more likely to have a membership of the misinformation cluster in this instance. This finding is a generalisation of the insights gained in the previous subsection when looking at misinformation cluster' opinion leader's inactivity.

4.3.2.4 Verified Account Analysis

To capture and evaluate the quality of users that are present within each cluster, we consider the number of accounts that are verified. We use 'Tweepy' to retrieve verified status of every user in the #deathtax debate and sort them on the basis of their cluster membership. 4.6 illustrates the proportion of users that have verified accounts in the debunking and misinformation cluster, respectively.

Insights. From these results, we can see that just under 2.8% of the debunking cluster users have verified accounts. This number is noticeably lower for the misinformation cluster at less than 1%. Hence, we find that there are more verified – and by extension, reliable – users within the debunking cluster than there are in the misinformation cluster [Vaidya et al., 2019].

The table also depicts the total number of verified accounts in the network to be 165. Given that 139 of these verified accounts belong to the debunking cluster users, we can deduce that 85% of all verified accounts in the network are accounts with the debunking cluster membership.

Summary. To summarise, through comparing the verified status of the users from each cluster, we find that a user who is verified is more likely to not be spreading mis-/dis-information. Conversely, we find that a user who is actively spreading mis-/dis-information is very likely to have an account that is unverified – which is validated by our findings from observing the low proportion of misinformation opinion leaders with verified accounts.

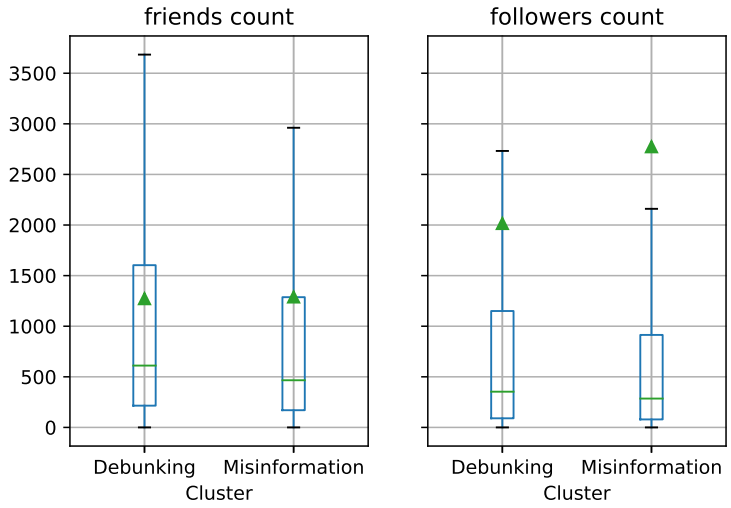


Figure 4.15: Systematic differences in friends and followers count of users

4.3.2.5 Account Feature Analysis

Account feature analysis is an approach where we look at account characteristics of users such as followers count, friends count and verified status. We begin with sorting the users according to their cluster membership – debunking or misinformation. Using the #deathtax user dataset, we obtain account details of each user. Finally, graphing these for each cluster yields 4.15. Note that we opted not to display outliers in the boxplot in 4.15 to avoid noise when presenting our results.

Insights. From 4.15, we can see that the average friends count, indicated by the green, for the misinformation cluster is slightly higher than that for the debunking cluster. This also holds true for the average followers count — where the misinformation cluster has a much higher average than the debunking cluster. Additionally, we can see that, with respect to followers count, the average follower count for the misinformation cluster sits outside the maximum which is indicative of the presence of some actors in the misinformation cluster possessing an unusually large following compared to their fellow users.

Comparing the boxplots, for both friends and followers counts, between the two cluster also yields that there is a greater variability in both the features for the debunking cluster in comparison to the misinformation cluster. This can be interpreted as the users in the debunking cluster having a more natural distribution of friends and followers count.

Summary. A comparison of account features such as friends and followers count between the two clusters shows that the misinformation cluster users on average have a slightly bigger friends count and a significantly bigger followers count on

average than the users within the debunking cluster. Furthermore, we find that there is less variability in the distribution of these features for the misinformation cluster users and, in fact, the features for these users are too in-agreement to be a natural distribution. Furthermore, this hint of abnormality is in line with the evidence of misinformation cluster's inorganic behaviour that we gathered from content analysis in 4.3.1.

4.3.3 Network Analysis

The third type of analysis we perform is network-level analysis where we explore characteristics of the network itself. Social network analysis is helpful when characterising the way a community is organised and how the members of this community are connecting. As a part of our network analysis, we look at various centrality measures of our #deathtax interaction network. Centrality measures are tools that can be used to understand graphs, networks and the role of each node better. There are a plethora of centrality measures but the ones we utilise for this case study are: degree, betweenness, closeness and eigenvector centrality. We also investigate how users are connected within each cluster using the theory of triads which considers every permutation of a triplet of users to look at the analyse of links between them. To aid in the computation of the centrality measures and triads, we use the 'networkx' package. This package is a tool which is designed to calculate various properties of a network or a graph. Note, in the following sections, the graphs are all log-plots where the x axis is the log (base 10) of the particular centrality measure being presented. The graphs are also all probability density plots estimated using the KDE functionality of 'seaborn' package.

4.3.3.1 In-Degree Centrality

Degree centrality is defined as the number of links incident upon a node (i.e., the number of edges that a node has). In-degree is a count of the number of directed edges that are incident on the node. To calculate the in-degree centrality for each cluster, we use the 'in_degree_centrality' functionality of the networkx package which calculates the in-degree centrality of a given graph's nodes. We take the results of passing our debunking and misinformation clusters through this function and plot them which yields 4.16. It is important to note here that multiple instances of a user retweeting another user is still represented by a single directed edge between the users.

Insights. In 4.16, we can see the log-curves for both of the clusters that represent the in-degree centrality distribution of the nodes in each cluster. It is evident from this graph that the misinformation curve is shifted towards the right relative to the debunking curve. This can be interpreted as the nodes in the misinformation cluster

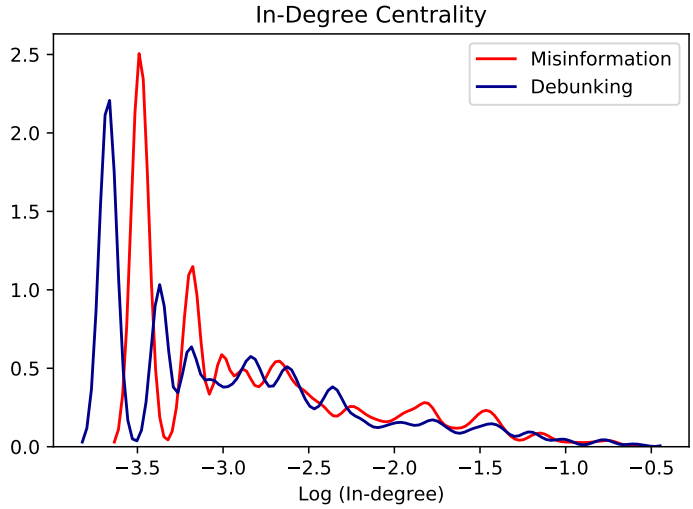


Figure 4.16: In-degree centrality measure for each cluster

having a higher in-degree centrality than the nodes in the debunking cluster. This, in our context, signifies that the misinformation users are retweeted by a significantly greater proportion of their fellow members and this implies that there is a greater inter-connectivity between users within the misinformation cluster. In contrast, debunking cluster users exhibit less of this type of behaviour – where they retweet only some of their fellow cluster members and so there is far less inter-connectivity between the users in the debunking cluster.

Summary. The main takeaway from comparing the in-degree centrality of the two clusters is that a user in the misinformation cluster, on average, retweets a wider group of their fellow users which achieves the effect of boosting connectivity of the cluster. That is, we find that the misinformation cluster users exclusively retweet each other.

4.3.3.2 Betweenness Centrality

Betweenness centrality quantifies the sum of fractions of the times a node acts as a bridge along the shortest path between two other nodes. Mathematically, this can be expressed as:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{4.1}$$

where $\sigma(s,t)$ is the total number of shortest paths between s and t and $\sigma(s,t|v)$ is the number of shortest paths that pass through our node in question which in this

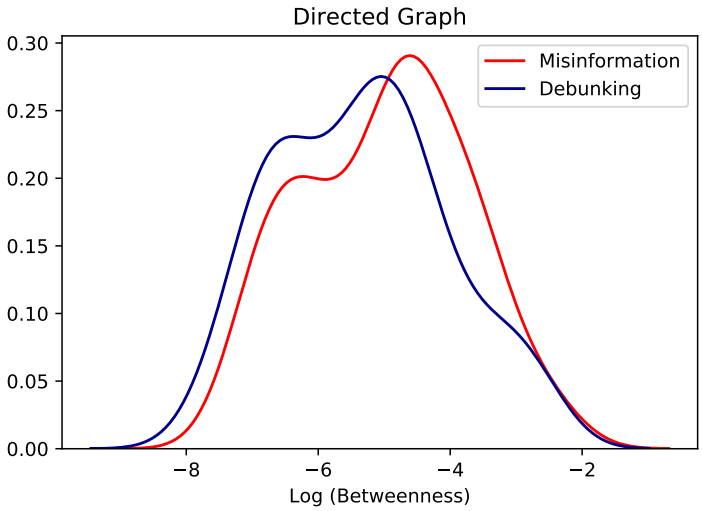


Figure 4.17: Betweenness centrality measure for each cluster

equation is represented by v.

We use the 'betweenness_centrality' tool within the networkx package to calculate the betweenness centrality for each node in both the debunking and misinformation clusters and graph the results as can be seen in 4.17. To obtain this graph, we simply use the tool on each of our clusters.

Insights. From the graph in 4.17, we can see that there are two peaks for both the debunking and misinformation curves. The first peak for the debunking cluster is higher than the misinformation cluster's. However, the misinformation curve is shifted towards the right relative to the debunking curve. Furthermore, the two peaks of the misinformation cluster's curve are more distinct in comparison to the debunking cluster's peak.

We can draw a few conclusions from this, the first being that the community of users within the misinformation cluster is made up of a two tier system. This two tier systems is reflected through the two peaks of the curve where the the higher peak is representative of the central users – the opinion leaders - and the lower peak of the two is representative of the rest of the misinformation community that is made up of the followers. We also have that the majority of the nodes in the misinformation cluster lie on more shortest paths between two other nodes in their cluster than nodes in the debunking cluster do. This means that more information is able to be passed through a node in the misinformation cluster than a node in the debunking cluster.

The significance of this finding becomes more apparent when we look at what this means for user involvement in a retweeting chain. We can take a retweeting chain to be a chain of retweeting events that starts with one user and ends with another user. A higher betweenness centrality value for the misinformation cluster then suggests

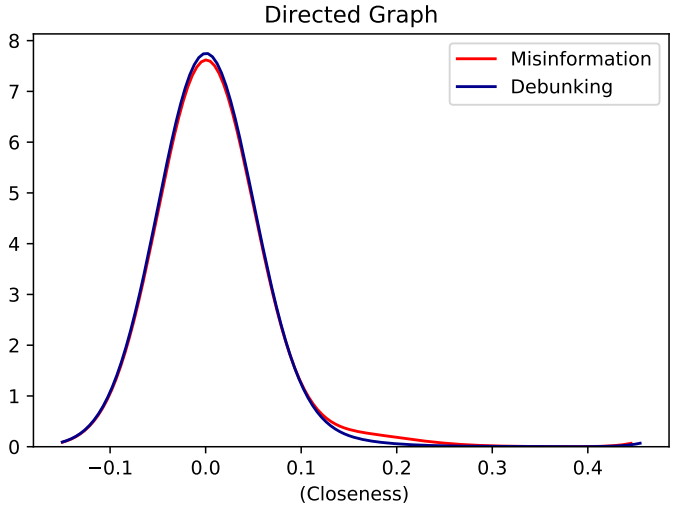


Figure 4.18: Closeness centrality measure for each cluster

that a user within this cluster is more likely to be involved in a greater number of the smallest retweeting chains between any two of its fellow users.

Furthermore, using 4.1 we can infer that the betweenness centrality of a cluster is proportional to the number of pair of nodes in a network. This means that a network with more nodes (and hence, pairs of nodes), is likely to have a higher betweenness centrality than smaller network. From 4.5, we can see that the size of the misinformation cluster is approximately 60% of the size of the debunking cluster. Given that the size of the misinformation cluster is almost half of that of the debunking cluster, we expect the betweenness centrality to be significantly lower for the misinformation cluster. However, the unexpectedly larger observed values for betweenness centrality of the misinformation cluster in 4.17 is telling of organised behaviour.

Summary. We compute and compare the betweenness centrality of both clusters as a part of our network-level analysis. This leads to the finding that, in general, the misinformation cluster nodes have higher betweenness centrality values despite the cluster being smaller in size. We suspect this is a sign of inorganic organisation of the misinformation users in their cluster which seems to be in agreement with the similar conclusions that we drew based on the insights gathered from content and user analysis.

4.3.3.3 Closeness Centrality

Closeness centrality indicates how close a node is to all other nodes within the same network. Thus, the more central a node is, the closer it is to all other nodes. Closeness centrality is mathematically defined as:

$$c(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{4.2}$$

where we have n number of nodes in a network or graph and $d(v,u)$ is the shortest distance between the nodes u and v. Note, that this version of closeness centrality is normalised and does not scale with the size of the cluster. However, there is a version of closeness centrality which is not normalised and it is mathematically defined as:

$$c(u) = \frac{1}{\sum_{v=1}^{n-1} d(v,u)} \tag{4.3}$$

Between the two variants, we choose to operate with the normalised variant of the closeness centrality on our network to ensure standardisation.

We use the 'closeness_centrality' functionality of the networkx package to obtain this measure for all of the nodes in both the clusters and plot the results as seen in 4.18.

Insights. From this graph, we can see that both the cluster curves are almost identical. The only points of difference between the two curves is towards the right tail where the misinformation cluster's curve is higher than the debunking cluster's. This means that, while for the most part, the nodes in both the clusters are similarly close to their fellow cluster members, there is a bigger group of nodes in in the misinformation cluster for log closeness values between $0.1 - 0.3$ than for the debunking cluster.

The implication here is that there are slightly more users in the misinformation cluster that are more central – i.e. a user in misinformation cluster is closer to all other users in that cluster where proximity is defined by the number of edges connecting users. More generally, this means that the users that spread mis-/dis-information are more likely to retweet or be retweeted by a connection-of-a-connection.

Summary. The findings from evaluating the closeness centrality of our clusters, in essence, support the findings we obtain from looking at the other centrality measures. Once again we find evidence that the misinformation cluster users are more connected to each other and are more central in their cluster than the users within the debunking cluster.

4.3.3.4 Eigenvector Centrality

Eigenvector centrality is a measure of level of influence of a node within a network and is computed through finding the centrality of a node based on its neighbours' centrality. Mathematically, our network can be expressed as an adjacency matrix A. An adjacency matrix is a matrix representation of a network where the (i,j) element of the matrix signifies the number of direct links that exist between node i and node

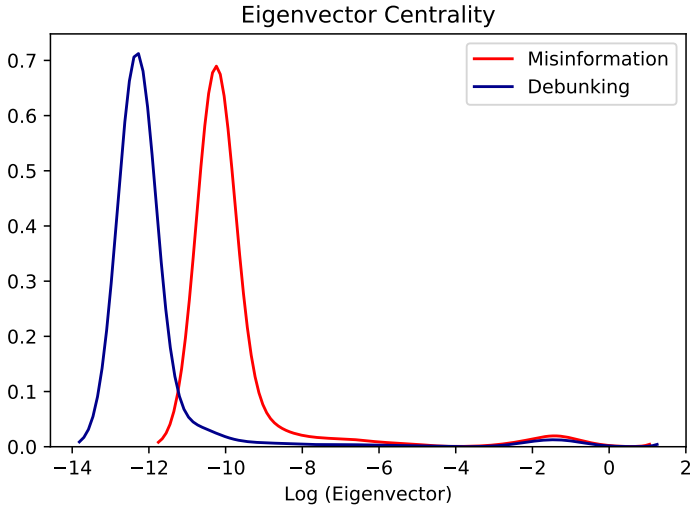


Figure 4.19: Eigenvector centrality measure for each cluster

j. We can then define x as the eigenvector of the adjacency matrix A if:

$$Ax = \lambda x \tag{4.4}$$

where λ is the eigenvalue of A. The way eigenvector centrality of a node n is calculated by 'networkx' is through inspecting the n^{th} element of the eigenvector x.

We use 'eigenvector_centrality' function from the networkx package to find this measure for every node in both the debunking and misinformation clusters and graph the results which can be seen in 4.19.

Insights. We can see from this graph that the misinformation cluster's density graph is shifted towards the right compared to the debunking cluster. This means that the nodes in the misinformation cluster have much higher values for eigenvector centrality than the nodes belonging to the debunking cluster. From this result, we can infer that the nodes within the misinformation cluster tend to have a significantly higher level of influence in their network when compared to the debunking cluster.

Since for eigenvector centrality, connections with nodes of higher importance are prioritised over connections with nodes of lower influence, a cluster with a large number of highly influential nodes implies more cohesiveness between the users [Belau, 2014]. Hence, a universally high level of influence of nodes within the misinformation cluster suggests that there is an abnormally similar level of influence across the users – because to obtain a high eigenvector centrality, more connections to other nodes with high influence are required. However this is hard to achieve in organic online setting, such as retweeting on Twitter, as it requires level of collaboration which is challenging to achieve without premeditating an effort.

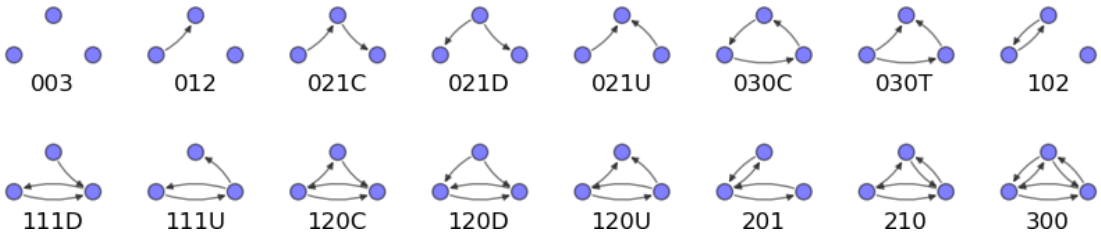


Figure 4.20: 16 types of triads according to social network theory

Summary. The eigenvector centrality characterises the level of influence a node has within a network. We find that the misinformation cluster has a remarkably high eigenvector centrality for majority of its nodes when compared to the debunking cluster. We can derive that this unusually large centrality value likely indicates a premeditated effort by the misinformation users to spread mis-/dis-information.

4.3.3.5 Triads

Theory of triads is a tool which is often used in social network analysis to get an in-depth analysis of how nodes, or users in our case, are connected within a network. A triad is a set of three nodes which represents the links, or lack thereof, between the said nodes. In a directed network, there are sixteen types of triads as seen in 4.20 [Staiano et al., 2012]. This tool is computed through obtaining every single permutation of a set of three nodes in a network to form a triad and then sorting each triad into one of the 16 types to analyse the type and frequency of interactions that are present.

We use the 'triadic_census' tool within the 'networkx' package to acquire a snapshot of the count of each of the 16 types of triads that exist in our clusters. We next plot the results in form of a bar graph where the x axis represents the type of triad – one of the 16 from 4.20 – and the y axis represents the fraction of total number of triads that makes up the particular type.

Insights. We can center the discussion of the results of the triadic consensus of each cluster around three factors on the basis of types of connections between users – isolates, reciprocity and transitivity. The first category, isolates, focuses on the types of triads where no links exist between any three users. The second category of reciprocity looks at the types of triads where one or two links exist between the triplet of nodes. Although reciprocity is a measure that is generally characteristic of a dyadic relation, we use it to look at whether the (retweeting) interactions are being reciprocated for our triads of users [Solano and Dunnam, 1985]. Lastly, the third category of transitivity is concerned with set of three nodes with three or more established links.

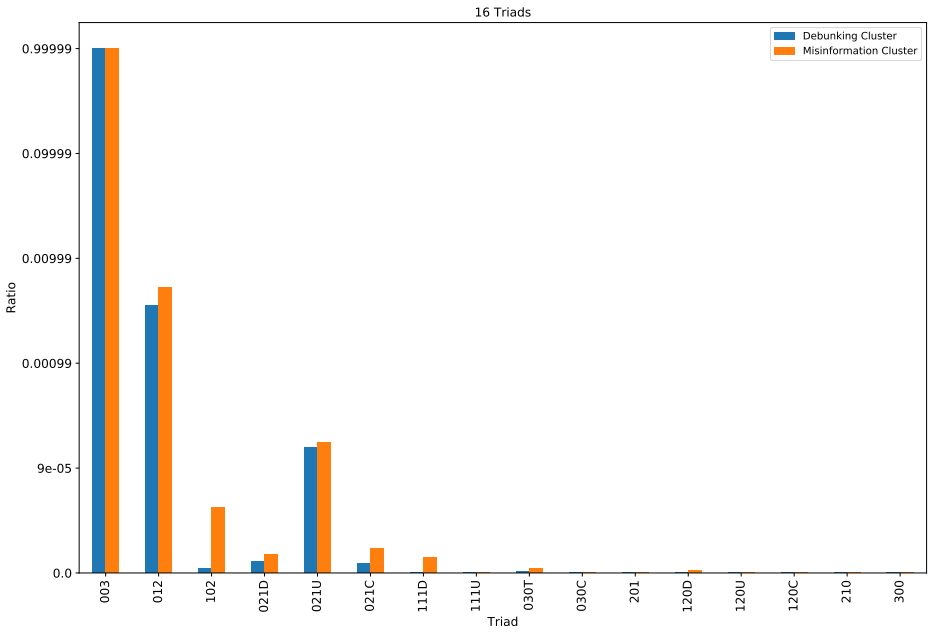


Figure 4.21: Triadic census

Isolates. We consider isolates to be the triadic set of users that are not connected to each other – i.e. a user triplet which is completely disconnected. There is one such type of triad, of the sixteen, that falls within this category: 003. From 4.21, we can see that there is an almost similar proportion of this type of triad in each cluster, with the debunking cluster consisting of a slightly higher proportion of this triad type. We can gather that there is a similar distribution of completely disconnected triplets of users in both the clusters – i.e. the distribution of user triplets that do not retweet each other is almost identical. Although this is the most commonly occurring type of triad in both the clusters, it is the two succeeding categories which highlight the contrast between how users connect within each cluster.

Reciprocity. When looking at reciprocity, we consider triads where one or two links exist between the set of three users. Using 4.20, we can deduce that there are five types of triads that are relevant: 012, 021C, 021D, 021U and 102. In context of these five triads, reciprocity can be defined as whether a user who retweets another in the triad is retweeted back. While there is only one type of triad (102) that fully encapsulates reciprocity between the users, the other four triad types also give us an insight into varying degree of reciprocity.

4.21 demonstrates that significantly higher proportion of each of these five triad types are present in the misinformation cluster than in the debunking cluster. This is particularly the case for 102 where we can observe that the misinformation cluster users are much more likely to get retweeted back by a user that they retweeted – this is indicative of a high level of reciprocity within the misinformation cluster users.

With respect to the other four triad types, a higher proportion of these triads being present within the misinformation cluster generally suggests that the users who spread misinformation are not only more likely to retweet one or more of their peers, but also are more likely to be retweeted by another user.

Transitivity. To evaluate the level of transitivity within each cluster, we need at least three ties to exist between users. Hence, we focus on the types of triads that consist of three or more connections between users and 4.20 depicts that there are ten such triad types. Complete transitivity within the triads can be defined as the existence of a link directed from user A to user B, a link from user B to user C and a third link from user A to user C – i.e. triad type 030T. The other nine types of remaining triads possess a varying level of transitivity. However, seven of these are not present in either of the cluster and so we will focus on the following three: 111D, 030T and 120D.

From 4.21, we can see that two of these triad types – 111D and 120D – are unique to the misinformation cluster. We can infer from this that there are groups of misinformation cluster users that possess the quality of reciprocity between two users in addition to one or both of these users also being retweeted by a third user. We can also note that there are more cases of 030T present in the misinformation cluster than in the debunking cluster. The collectively higher proportion of these three types

of triads in the misinformation cluster reflects that the cluster is partially transitive – more so than the debunking cluster.

Summary. The triadic census allows us to understand the distribution of the sixteen types of triads that can exist within a directed social network. Through looking at each of these types of triads, we find that the misinformation cluster has more reciprocity between its users and is significantly more partially-transitive than the debunking cluster. Essentially, this means that the misinformation cluster users not only exclusively retweet each other but also are more likely to retweet a connection-of-a-connection (edge separation of 2) which is also supported by our findings from evaluating closeness centrality of the cluster.

4.3.4 Diffusion Analysis

The fourth type of analysis that we perform is diffusion analysis. Having looked at the content, the users and the organisation of the network, the logical next step is to analyse how this discussion spread. We look at the tweeting behaviour and the diffusion patterns of the #deathtax discussion that occurred online on Twitter.

4.3.4.1 Exposure over time

We begin the investigation by examining when the misinformation and debunking narratives start to surface. We do this by looking at the cumulative sum of exposure over time that is achieved by each narrative. Exposure, in this research, is defined to be the number of people that were exposed to a tweet which is optimistically calculated to be the follower-ship of a tweet's author. This is repeated for all of the tweets to produce a timeline of exposure which can be seen in 4.22.

Insights. This graph illustrates that there was an early spread of the misinformation cluster narrative in February 2019. This narrative rapidly gains traction until the election day (indicated by the solid black line) and then plateaus immediately after.

On the other hand, we can see that the debunking cluster's narrative starts to appear on Twitter towards the end of April and gains exponential exposure subsequent to the day of election. This narrative continues to gain traction long after the elections and reaches a significantly larger number of people than the misinformation cluster's narrative after the elections have concluded.

Summary. We can piece together that the users perpetuating the #deathtax rumour began circulating their narrative well in-advance to the election day and the debunking narrative was created as a response to the misinformation that was being spread about the #deathtax. Hence, the misinformation narrative seems to be more of an organised effort to spread misinformation about the death tax and the debunking

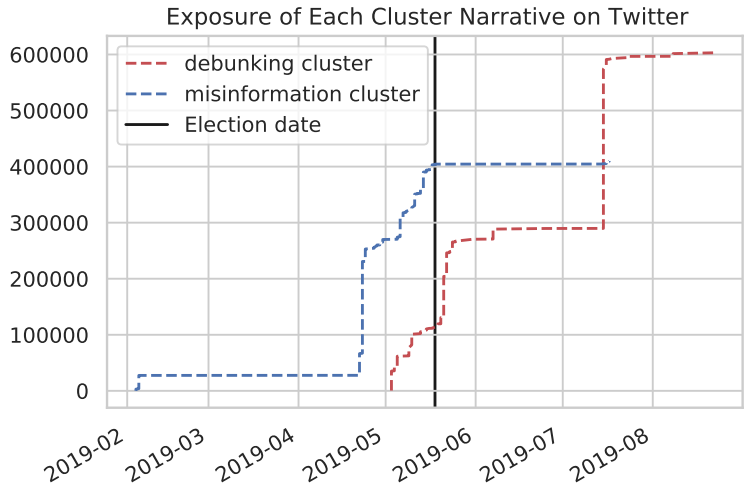


Figure 4.22: Exposure over time of each narrative

narrative seems to have originated as a reaction to this effort – as so far also observed in all three other levels of analysis.

4.3.4.2 Retweeting Patterns

The next tool we use to perform diffusion analysis is retweeting patterns. We look at the retweeting behaviour of the users belonging to each cluster. We do this by looking at every single tweet from each cluster and its retweet and comparing the time it took a user to retweet the original tweet. We repeat this for all tweets and plot the results that can be seen in 4.23. 4.23 is a log plot of the distribution of time differences between a tweet and its retweet where the x axis is log, base 10, in hours of these time differences.

Insights. From the graph, we can tell there is a difference in retweeting behaviour of the users from the two clusters. From 4.23, we can see that the misinformation distribution curve has a fatter tail, and (visually) a larger σ value, compared to the debunking cluster.

From this, we can observe that the users within the misinformation cluster not only retweet suspiciously quick but also take longer to retweet. The suspiciously quick retweeting can be attributed to making content go viral and the suspiciously long time difference can be attributed to the intention of pushing old messages back onto people's timelines. We find that the users in misinformation cluster have an abnormal retweeting behaviour compared to users belonging to the debunking cluster.

Summary. Through examining the retweeting behaviour of each cluster's users, we

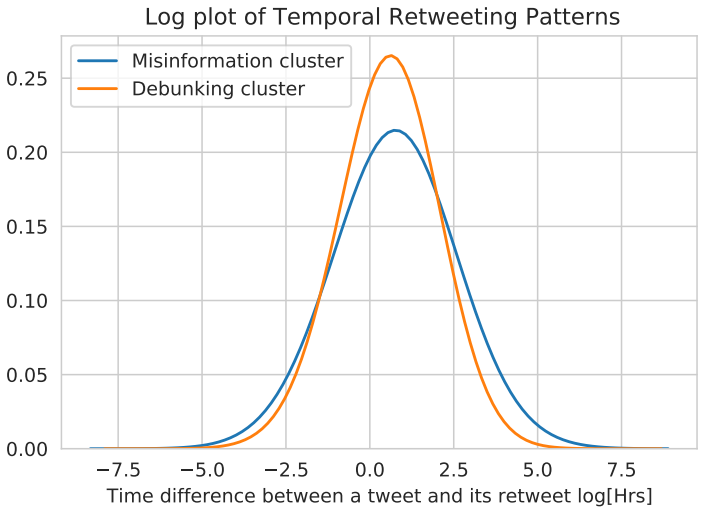


Figure 4.23: Temporal retweeting patterns on a log scale of each cluster

find that the misinformation cluster users exhibit an unusual twofold behaviour. We find that unnaturally quick retweeting times as well as abnormally long retweeting times are both characteristic of how these users retweet one another.

4.3.4.3 Diffusion Cascades

The last type of analysis we perform to understand the diffusion profile of our network is through inspecting diffusion cascades for each of the clusters. A cascade, sometimes known as an information cascade, is an event where several people sequentially make similar decisions. In the context of this dataset, the event is a chain of tweets – i.e. a tweet is retweeted. The cascade data we use is described in 4.2.

We take this cascade data and plot the Complementary Cumulative Distribution Function (CCDF) as can be seen in 4.24. CCDF is mathematically defined as:

$$\bar{F}_X(x) = P(X > x) = 1 - F_X(x) \tag{4.5}$$

where $F_X(x)$ is the Cumulative Distribution Function (CDF). CCDF is the opposite of CDF, i.e. it calculates how frequently a randomly distributed variable is above a certain threshold X. In 4.24, the x axis is the size of each cascade and the y axis is the CCDF of the cascade sizes that were calculated.

Insights. This graph demonstrates that the misinformation cluster has more cascades that are smaller in size (< 80) compared to the debunking cluster. This is reversed for cascade sizes greater than 80 where the debunking cluster has more

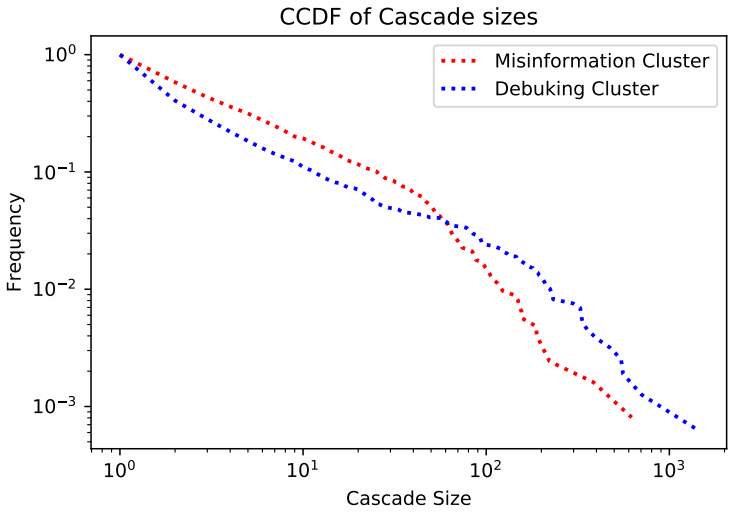


Figure 4.24: CCDF of the cascade sizes of each cluster

of these cascades compared to the misinformation cluster. Hence, we find that the users in the misinformation cluster are retweeted frequently but by a smaller groups of users from the same cluster. On the other hand, the users within the debunking cluster are retweeted more frequently by a larger group of other debunking cluster members.

Summary. Analysing the size of retweet cascades of both clusters, we find that existence of smaller, but frequent in number of, diffusion cascades is a feature that can be traced back to users who spread mis-/dis-information. Hence, a user who is involved in numerous small diffusion cascades can likely be categorised as a suspicious actor. Note, this finding is in agreement with the results obtained from computing the betweenness centrality where we draw a similar conclusion.

4.4 Link to Traditional Media

After analysing our network in depth, we link the spread of misinformation on online social media platforms to traditional media to understand how misinformation has been leaking into media such as news and articles. We compare the timeline of #deathtax's appearance on traditional media versus social media. We also look at how traditional news sources are used by users in both the clusters to evaluate the reliability of each narrative.

4.4.1 A Timeline

We begin by mapping out a timeline of first mentions of #deathtax on news and articles and comparing it to when #deathtax first started to circulate on social media. We also research news articles and news programs that aired on television and find that the first mention of #deathtax was in an article that was released by the Daily Telegraph on the 21^{st} July 2018 which stated that the Australian Council of Trade Unions supported an inheritance tax.

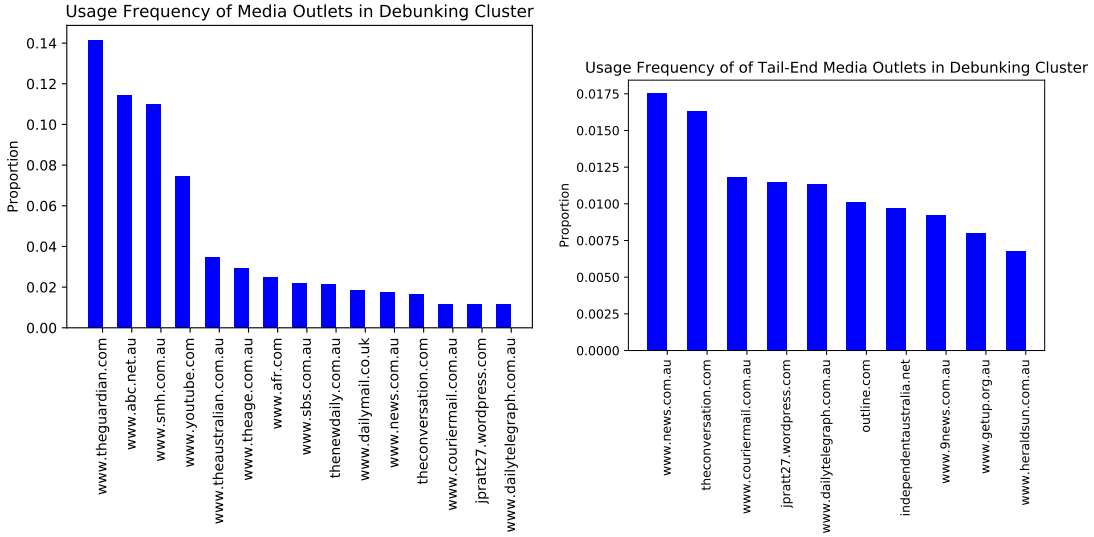
A day later, on the 22^{nd} July 2018, there was a segment on the popular Australian morning program 'Sunrise' which featured a discussion on 'Should there be death duties?'. On 24^{th} January 2019 there was an official release by Josh Frydenberg, then Australian Treasurer, warning about Labor's supposed Death Tax. This official release by Frydenberg was picked up and circulated by misinformation cluster users. It seems to be the case that the first mention of #deathtax appears in traditional media before it leaks onto social media.

4.4.2 References of Media outlets

We attempt to characterise how each cluster uses traditional media to perpetuate their respective arguments. We use the URL links data, described in 4.2, consisting of news and article links from the tweets belonging to each cluster to analyse the media sources referenced. Then we graph the fifteen most frequent media outlets referenced by the debunking cluster in 4.25(b) and by the misinformation cluster in 4.26(a). To capture the full picture, we also depict the ten least frequently referenced media outlets – for the debunking cluster, this can be found in 4.25(b) and for misinformation cluster, it is 4.26(b). The results for this are all in form of a bar graph where the y axis indicates the relative usage, i.e. the usage compared to all the other media sources.

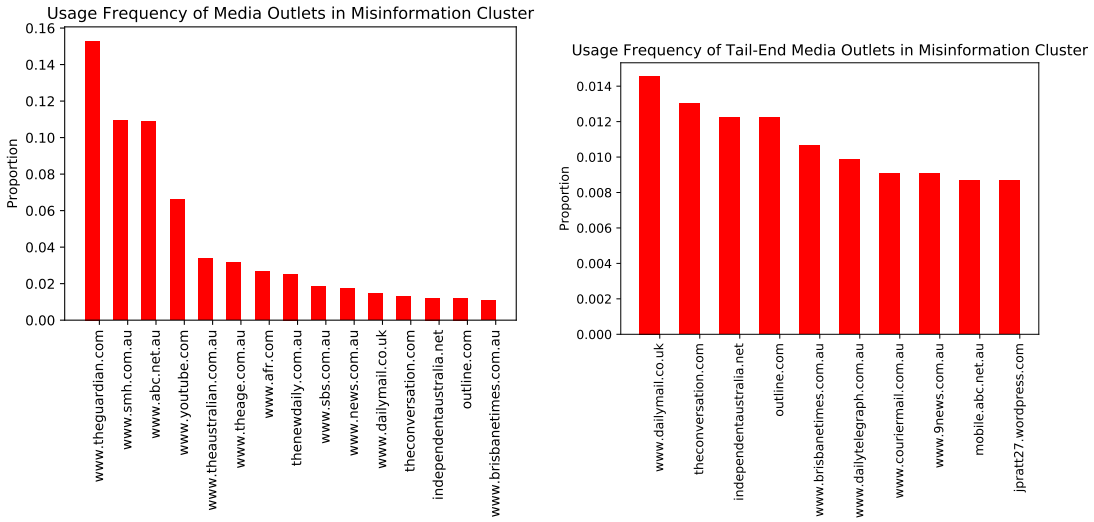
Debunking Cluster. We find, from 4.25(a), the ten most frequently sourced media outlets by the users in the debunking cluster are all reliable and popular news sources with the likes of 'The Guardian' and 'Australian Financial Review'. The anomaly, in terms of credibility, lies in the last five of the fifteen most sourced outlets - 'Wordpress'. It is noteworthy, that the three most frequently quoted media – 'The Guardian', 'ABC' and 'Sydney Morning Herald' – outlets all have relatively high usage compared to every other news source and have high credibility as sources.

On the other hand, looking at the least frequently used news sources in 4.25(b), tells us that even the lesser referenced media outlets such as '9news' are reliable. There are a few anomalies with 'getup' and again 'wordpress'. However, overall we can verify that the credibility and reliability of the news sources quoted by the users within the debunking cluster is high.



(a) 15 most frequently linked media sources (b) 15 least frequently linked media sources

Figure 4.25: Usage of traditional media by debunking cluster users



(a) 15 most frequently linked media sources (b) 15 least frequently linked media sources

Figure 4.26: Usage of traditional media by misinformation cluster users

Misinformation Cluster. Looking at 4.26(a), in the case of the misinformation cluster, we can see that the top three most frequently sourced media outlets are the same as for the debunking cluster, with the order of 'Sydney Morning Herald' and 'ABC' swapped. We find that the ten most frequently referred to media sources for the misinformation cluster looks quite similar to the ten most frequently quoted sources for the debunking cluster. The real difference between the usage of media outlets of the two clusters lies in the right-end of the bar graph 4.26(a) for the misinformation cluster. For example, the misinformation cluster users quote 'thenewsdaily' more than 'sbs'. The opposite, however, is true for the debunking cluster where links from 'sbs' are referred to more, than articles/news from 'thenewsdaily'.

From 4.26(b), we can see that the users in the misinformation clusters refer to some unreliable sources such as "wordpress" - similar to the users in the debunking cluster. Also similar to the debunking cluster, the credibility of these least frequently used sources is relatively good.

Summary. Overall, the major difference we observe between the usage of media outlets between the two clusters is how frequently each outlet is used rather than which outlets are linked.

4.5 Summary

To summarise, in this chapter, we present a list of instances of online misinformation spread during the 2019 Australian elections in 4.1. We discuss instances such as 'Misleading Signs' and 'Stop Adani' billboard before introducing 'Death tax' – the chosen instance to deep-dive into. We also discuss using hashtags and keyword search on phrases 'death tax' and 'inheritance tax' to filter the #auspol dataset in 4.2. In this section, we first introduce our network of Twitter users' interaction in 4.7 and exemplify the existence of two different clusters of users within the network – one more densely packed compared to the other.

Content Analysis. We then elaborate on how we conduct the four-tiered analysis on the #Deathtax dataset in 4.3 – beginning with content analysis on the interaction map. We dive into the message content of the tweets belonging to each cluster and identify one of the clusters as the misinformation cluster – where the users attempt to perpetuate misinformation about #deathtax – and the other cluster as the debunking cluster – where the users post tweets pertaining to falsehood of the #deathtax claims. We also generate word clouds for the content of the tweets and the hashtags for each of the cluster and find that the misinformation cluster users interleave #deathtax discussions to other topical political issues such as 'coal' and 'defence funding' to inject any pre-existing bias of those topics into the #deathtax debate. The debunking cluster text word cloud, on the other hand, demonstrates that the users within that cluster try to convey that #deathtax is 'fake news' and a 'scare campaign' and focus only on the death tax.

Another interesting finding lies within how misinformation users' leverage hashtags to attract as much traction possible to their cause – through excessively tagging their tweets with popular news sources. Lastly, we also perform sentiment analysis on the tweet content of each cluster using analysers such as VADER and TextBlob. Results obtained from TextBlob suggest that the tweets originating from the misinformation cluster are carefully produced to give an impression of objectivity. Furthermore, insights from sentiment analysis show that these users attempt to adopt an almost neutral tone in their message – simultaneously shying away from use of strong emotion in their tweets. In contrast, the users that belong to the debunking cluster appear to be more liberal in their expression, of both negative and positive emotion, and the reactionary nature of their narrative's delivery supports the higher level of subjectivity that is detected in their tweets.

User Analysis. We also present the second analytical lens, the user analysis, where user-level analysis is performed on each of our clusters. We begin by identifying users from each cluster that have been at the core of the #deathtax debate on Twitter and are retweeted the most – we call these users opinion leaders. We show that 60% of the misinformation cluster's opinion leaders have inactive accounts whereas none of the debunking cluster's opinion leaders have inactive accounts. Furthermore, only 20% of the misinformation opinion leaders have verified accounts but this number is significantly higher for the debunking cluster at 60%. Furthermore, we find that the opinion leaders of the debunking cluster have a significantly higher number of status counts and have accounts that were established over a decade ago. However, the misinformation cluster opinion leaders tend to have a much lower count of statuses/tweets posted and have relatively new accounts – created within the last 5 years.

We present our findings from investigating account activity of the users from each cluster where we uncover that 63% of all inactive or suspended accounts in the network belong to the misinformation cluster. Additionally, we find that approximately 85% of the verified users in the network belong to the debunking cluster and the users within the misinformation cluster are far less likely to have a verified account. Lastly, for user-level analysis, we also show that the users within the misinformation cluster have a slightly higher mean friends count and a noticeably higher mean followers count. Additionally, we find the comparatively low variability in these account features of misinformation cluster users to be a further indication of inorganic behaviour.

Network Analysis. Next, we present our third analytical tool to scrutinise the network with – network-level analysis. Here, we show that there are noteworthy variations in the way the two clusters are organised. We use centrality measures – such as in-degree, betweenness, closeness as well as eigenvector centrality – and triads to extract these differences for each cluster. We find that the in-degree centrality for the nodes in the misinformation cluster is higher than the in-degree for the nodes in the debunking cluster which reveals that the users in the misinforma-

tion cluster exclusively retweet each other far more than the users in the debunking cluster. Additionally, we find this to be a similar case for the betweenness centrality, where the abnormally high betweenness values for the misinformation cluster users signifies that the misinformation cluster users are more connected. Furthermore, this also provides evidence of ingenuity due to the discrepancy in the high betweenness centrality values and the smaller cluster size.

Further validating the earlier results, we also demonstrate that both the closeness and eigenvector centrality measures for the misinformation cluster nodes are higher. For the case of closeness centrality, this further solidifies the finding that the cluster is strongly connected in nature. We also find that the universally high values of eigenvalue centrality observed within the misinformation cluster hints at existence of premeditation in the cluster organisation. Finally, we use social network theory of triads to show that the misinformation cluster is not only more connected but is also partially-transitive.

Diffusion Analysis. The fourth cornerstone that we detail in 4.3 is the diffusion analysis where we document looking at the diffusion pattern of the two clusters. We present the investigation into the exposure each cluster's narrative gets over time and show that the misinformation cluster began spreading their narrative much earlier in February 2019 before abruptly ceasing soon after the election day. On the contrary, the debunking cluster begins tweeting about #deathtax a few months after the misinformation cluster and the narrative gains the maximum traction some time after the election day. We demonstrate that the misinformation cluster begins perpetuating the misinformation about #deathtax early and before suspiciously stopping immediately after the election day which reiterates the presence of an organised effort at spreading misinformation.

We also document our examination into the retweeting behaviour of each cluster's users and show that the misinformation cluster users exhibit two types of behaviour – where they are not only suspiciously quick in retweeting other users but also retweet other users or themselves after a long time, as if to push old messages back onto other user's timelines. In comparison, the debunking cluster's users have a more natural retweeting behaviour where the distribution curve has one mode – the behaviour is consistent. Lastly, for the diffusion analysis, we also show that the diffusion cascade differ for each cluster and the misinformation cluster has a larger number of small-sized diffusion cascades whereas the debunking cluster has a greater number of larger-sized cascades. The overarching conclusion we reach is that at each level of the four types of analysis, there is evidence of inorganic behaviour as well as premeditation in the organisation of the misinformation cluster.

Traditional Media. Finally, we explore how the traditional media is connected to the #deathtax misinformation campaign in 4.4. We investigate the timeline of the appearance of the death tax on traditional media and find that it first appeared on traditional media outlets such as news programs and articles before trickling into social media posts such as Twitter and Facebook. We also analyse the use of

traditional media sources within the Twitter debate through analysing the links to these media outlets in our #deathtax tweets. Through this, we find that the two clusters' references to media outlets in their tweets are similar – both in credibility and political polarity.

Chapter 5

Conclusion

Mis-/dis-information spread has been a topic of rising concern shared by the research community, governments around the world and social media giants alike. Unbridled circulation of mis-/dis-information on social media platforms can allow for serious consequences, for us and our society, to fester. Current relevant works on this problem offer limited counsel on a unified approach to aid in detection of mis-/dis-information. Furthermore, there is also a gap in the existing literature, culminated by little research effort that has been allocated to finding a standardised set of features or tell-tales of a classic mis-/dis-information campaign.

In this research, we studied if organized disinformation can be identified and detected through addressing the following four research questions we posed at the beginning:

1. Are instances of mis-/dis-information in the Australian context? If so, what do these instances look like?
2. What is an all-encompassing analytical approach that we can use to examine these instances of mis-/dis-information spread?
3. What is a set of characteristics of a mis-/dis-information campaign?
4. How is traditional media linked to social media when looking at mis-/dis-information spread?

Within our research, we addressed the first question by compiling a list of cases of attempted opinion manipulation during the 2019 Australian elections and found 'death tax' to be the most prominent of the cases. We also find that, in this instance, there are two clusters of users – one community that attempts to spread misinformation about #deathtax; and the other which attempts to debunk the myth.

We addressed the second question by introducing our four-cornerstone approach to the analysis that we conduct on the #deathtax dataset. This approach constitutes of analysing four different levels of the dataset which are: content level, user level,

network level and diffusion level. For **content analysis**, we consider the content of the messages that are posted in the debate; as well as generate word frequency clouds and conduct sentiment analysis. As a part of a **user-level** analysis, we looked at the account characteristics, such as account activity/inactivity and friends and followers count, of all users who are involved in the debate on Twitter. We also identify opinion leaders, i.e. users who are most retweeted on the subject, for each side of the debate and further compare their account features such as verified status.

For **network-level** analysis, we considered the network of users and evaluated the centrality measures, e.g. in-degree, closeness, betweenness and eigenvector. We also employed the theory of triads to understand the relationship between users and how they are connected. For the fourth level, **diffusion analysis**, we studied the diffusion patterns, by looking at retweeting behaviour of users, and compared the information cascades of each side of the discussion. We also mapped and compared the traction gained over time by each side of the #deathtax debate.

We addressed the third question by obtaining findings and results from applying the novel four-cornerstone analysis approach to my dataset. We illustrated that there are a set of noticeable features characteristic of a mis-/dis-information campaign that are highlighted when compared with the debunking cluster. We showed that, at the content level, the users intending to perpetuate misinformation tend to post messages infusing multiple other controversial and political issues. At the user level, we found that the users that belong to the misinformation cluster are more likely to have suspended or inactive accounts and less likely to have verified accounts.

Through conducting network analysis, we demonstrated that the users in the misinformation cluster are more connected and have a network structure that appears to be more organised in comparison to the debunking cluster. Furthermore, modelling the exposure gained over time, as a part of diffusion analysis, showed that the misinformation about the death tax was being circulated some months prior to the election day and this ceased immediately after the election day which hinted at an organised effort. Finally, through analysing retweeting patterns, we found that the retweeting behaviour of users in the misinformation cluster is twofold: suspiciously quick retweeting times, as if to bombard other users' timelines with their messages, and suspiciously long retweeting times – with the likely intention of pushing old messages back onto users' timelines.

Finally, we addressed the last question by linking the coverage of #deathtax on traditional media and social media which showed that the first mention of the death tax was on news programs before leaking into the social media sphere. Additionally, we showed that the usage of traditional media outlets as sources in tweets is similar for both debunking and misinformation clusters users.

In lieu of the above, we present our main contributions of this research to be:

1. Exploring a case study with an Australian context – performing analysis on an Australian dataset.
2. Compiling, and gathering evidence for, a list of cases of mis-/dis-information campaigns in Australia.
3. Constructing a four-tiered approach which can be utilised to analyse data with multiple lenses
4. Demonstrating evidence of inorganic behaviour in relation to the online discourse of #deathtax campaign.
5. Extracting and presenting noteworthy features of an instance of a mis-/dis-information campaign
6. Exploring the part traditional media plays in the spread of mis-/dis-information on social media.

Although we have made considerable strides in addressing the research questions in this thesis, further work is required before our findings can be realistically adopted.

5.1 Future Work

Despite the volume of work explored in this research, there still remains much potential for further exploration in this field. For example, this research explores mis-/dis-information spread in one instance – the #deathtax. A potential extension to this research could be to emulate the four-cornerstone analysis presented in this paper on various other instances with careful consideration of multidisciplinary contexts. For example, this could take form of looking at instances of mis-/dis-information spread in a medical or sociological context rather than another political. Replication of this work on other instances will allow for comparison of findings and insights. Consistency of these findings across the instances can be used to validate if they can be generalised. Additionally, these generalised findings can be collated to form a set of defining features of a disinformation campaign.

Another potential avenue for future work is a more in-depth analysis into libraries and packages that can be used integrated our approach to obtain enhanced results. Due to time constraints, we did not have enough time to compare or evaluate a wider set of tools for our base analysis. For example, within sentiment analysis, we only compare VADER and TextBlob instead of a wider range. Furthermore, each of these tools can be configured using a variety of analysers and lexicons. Hence, finding the tools and their configurations that produce the best results is an example of further research that can be undertaken to improve and build on our results.

Another research avenue, with significant implications and contributions, that can be explored as an extension to this research is to use our findings, and any further findings from duplication of this work on other case studies, to build a model or an agent which can be utilised to detect or identify disinformation campaigns in real-time. For example, the generalised findings from our research can be used as input features to a binary neural net classifier that can be used to classify whether a network of users are attempting to spread disinformation. Furthermore, this can be extended to give a score, instead of a binary prediction, to a network indicative of the likelihood of it being an instance of disinformation campaign. These models can be embedded into social media platforms like Twitter to automate the real-time detection of disinformation spread.

Also due to the time constraints, our exploration of disinformation flow between traditional and social media was limited to a basic level. There is large scope for further analysis into understanding how disinformation flows from traditional to social media or conversely flow of disinformation from social to traditional media. This can be further extended to explore questions such as how mutation of narratives and content occurs as it flows between traditional and social media.

Other potential avenues include introducing depth and complexity into each facet of the four-tiered analysis approach – for e.g. using the theory of Natural Language Processing (NLP) as one of the sub-components of content-level analysis.

Bibliography

Aboufarw, K.; Grigorev, A.; and Mihaita, A., 2022. Traffic accident risk forecasting using vision transformers,. In *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*. (cited on page 9)

Allcott, H. and Gentzkow, M., 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31, 2 (2017), 211–36. (cited on page 7)

Belau, J., 2014. Consequences of connection failure-centrality and the importance for cohesion. (2014). (cited on page 43)

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008, 10 (2008), P10008. (cited on page 23)

Bode, L. and Vraga, E. K., 2018. See something, say something: correction of global health misinformation on social media. *Health communication*, 33, 9 (2018), 1131–1140. (cited on page 1)

Bovet, A.; Morone, F.; and Makse, H. A., 2018. Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump. *Scientific reports*, 8, 1 (2018), 1–16. (cited on page 7)

Bradshaw, S. and Howard, P. N., 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project*, 1 (2018). (cited on page 1)

Bruns, A. and Burgess, J., 2011. The use of twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European consortium for political research (ECPR) general conference 2011*, 1–9. The European Consortium for Political Research (ECPR). (cited on page 28)

Chen, X., 2016. The influences of personality and motivation on the sharing of misinformation on social media. *IConference 2016 Proceedings*, (2016). (cited on page 1)

Dawson, N.; Rizoiu, M.-A.; Johnston, B.; and Williams, M. A., 2019. Adaptively selecting occupations to detect skill shortages from online job ads. In *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 1637–1643. IEEE, Los Angeles, CA, USA. doi:10.1109/BigData47090.2019.9005967. http://arxiv.org/abs/1911.02302https://ieeexplore.ieee.org/document/9005967/. (cited on page 9)

Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W., 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113, 3 (2016), 554–559. (cited on page 8)

Figueira, A.; Guimaraes, N.; and Torgo, L., 2018. Current state of the art to detect fake news in social media: Global trendings and next challenges. In *WEBIST*, 332–339. (cited on page 2)

Forelle, M.; Howard, P. N.; Monroy-Hernández, A.; and Savage, S., 2015. Political bots and the manipulation of public opinion in venezuela. *Available at SSRN 2635800*, (2015). (cited on pages 2 and 7)

Grigorev, A.; Mihaita, A.; Saleh., K.; and Picardi, M., 2022a. Traffic incident duration prediction via a deep learning framework for text description encoding. In *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*. (cited on page 9)

Grigorev, A.; Mihaita, A.-S.; Lee, S.; and Chen, F., 2022b. Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation. *Transportation Research Part C: Emerging Technologies*, 141 (2022), 103721. doi:https://doi.org/10.1016/j.trc.2022.103721. https://www.sciencedirect.com/science/article/pii/S0968090X22001589. (cited on page 9)

Gupta, A.; Lamba, H.; and Kumaraguru, P., 2013. $1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter. In *2013 APWG eCrime researchers summit*, 1–12. IEEE. (cited on pages 2 and 5)

Hu, Y., 2005. Efficient, high-quality force-directed graph drawing. *Mathematica journal*, 10, 1 (2005), 37–71. (cited on page 20)

Hutto, C. and Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8. (cited on page 29)

Issa, F.; Monticolo, D.; Gabriel, A.; and Mihăiţă, A., 2014. An intelligent system based on natural language processing to support the brain purge in the creativity process. *IAENG International Conference on Artificial Intelligence and Applications (ICAIA'14) Hong Kong*, (Mar. 2014). (cited on page 9)

Jones, M. O., 2019. The gulf information war| propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. *International journal of communication*, 13 (2019), 27. (cited on page 33)

Kalpokas, I., 2017. Information warfare on social media: a brand management perspective. *Baltic Journal of Law & Politics*, 10, 1 (2017), 35–62. (cited on page 2)

Kong, Q.; Rizoiu, M.-A.; Wu, S.; and Xie, L., 2018. Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 175–178. ACM Press, Lyon, France. doi:10.1145/3184558.3186972. https://arxiv.org/abs/1801.04117http://dl.acm.org/citation.cfm?doid=3184558.3186972. (cited on page 9)

Kong, Q.; Rizoiu, M. A.; and Xie, L., 2020. Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *International Conference on Information and Knowledge Management, Proceedings*, 645–654. ACM, New York, NY, USA. doi:10.1145/3340531.3411861. https://arxiv.org/pdf/2001.11132.pdfhttps://dl.acm.org/doi/10.1145/3340531.3411861. (cited on page 9)

Mao, T.; Mihaita, A.; and Cai, C., 2019. Traffic signal control optimisation under severe incident conditions using genetic algorithm. *Proc. of ITS World Congress (ITSWC 2019), Singapore*, (Oct. 2019). (cited on page 9)

Mao, T.; Mihăită, A.-S.; Chen, F.; and Vu, H. L., 2022. Boosted genetic algorithm using machine learning for traffic control optimization. *Trans. Intell. Transport. Sys.*, 23, 7 (jul 2022), 7112–7141. doi:10.1109/TITS.2021.3066958. https://doi.org/10.1109/TITS.2021.3066958. (cited on page 9)

McSwiney, J., 2020. Social networks and digital organisation: far right parties at the 2019 australian federal election. *Information, Communication & Society*, (2020), 1–18. (cited on pages 2 and 6)

Mihaita, A.; LI, H.; and Rizoiu, M., 2020a. Traffic congestion anomaly detection and prediction using deep learning. doi:arXiv:2006.13215. (cited on page 9)

Mihaita, A. S.; Benavides, M.; Camargo, C.; and Cai, C., 2019a. Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models. *International Journal of Intelligent Transportation System Research (IJITSR)*, 17, 2 (2019), 125–141. doi:DOI:10.1007/s13177-018-0160-z. https://link.springer.com/article/10.1007/s13177-018-0160-z. (cited on page 9)

Mihaita, A. S.; Dupont, L.; Cherry, O.; Camargo, M.; and Cai, C., 2018. Air quality monitoring using stationary versus mobile sensing units: a case study from lorraine, france. *Proc. of ITS World Congress (ITSWC 2018), Copenhagen, Denmark*, (Sep. 2018). (cited on page 9)

Mihaita, A.-S.; Li, H.; He, Z.; and Rizoiu, M.-A., 2019b. Motorway Traffic Flow Prediction using Advanced Deep Learning. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 1683–1690. IEEE, Auckland, New Zealand. doi:10.1109/ITSC.2019.8916852. https://ieeexplore.ieee.org/document/8916852/. (cited on page 9)

Mihaita, A.-S.; Liu, Z.; Cai, C.; and Rizoiu, M.-A., 2019c. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. In *Pro-*

ceedings of the 26th ITS World Congress, 1–12. Singapore. http://arxiv.org/abs/1905.12254. (cited on page 9)

Mihaita, A.-S.; Papachatgis, Z.; and Rizoiu, M.-A., 2020b. Graph modelling approaches for motorway traffic flow prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* (Rhodes, 2020), 1–8. IEEE Press. doi:10.1109/ITSC45102.2020.9294744. https://doi.org/10.1109/ITSC45102.2020.9294744. (cited on page 9)

Mihaylov, T.; Georgiev, G.; and Nakov, P., 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the nineteenth conference on computational natural language learning*, 310–314. (cited on pages 2, 5, and 6)

Mihăiţă, A.; Camargo, M.; and Lhoste, P., 2014. Evaluating the impact of the traffic reconfiguration of a complex urban intersection. *10th International Conference on Modelling, Optimization and Simulation (MOSIM 2014), Nancy, France, 5-7 November 2014*, (Nov. 2014). (cited on page 9)

Mihăiţă, A. S.; Tyler, P.; Menon, A.; Wen, T.; Ou, Y.; Cai, C.; and Chen, F., 2017. An investigation of positioning accuracy transmitted by connected heavy vehicles using dsrc. *Transportation Research Board - 96th Annual Meeting, Washington, D.C.*, (Jan. 2017). (cited on page 9)

Mihăiţă, S. and Mocanu, S., 2011. An energy model for event-based control of a switched integrator. *IFAC Proceedings Volumes*, 44, 1 (2011), 2413–2418. doi:https://doi.org/10.3182/20110828-6-IT-1002.02082. https://www.sciencedirect.com/science/article/pii/S1474667016439741. 18th IFAC World Congress. (cited on page 9)

Mishra, S.; Rizoiu, M.-A.; and Xie, L., 2018. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, 1–10. Stanford, CA, USA. https://arxiv.org/pdf/1804.02101.pdf. (cited on page 9)

Monticolo, D. and Mihăiţă, A., 2014. A multi agent system to manage ideas during collaborative creativity workshops. *International Journal of Future Computer and Communication (IJFCC)*, 3, 1 (Feb. 2014), 66–70. doi:10.7763/IJFCC.2014.V3.269. (cited on page 9)

Prier, J., 2017. Commanding the trend: Social media as information warfare. *Strategic Studies Quarterly*, 11, 4 (2017), 50–85. (cited on page 2)

Quattrociocchi, W.; Conte, R.; and Lodi, E., 2011. Opinions manipulation: Media, power and gossip. *Advances in Complex Systems*, 14, 04 (2011), 567–586. (cited on page 8)

Rizoiu, M.-A.; Graham, T.; Zhang, R.; Zhang, Y.; Ackland, R.; and Xie, L., 2018. # debatenight: The role and influence of socialbots on twitter during the 1st 2016 us

presidential debate. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12. (cited on page 6)

Rizoiu, M. A. and Velcin, J., 2011. Topic extraction for ontology learning. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* (Eds. W. Wong; W. Liu; and M. Bennamoun), 38–60. IGI Global. ISBN 9781609606251. doi:10.4018/978-1-60960-625-1.ch003. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-625-1.ch003. (cited on page 9)

Rizoiu, M.-A. and Xie, L., 2017. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *International AAAI Conference on Web and Social Media (ICWSM '17)*, 182–191. Montréal, Québec, Canada. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553https://arxiv.org/pdf/1703.01012.pdf. (cited on page 9)

Rizoiu, M. A.; Xie, L.; Caetano, T.; and Cebrian, M., 2016. Evolution of privacy loss in Wikipedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, 215–224. ACM, ACM Press, New York, New York, USA. doi:10.1145/2835776.2835798. http://dl.acm.org/citation.cfm?doid=2835776.2835798http://arxiv.org/abs/1512.03523http://dx.doi.org/10.1145/2835776.2835798. (cited on page 9)

Shaffiei, S.; Mihaita, A.; and Cai, C., 2019. Demand estimation and prediction for short-term traffic forecasting in existence of non-recurrent incidents. *Proc. of ITS World Congress (ITSWC 2019), Singapore*, (Oct. 2019). (cited on page 9)

Shafiei, S.; Mihaita, A.; Nguyen, H.; Bentley, C. D. B.; and Cai, C., 2020. Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation. In *Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.* doi:http://hdl.handle.net/10453/138721. (cited on page 9)

Shafiei, S.; Mihăiţă, A.-S.; Nguyen, H.; and Cai, C., 2022. Integrating data-driven and simulation models to predict traffic state affected by road incidents. *Transportation Letters*, 14, 6 (2022), 629–639. doi:10.1080/19427867.2021.1916284. https://doi.org/10.1080/19427867.2021.1916284. (cited on page 9)

Shin, J.; Jian, L.; Driscoll, K.; and Bar, F., 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83 (2018), 278–287. (cited on pages 2 and 6)

Shu, K.; Wang, S.; and Liu, H., 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 430–435. IEEE. (cited on page 5)

Solano, C. H. and Dunnam, M., 1985. Two's company: Self-disclosure and reciprocity in triads versus dyads. *Social Psychology Quarterly*, (1985), 183–187. (cited on page 44)

Staiano, J.; Lepri, B.; Aharony, N.; Pianesi, F.; Sebe, N.; and Pentland, A., 2012. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, 321–330. (cited on page 44)

Starbird, K.; Maddock, J.; Orand, M.; Achterman, P.; and Mason, R. M., 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings*, (2014). (cited on page 7)

Thomas, T. L., 1996. Deterring information warfare: a new strategic challenge. *The US Army War College Quarterly: Parameters*, 26, 4 (1996), 12. (cited on page 2)

Unwin, J. T.; Routledge, I.; Flaxman, S.; Rizoiu, M. A.; Lai, S.; Cohen, J.; Weiss, D. J.; Mishra, S.; and Bhatt, S., 2021. Using hawkes processes to model imported and local malaria cases in near-elimination settings. *PLoS Computational Biology*, 17, 4 (apr 2021), e1008830. doi:10.1371/JOURNAL.PCBI.1008830. http://medrxiv.org/content/early/2020/07/17/2020.07.17.20156174.abstracthttps://dx.plos.org/10.1371/journal.pcbi.1008830. (cited on page 9)

Vaidya, T.; Votipka, D.; Mazurek, M. L.; and Sherr, M., 2019. Does being verified make you more credible? account verification's effect on tweet credibility. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. (cited on page 36)

Vraga, E. K. and Bode, L., 2018. I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, 21, 10 (2018), 1337–1353. (cited on page 1)

Wardle, C.; Derakhshan, H.; et al., 2018. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco*, (2018), 43–54. (cited on page 1)

Wen, T.; Mihăiţă, A.-S.; Nguyen, H.; Cai, C.; and Chen, F., 2018. Integrated incident decision-support using traffic simulation and data-driven models. *Transportation Research Record*, 2672, 42 (2018), 247–256. doi:10.1177/0361198118782270. https://doi.org/10.1177/0361198118782270. (cited on page 9)

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2019. Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW (nov 2019), 1–25. doi:10.1145/3359285. http://dl.acm.org/citation.cfm?doid=3371885.3359285. (cited on page 9)

Wu, S.; Rizoiu, M. A.; and Xie, L., 2020. Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, 715–725. https://arxiv.org/abs/2003.09557. (cited on page 9)

Zhang, R.; Walder, C.; and Rizoiu, M.-A., 2020. Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 04 (apr 2020), 6803–6810. doi:10.1609/aaai.v34i04.6160. http://arxiv.org/abs/1905.10496https://aaai.org/ojs/index.php/AAAI/article/view/6160. (cited on page 9)

Zhao, D.; Mihaita, A.; Ou, Y.; Shafiei, S.; Grzybowska, H.; Qin, K.; Tan, G.; and Li, M., 2022. Real-time attention-augumented spatio-temporal networks for video-based driver activity recognition. In *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*. (cited on page 9)