#ArsonEmergency and Australia's "Black Summer": Polarisation and misinformation on social media

Derek Weber[1,2][0000−0003−3830−9014], Mehwish Nasim[1,3,5,6][0000−0003−0683−9125], Lucia Falzon[4,2][0000−0003−3134−4351], and Lewis Mitchell[1,5][0000−0001−8191−1997]

[1] University of Adelaide, South Australia, Australia
{derek.weber,lewis.mitchell}@adelaide.edu.au
[2] Defence Science and Technology Group, Adelaide, Australia
derek.weber@dst.defence.gov.au
[3] Data61, Commonwealth Science and Industry Research Organisation, Australia
mehwish.nasim@data61.csiro.au
[4] School of Psychological Sciences, University of Melbourne
lucia.falzon@unimelb.edu.au
[5] ARC Centre of Excellence for Mathematical and Statistical Frontiers, Australia
[6] Cyber Security Cooperative Research Centre

Abstract. During the summer of 2019-20, while Australia suffered unprecedented bushfires across the country, false narratives regarding arson and limited backburning spread quickly on Twitter, particularly using the hashtag *#ArsonEmergency*. Misinformation and bot- and troll-like behaviour were detected and reported by social media researchers and the news soon reached mainstream media. This paper examines the communication and behaviour of two polarised online communities before and after news of the misinformation became public knowledge. Specifically, the *Supporter* community actively engaged with others to spread the hashtag, using a variety of news sources pushing the arson narrative, while the *Opposer* community engaged less, retweeted more, and focused its use of URLs to link to mainstream sources, debunking the narratives and exposing the anomalous behaviour. This influenced the content of the broader discussion. Bot analysis revealed the active accounts were predominantly human, but behavioural and content analysis suggests Supporters engaged in trolling, though both communities used aggressive language.

Keywords: Social Media · Information Campaigns · Polarisation · Misinformation · Crisis.

1 Introduction

People share an abundance of useful information on social media during a crisis situation [6, 5]. This information, if analysed correctly, can rapidly reveal population-level events such as imminent civil unrest, natural disasters, or accidents [26]. Not all content is helpful, however: different entities may try to

popularise false narratives using sophisticated social bots and/or humans. The spread of such misinformation not only makes it difficult for analysts to use Twitter data for public benefit [21] but may also encourage large numbers of people to believe false narratives, which may then influence public policy and action, and can be particularly dangerous during crises [18].

This paper presents a case study of the dynamics of misinformation propagation during one such crisis. The 2020 Australian 'Black Summer' bushfires burnt over 16 million hectares, destroyed over 3,500 homes, and caused at least 33 human and a billion animal fatalities[1], and attracted global media attention. We show that:

- Significant Twitter discussion activity accompanied the Australian bushfires, influencing media coverage.
- In the midst of this, narratives of misinformation began to circulate on social media, including that:
  - the bushfires were caused by arson;
  - preventative backburning efforts were reduced due to green activism;
  - Australia commonly experiences such bushfires; and
  - climate change is not related to bushfires.

All of these narratives were refuted, e.g., the arson figures being used were incorrect[2], preventative backburning has limited effectiveness[3], the fires are "unprecedented"[4], and climate change is, in fact, increasing the frequency and severity of the fires[5]. The Twitter discussion surrounding the bushfires made use of many hashtags, but according to research by Graham & Keller [13] reported on ZDNet [25], the arson narrative was over-represented on $\#ArsonEmergency$, likely created as a counter to the pre-existing $\#ClimateEmergency$ [2]. Furthermore, their research indicated that $\#ArsonEmergency$ was being boosted by bots and trolls. This attracted widespread media attention, with most coverage debunking the arson conspiracy theory. This case thus presents an interesting natural experiment: the nature of the online narrative before the publication of the ZDnet article and then after these conspiracy theories were debunked.

We offer an exploratory mixed-method analysis of the Twitter activity using the term 'ArsonEmergency' around ($\pm$7 days) the publication of the ZDNet

[1] https://www.abc.net.au/news/2020-02-19/australia-bushfires-how-heat-and-drought-created-a-tinderbox/11976134

[2] https://www.abc.net.au/radionational/programs/breakfast/victorian-police-reject-claims-bushfires-started-by-arsonists/11857634

[3] https://www.theguardian.com/australia-news/2020/jan/08/hazard-reduction-is-not-a-panacea-for-bushfire-risk-rfs-boss-says.

[4] The Australian Academy of Science's statement: https://www.science.org.au/news-and-events/news-and-media-releases/statement-regarding-australian-bushfires

[5] Science Brief, on 14 January 2020, reports on a survey of 57 papers on the matter conducted by researchers from the University of East Anglia, Imperial College, London, Australia's CSIRO, the Univerity of Exeter and the Met Office Hadley Centre, Exeter: https://sciencebrief.org/briefs/wildfires

article, including comparison with another prominent contemporaneous bushfire-related hashtag, *#AustraliaFire*. A timeline analysis revealed three phases of activity. Social network analysis of retweeting behaviour identifies two polarised groups of Twitter users: those promoting the arson narrative, and those exposing and arguing against it. These polarised groups, along with the unaffiliated accounts, provide a further lens through which to examine the behaviour observed. A content analysis highlights how the different groups used hashtags and other sources to promote their narratives. Finally, a brief analysis of bot-like behaviour then seeks to replicate Graham & Keller's findings [13].

Our contribution is two-fold: 1) we offer an original, focused dataset from Twitter at a critical time period covering two eras in misinformation spread[6]; and 2) insight into the evolution of a misinformation campaign relating to the denial of climate change science and experience in dealing with bushfires.

1.1 Related Work

The study of Twitter during crises is well established [6, 5, 11], and has provided recommendations to governments and social media platforms alike regarding its exploitation for timely community outreach. The continual presence of trolling and bot behaviour diverts attention and can confuse the public at times of political significance [15, 7, 21, 22] as well as creating online community-based conflict [16, 8] and polarisation [12].

Misinformation on social media has also been studied [17]. In particular, the disinformation campaign against the White Helmets rescue group in Syria is useful to consider here [24]. Two clear corresponding clusters of pro- and anti-White Helmet Twitter accounts were identified and used to frame an investigation of how external references to YouTube videos and channels compared with videos embedded in Twitter. They found the anti-White Helmet narrative was consistently sustained through "sincere activists" and concerted efforts from Russian and alternative news sites. These particularly exploited YouTube to spread critical videos, while the pro-White Helmet activity relied on the White Helmets' own online activities and sporadic media attention. This interaction between supporter and detractor groups and the media may offer insight into activity surrounding similar crises.

1.2 Research Questions

Motivated by our observations, we propose the following research questions about Twitter activity during the 2019-20 Australian bushfire period:

RQ1 To what extent can an online misinformation community be discerned?
RQ2 How did the spread of misinformation differ between the identified phases, and did the spread of the hashtag *#ArsonEmergency* differ from other emergent discussions (e.g., *#AustraliaFire*)?

[6] https://github.com/weberdc/socmed_sna

RQ3 How does the online behaviour of those who accept climate science differ from those who refute or question it? How was it affected by media coverage exposing how the *#ArsonEmergency* hashtag was being used?

RQ4 To what degree was the spread of misinformation facilitated or aided by troll and/or automated bot behaviour?

In the remainder of this paper, we describe our mixed-method analysis and the datasets used. A timeline analysis is followed by the polarisation analysis. The revealed polarised communities are compared from behavioural and content perspectives, as well as through bot analysis. Answers to the research questions are summarised and we conclude with observations and proposals for further study of polarised communities.

2 Dataset and Timeline

The primary dataset, 'ArsonEmergency', consists of 27,456 tweets containing this term posted by 12,872 unique accounts from 31 December 2019 to 17 January 2020. The tweets were obtained using Twitter's Standard search Application Programming Interface (API)[7] by combining the results of searches conducted with Twarc[8] on 8, 12, and 17 January. As a contrast, the 'AusFire' dataset comprises tweets containing the term 'AustraliaFire' over the same period, made from the results of Twarc searches on 8 and 17 January. 'AusFire' contains 111,966 tweets by 96,502 accounts. Broader searches using multiple related terms were not conducted due to time constraints and in the interests of comparison with Graham & Keller's findings [13]. Due to the use of Twint[9] in that study, differences in dataset were possible, but expected to be minimal. Differences in datasets collected simultaneously with different tools have been previously noted [27]. Live filtering was also not employed, as the research started after Graham & Keller's findings were reported.

This study focuses on about a week of Twitter activity before and after the publication of the ZDNet article [25]. Prior to its publication, the narratives that arson was the primary cause of the bushfires and that fuel load caused the extremity of the blazes were well known in the conservative media [2]. The ZDnet article was published at 6:03am GMT (5:03pm AEST) on 7 January 2020, and was then reported more widely in the MSM morning news, starting around 13 hours later. We use these temporal markers to define three dataset phases:

- *Phase 1*: Before 6am GMT, 7 January 2020;
- *Phase 2*: From 6am to 7pm GMT, 7 January 2020; and
- *Phase 3*: After 7pm GMT, 7 January 2020.

[7] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets
[8] https://github.com/DocNow/twarc
[9] https://github.com/twintproject/twint

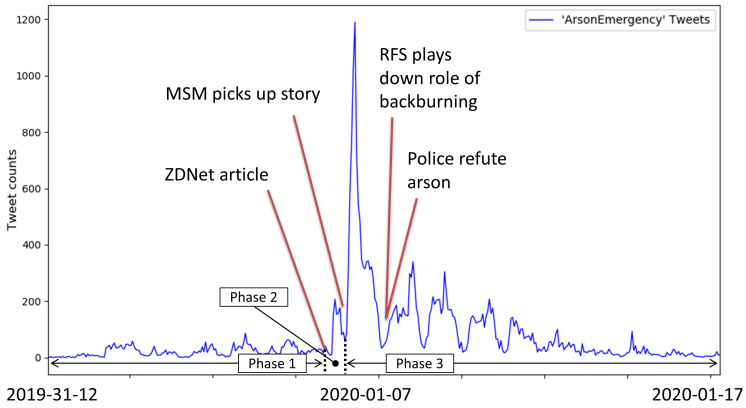


Fig. 1. Tweet activity in the 'ArsonEmergency' dataset, annotated with notable real-world events and the identified phases.

Figure 1 shows the number of tweets posted each hour in the 'ArsonEmergency' dataset, and highlights the phases and notable events including: the publication of the ZDNet article; when the story hit the MSM; the time at which the Rural Fire Service (RFS) and Victorian Police countered the narratives promoted on the *#ArsonEmergency* hashtag; and the clear subsequent day/night cycle. The RFS and Victorian Police announcements countered the false narratives promoted in political discourse in the days prior.

Since late September 2020, Australian and international media had reported on the bushfires around Australia, including stories and photos drawn directly from social media, as those caught in the fires shared their experiences. No one hashtag had emerged to dominate the online conversation and many were in use, including *#AustraliaFires*, *#ClimateEmergency*, *#bushfires*, and *#AustraliaIsBurning*.

The use of *#ArsonEmergency* was limited in Phase 1, with the busiest hour having around 100 tweets, but there was an influx of new accounts in Phase 2. Of all 927 accounts active in Phase 2 (responsible for 1,207 tweets), 824 (88.9%) of them had not posted in Phase 1 (which had 2,061 active accounts). Content analyses revealed 1,014 (84%) of the tweets in Phase 2 were retweets, more than 60% of which were retweets promoting the ZDNet article and the findings it reported. Closer examination of the timeline revealed that the majority of the discussion occurred between 9pm and 2am AEST, possibly inflated by a single tweet referring to the ZDNet article (at 10:19 GMT), which was retweeted 357 times. In Phase 3, more new accounts joined the conversation, but the day/night cycle indicates that the majority of discussion was local to Australia (or at least its major timezones).

The term 'ArsonEmergency' (sans '#') was used for the Twarc searches, rather than '#ArsonEmergency', to capture tweets that did not include the hashtag but were relevant to the discussion. Of the 27,546 tweets in the 'Arson-

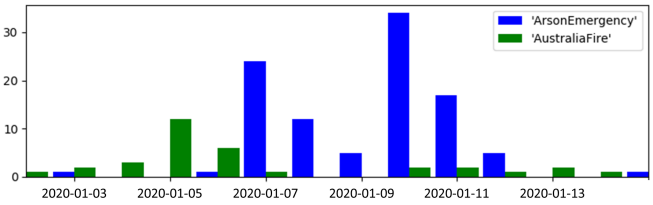


Fig. 2. Counts of tweets using the terms 'ArsonEmergency' and 'AustraliaFire' without a '#' symbol from 2–15 January 2020 in meta-discussion regarding each term's use as a hashtag (counts outside were zero).

Emergency' dataset, only 100 did not use it with the '#' symbol, and only 34 of the 111,966 'AustraliaFire' tweets did the same. Figure 2 shows the emergence of the reflexive discussion generated by those conversing about the discussion on *#ArsonEmergency* without promulgating the hashtag itself.

3 Polarisation in the Retweet Network

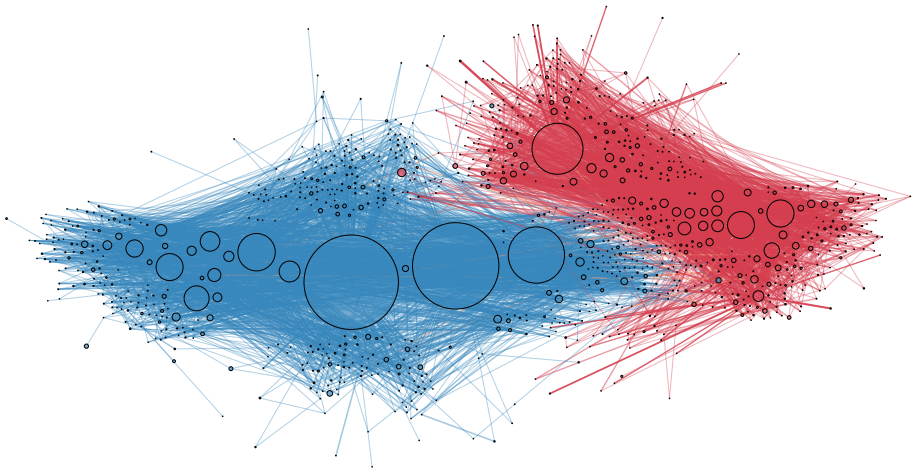


Fig. 3. Polarised retweets graph about the arson theory. Left(blue): *Opposers*, right(red): *Supporters* of the arson narrative. Nodes represent users. An edge between two nodes means one retweeted the tweet of the other. Node size corresponds to degree centrality.

There is no agreement on whether retweets imply endorsement or alignment. Metaxas *et al.* [19] studied retweeting behaviour in detail by conducting user surveys and studying over 100 relevant papers referring to retweets. Their findings conclude that when users retweet, it indicates interest and agreement as

well as trust in not only the message content but also in the originator of the tweet. This opinion is not shared by some celebrities and journalists who put a disclaimer on their profile: "retweets ≠ endorsements". Metaxas *et al.* [19] also indicated that inclusion of hashtags strengthens the agreement, especially for political topics. Other motivations, such as the desire to signal to others to form bonds and manage appearances [10], serve to further imply that even if retweets are not endorsements, we can assume they represent agreement or an appeal to likemindedness at the very least.

We conducted an exploratory analysis on the retweets graph shown in Figure 3. The nodes indicate Twitter accounts. An edge between two accounts shows that one retweeted a tweet of the other. Using conductance cutting [4], we discovered two distinct well-connected communities, with a very low number of edges between the two communities. Next, we selected the top ten accounts from each community based upon the degree centrality (most retweeted), manually checked their profiles, and hand labelled them as *Supporters* and *Opposers* of the arson narrative[10]. The accounts have been coloured accordingly in Figure 3: red nodes are accounts that promoted the narrative, while blue nodes are accounts that opposed them.

$\#ArsonEmergency$ had different connotations for each community. Supporters used the hashtag to reinforce their existing belief about climate change, while Opposers used this hashtag to refute the arson theory. The arson theory was a topic on which people held strong opinions resulting in the formation of the two strongly connected communities. Such polarised communities typically do not admit much information flow between them, hence members of such communities are repeatedly exposed to similar narratives, which further strengthens their existing beliefs. Such closed communities are also known as *echo chambers*, and they limit people's information space. The retweets tend to coalesce within communities, as has been shown for Facebook comments [20].

These two groups, Supporters and Opposers, and those users unaffiliated with either group, are used to frame the remainder of the analysis in this paper.

3.1 Behaviour

User behaviour on Twitter can be examined through the features used to connect with others and through content. Here we consider how active the different groups were across the phases of the collection, and then how that activity manifested itself in the use of mentions, hashtags, URLs, replies, quotes and retweets.

Considering each phase (Table 1) Supporters used $\#ArsonEmergency$ nearly fifty times more often than Opposers, which accords with Graham & Keller's findings that the false narratives were significantly more prevalent on that hashtag compared with others in use at the time [25, 13]. In Phase 2, during the Australian night, Opposers countered with three times as many tweets as Supporters, including fewer hashtags, more retweets, and half the number of replies,

[10] Labelling was conducted by the first two authors independently and then compared.

Table 1. Activity of the polarised retweeting accounts, by interaction type broken down by phase.

	Group		Tweets	Accounts	Hashtags	Mentions	Quotes	Replies	Retweets	URLs
Phase 1	Supporters	*Raw count*	1,573	360	2,257	2,621	185	356	938	405
		Per account	4.369	—	1.435	1.666	0.118	0.226	0.596	0.257
	Opposers	*Raw count*	33	21	100	35	8	2	20	9
		Per account	1.571	—	3.030	1.061	0.242	0.061	0.606	0.273
Phase 2	Supporters	*Raw count*	121	77	226	159	11	29	74	24
		Per account	1.571	—	1.868	1.314	0.091	0.240	0.612	0.198
	Opposers	*Raw count*	327	172	266	476	7	14	288	31
		Per account	1.901	—	0.813	1.456	0.021	0.043	0.881	0.095
Phase 3	Supporters	*Raw count*	5,278	474	7,414	7,407	593	1,159	3,212	936
		Per account	11.135	—	1.405	1.403	0.112	0.220	0.609	0.177
	Opposers	*Raw count*	3,227	585	3,997	3,617	124	95	2,876	359
		Per account	5.516	—	1.239	1.121	0.038	0.029	0.891	0.111
Overall	Supporters	*Raw count*	6,972	497	9,897	10,187	789	1,544	4,224	1,365
		Per account	14.028	—	1.420	1.461	0.113	0.221	0.606	0.196
	Opposers	*Raw count*	3,587	593	4,363	4,128	139	111	3,184	399
		Per account	6.049	—	1.216	1.151	0.039	0.031	0.888	0.111

demonstrating different behaviour to Supporters, which actively used the hashtag in conversations. Content analysis confirmed this to be the case. This is evidence that Supporters wanted to promote the hashtag to promote the narrative. Interestingly, Supporters, having been relatively quiet in Phase 2, produced 64% more tweets in Phase 3 than Opposers, using proportionately more of all interactions except retweeting, and many more replies, quotes, and tweets spreading the narrative by using multiple hashtags, URLs and mentions. In short, Opposers tended to rely more on retweets, while Supporters engaged directly and were more active in the longer phases.

The concentration of narrative from certain voices requires attention. To consider this, Table 2 shows the degree to which accounts were retweeted by the different groups by phase. We can immediately see that in Phases 1 and 3 Supporters relied on the content of a few accounts, retweeting accounts approximately three to five times more often than Opposers. Interestingly, unaffiliated accounts (of which there were many more, as shown in Table 2), retweeted the same accounts more often in the first two phases, but only half as often in Phase 3, which implies that Supporters focused their narrative through a limited pool of accounts.

Of the top 41 retweeted accounts (those retweeted 100 times or more in the dataset), 17 were Supporters and 20 Opposers. Supporters were retweeted $5,487$ times (322.8 retweets per account), while Opposers were retweeted $8,833$ times (441.7 times per account). Together, affiliated accounts contributed 93.3% of the top 41's $15,350$ retweets, in a dataset with $21,526$ retweets overall, and the

Table 2. Retweeting activity in the dataset, by phase and group.

Phase	Supporters			Opposers			Unaffiliated		
	Retweets	Retweeted Accounts	Retweets per account	Retweets	Retweeted Accounts	Retweets per account	Retweets	Retweeted Accounts	Retweets per account
1	938	77	12.182	20	8	2.500	1,659	105	15.800
2	74	21	3.524	288	31	9.290	652	60	10.867
3	3,212	74	43.405	2,876	228	12.614	11,807	532	22.194

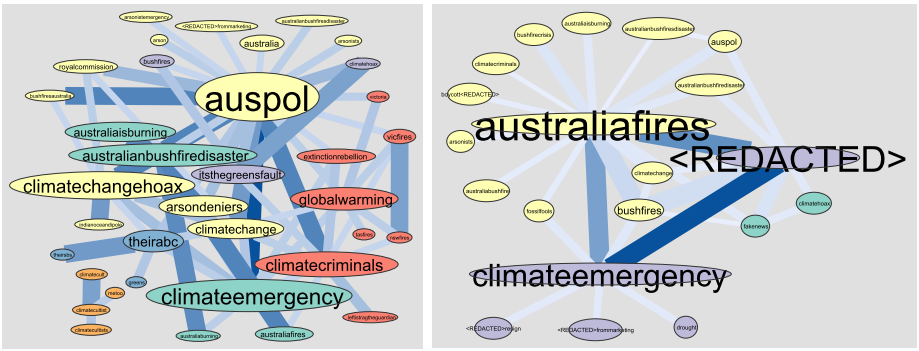
top 41 accounts were retweeted far more often than most. Thus Supporters and Opposers made up the majority of the most retweeted accounts.

3.2 Content

When contrasting the content of the two affiliated groups, we considered the hashtags and external URLs used. A hashtag can provide a proxy for a tweet's topic, and an external URL can refer a tweet's reader to further information relevant to the tweet, and therefore tweets that use the same URLs and hashtags can be considered related.

Hashtags. To discover *how* hashtags were used, rather than simply *which* were used, we developed co-mention graphs (Figure 4). Each node is a hashtag, sized by degree centrality; edges represent an account using both hashtags (not necessarily in the same tweet); the edge weight represents the number of such accounts in the dataset. Nodes are coloured according to cluster detected with the widely used Louvain method [3]. We removed the *#ArsonEmergency* hashtag (as nearly each tweet in the dataset contained it) as well as edges having weight less than 5. Opposers used a smaller set of hashtags, predominantly linking *#AustraliaFires* with *#ClimateEmergency* and a hashtag referring to a well-known publisher. In contrast, Supporters used a variety of hashtags in a variety of combinations, mostly focusing on terms related to 'fire', but only a few with 'arson' or 'hoax', and linking to *#auspol* and *#ClimateEmergency*. Manual review of Supporter tweets included many containing only a string of hashtags, unlike the Opposer tweets. Notably, the *#ClimateChangeHoax* node has a similar degree to the *#ClimanteChangeEmergency* node, indicating Supporters' skepticism of the science, but perhaps also attempts by Supporters to join or merge the communities.

Manual inspection of Supporter tweets revealed that replies often consisted solely of "*#ArsonEmergency*" (one Supporter replied to an Opposer 26 times in under 9 minutes with a tweet just consisting of the hashtag, although in six of the tweets @mentions of other influential Twitter accounts were also included). This kind of behaviour, in addition to inflammatory language in other Supporter replies, suggests a degree of aggression, though aggressive language was also noted among Opposers. Only 1.7% of Opposer tweets included more than 5 hashtags, while 2.8% of Supporter ones did, compared with 2.1% unaffiliated.



(a) Supporter hashtags. (b) Opposer hashtags.

Fig. 4. Co-mentioned hashtags of Supporters and Opposers. Hashtag nodes are linked when five or more accounts tweeted both hashtags, and are coloured by cluster. <REDACTED> hashtags include identifying information. Heavy edges (with high weight) are thicker and darker.

External URLs. URLs in tweets can be categorised as *internal* or *external*. Internal URLs refer to other tweets in retweets or quotes, while external URLs are often included to highlight something about their content, e.g., as a source to support a claim. By analysing the URLs, it is possible to gauge the intent of the tweet's author by considering the reputation of the source or the argument offered.

We categorised[11] the top ten URLs used most by Supporters, Opposers, and the unaffiliated across the three phases, and found a significant difference between the groups. URLs were categorised into four categories:

NARRATIVE Articles used to emphasise the conspiracy narratives by prominently reporting arson figures and fuel load discussions.

CONSPIRACY Articles and web sites that take extreme positions on climate change (typically arguing against predominant scientific opinion).

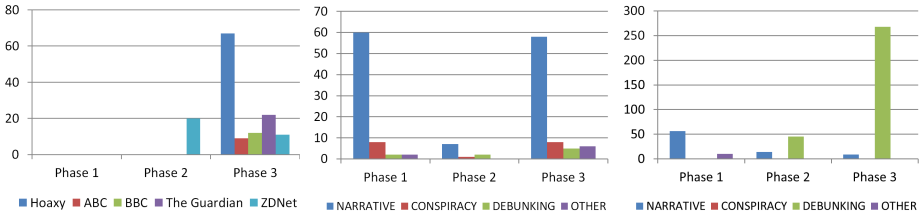
DEBUNKING News articles providing authoritative information about the bushfires and related misinformation on social media.

OTHER Other web pages.

URLs posted by Opposers were concentrated in Phase 3 and were all in the DEBUNKING category, with nearly half attributed to Indiana University's Hoaxy service [23], and nearly a quarter referring to the original ZDNet article [25] (Figure 5a). In contrast, Supporters used many URLs in Phases 1 and 3, focusing mostly on articles emphasising the arson narrative, but with references to a number of climate change denial or right wing blogs and news sites (Figure 5b).

[11] Categorisation was conducted by two authors and confirmed by the others.

Figure 5c shows that the media coverage changed the content of the unaffiliated discussion, from articles emphasising the arson narratives in Phase 1 to Opposer-aligned articles in Phase 3. Although the activity of Supporters in Phase 3 increased significantly, the unaffiliated members appeared to refer to Opposer-aligned external URLs much more often.



(a) Opposer URLs. (b) Supporter URLs. (c) Unaffiliated URLs.

Fig. 5. URLs used by Opposers, Supporters and unaffiliated accounts.

Supporters used many more URLs than Opposers overall (1,365 to 399) and nearly twice as many external URLs (390 to 212). Supporters seemed to use many different URLs in Phase 3 and overall, but focused much more on particular URLs in Phase 1. Of the total number of unique URLs used in Phase 3 and overall, 263 and 390, respectively, only 77 (29.3%) and 132 (33.8%) appeared in the top ten, implying a wide variety of URLs were used. In contrast, in Phase 1, 72 of 117 appeared in the top ten (61.5%), similar to Opposers' 141 of 212 (66.5%), implying a greater focus on specific sources of information. In brief, it appears Opposers overall and Supporters in Phase 1 were focused in their choice of sources, but by Phase 3, Supporters had expanded their range considerably.

4 Botness Analysis

The analysis reported in ZDNet [25] indicated widespread bot-like behaviour by using `tweetbotornot`[12]. Our re-analysis of this finding had two goals: 1) attempt to replicate Graham & Keller's findings in Phase 1 of our dataset; and 2) examine the contribution of bot-like accounts detected in Phase 1 in the other phases. Specifically, we considered the questions:

- Does another bot detection system find similar levels of bot-like behaviour?
- Does the behaviour of any bots from Phase 1 change in Phases 2 and 3?

We evaluated 2,512 or 19.5% of the accounts in the dataset using Botometer [9], including all Supporter and Opposer accounts, plus all accounts that posted at least three tweets either side of Graham & Keller's analysis reaching the MSM.

[12] https://github.com/mkearney/tweetbotornot

Botometer [9] is an ensemble bot classifier for Twitter accounts, relying on over a thousand features drawn from six categories. It includes a "Complete Automation Probability" (CAP), a Bayesian-informed probability that the account in question is "fully automated". This does not accommodate hybrid accounts [14] and only uses English training data [21], leading some researchers to use conservative ranges of CAP scores for high confidence that an account is human (<0.2) or bot (>0.6) [22]. We adopt that categorisation.

Table 3. Botness scores and contribution to the discussion across the phases.

			Active accounts			Tweets contributed		
Category	CAP	Total	Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3
Human	$0.0 - 0.2$	2,426	898	438	1,931	2,213	674	11,700
Undecided	$0.2 - 0.6$	66	20	6	56	28	11	304
Bot	$0.6 - 1.0$	20	9	4	11	23	6	84

Table 3 shows that the majority of accounts were human and contributed more than any automated or potentially automated accounts. This contrast with the reported findings [25] may be due to a number of reasons. The CAP score is focused on non-hybrid, English accounts, whereas `tweetbotornot` may provide a more general score, taking into account troll-like behaviour. The content and behaviour analysis discussed above certainly indicates Supporters engaged more with replies and quotes, consistent with other observed trolling behaviour [16] or "sincere activists" [24]. The collection tool used, Twint, may have obtained different tweets to Twarc, as it explicitly avoids Twitter's APIs. It is possible its avoidance of the API reveals more bot-like behaviour. Finally, it is unclear what Graham & Keller's collection strategy was; if it focused on the particular accounts which drew their attention to *#ArsonEmergency* to begin with, it may not have included the wider range of behaviour evident in our dataset.

5 Discussion

We are now well-placed to address our research questions:

RQ1 *Discerning a misinformation-sharing community.* Analysis revealed two distinct polarised communities. The content posted by the most influential accounts in these communities shows Supporters were responsible for the majority of arson-related content, while Opposers countered the arson narrative.

RQ2 *Differences in the spread of misinformation across phases and other discussions.* Considering URL and hashtag use in Phase 1 and 3, while the number of active Supporters grew from 360 to 474, the number of unique external URLs they used more than doubled, from 117 to 263. This was possibly due to the increased traffic on *#ArsonEmergency*. The number of hashtags

increased from 182 hashtags used 2,257 times to 505 hashtags used 7,414 times. This implies Supporters attempted to connect *#ArsonEmergency* with other hashtag-based communities. In contrast, Opposer activity increased from 33 hashtags used 100 times to 182 hashtags used 3,997 times, but Figure 4b shows Opposers focused the majority of their discussion on a comparatively small number of hashtags.

RQ3 *Behavioural differences over time and the impact of media coverage.* Supporters were more active in Phase 1 and 3 and used more types of interaction than Opposers, especially replies and quotes, implying a significant degree of engagement, whether as trolls or as "sincere activists" [24]. Supporters also relied on a smaller pool of accounts for retweeting than Opposers in the same phases. Supporters' use of interaction types remained steady from Phase 1 to 3. While behaviour remained relatively similar, activity grew for both groups after the story reached the MSM. The vast majority of accounts shared articles debunking the false narratives. The ZDNet article also affected activity, spurring Opposers and others to share the analysis it reported.

RQ4 *Support from bots and trolls.* We found very few bots, but aggressive troll-like behaviour was observed in the Supporter community. Aggressive language was observed in both affiliated groups. Distinguishing deliberate baiting from honest enthusiasm (even with swearing), however, is non-trivial [24].

The *#ArsonEmergency* activity on Twitter in early 2020 provides a unique microcosm to study the growth of a misinformation campaign before and after it was widely known. Our study reveals the following:

- Two clear polarised communities with distinct behaviour patterns and use of content were present.
- Supporters were more active and more engaged. Opposers relied on retweeting more, and focused on a few prominent hashtags, while Supporters used many. This was possibly to widely promote their message, or due to non-Australian contributors being unfamiliar with which hashtags to use for an Australian audience.
- The majority of Phase 1 *#ArsonEmergency* discussion referred to articles relevant to the arson narratives, but after the story reached the MSM, only the Supporter community continued to use such links.
- The majority of unaffiliated accounts shifted focus from CCD narrative-related articles in Phase 1 to debunking sites and articles in Phase 3. It is unclear whether the change in behaviour was driven by accounts changing opinion or the influx of new accounts.
- The *#ArsonEmergency* growth rate followed a pattern similar to another related hashtag that appeared shortly before it (*#AustraliaFire*).
- The influence of bot accounts appears limited when analysed with Botometer [9]. It classified 0.8% (20 of 2,512) of accounts as bots, and 96.6% (2,426 of 2,512) of the remaining accounts confidently as human. Graham & Keller had found an even spread of bot scores, with an average score over 0.5. Only 20% of accounts had a score ≤ 0.2 and 46% ≥ 0.6 [25].

Further research is required to examine social and interaction structures formed by groups involved in spreading misinformation to learn more about how such groups operate and better address the challenge they pose to society. Future work will draw more on social network analysis based on interaction patterns and content [1] as well as developing a richer, more nuanced understanding of the Supporter community itself, including more content and behaviour analysis.

Acknowledgment

The work has been partially supported by the Cyber Security Research Centre Limited whose activities are partially funded by the Australian Government's Cooperative Research Centres Programme.

Ethics

All data was collected, stored and analysed in accordance with Protocols H-2018-045 and #170316 as approved by the University of Adelaide's human research ethics committee.

References

1. J. P. Bagrow, X. Liu, and L. Mitchell. Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3(2):122–128, 2019.
2. P. Barry. Broadcast 3rd February 2020: News Corps Fire Fight. *Media Watch, Australian Broadcasting Corporation*, 2020(1), Feb. 2020.
3. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
4. U. Brandes, M. Gaertler, and D. Wagner. Engineering graph clustering: Models and experimental evaluation. *ACM Journal of Experimental Algorithmics*, 12:1.1:1–1.1:26, 2007.
5. A. Bruns and J. Burgess. #qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods. Research Report 48241, ARC Centre of Excellence for Creative Industries and Innovation, Jan. 2012.
6. A. Bruns and Y. E. Liang. Tools and methods for capturing twitter data during natural disasters. *First Monday*, 17(4), Apr. 2012.
7. CREST. Russian interference and influence measures following the 2017 UK terrorist attacks. Policy Brief 17-81-2, Centre for Research and Evidence on Security Threats, Cardiff University, Dec. 2017.
8. S. Datta and E. Adar. Extracting inter-community conflicts in Reddit. In *ICWSM*, pages 146–157. AAAI Press, 2019.
9. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. BotOrNot: A system to evaluate social bots. In *WWW (Companion Volume)*, pages 273–274. ACM, 2016.
10. L. Falzon, C. McCurrie, and J. Dunn. Representation and analysis of Twitter activity: A dynamic network perspective. In *ASONAM*, pages 1183–1190. ACM, 2017.

11. T. Flew, A. Bruns, J. Burgess, K. Crawford, and F. Shaw. Social media and its impact on crisis communication: Case studies of Twitter use in emergency management in Australia and New Zealand. In *2013 ICA Shanghai Regional Conference: Communication and Social Transformation*, Nov. 2014.
12. V. R. K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Polarization on social media. In *WWW (Tutorial Volume)*. ACM, 2018.
13. T. Graham and T. R. Keller. Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight. https://theconversation.com/bushfires-bots-and-arson-claims-australia-flung-in-the-global-disinformation-spotlight-129556, Jan. 2020. (Accessed on 2020-02-07).
14. C. Grimme, D. Assenmacher, and L. Adam. Changing perspectives: Is it sufficient to detect social bots? In *HCI (13)*, volume 10913 of *Lecture Notes in Computer Science*, pages 445–461. Springer, 2018.
15. F. B. Keller, D. Schoch, S. Stier, and J. Yang. How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. In *ICWSM*, pages 564–567. AAAI Press, 2017.
16. S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky. Community interaction and conflict on the web. In *WWW*, pages 933–943. ACM, 2018.
17. S. Kumar and N. Shah. False information on web and social media: A survey. *CoRR*, abs/1804.08559, 2018.
18. E. Kušen and M. Strembeck. You talkin' to me? Exploring human/bot communication patterns during riot events. *Information Processing & Management*, 57(1):102126, 2020.
19. P. T. Metaxas, E. Mustafaraj, K. Wong, L. Zeng, M. O'Keefe, and S. Finn. What do retweets indicate? results from user survey and meta-review of research. In *ICWSM*, pages 658–661. AAAI Press, 2015.
20. M. Nasim, M. U. Ilyas, A. Rextin, and N. Nasim. On commenting behavior of Facebook users. In *HT*, pages 179–183. ACM, 2013.
21. M. Nasim, A. Nguyen, N. Lothian, R. Cope, and L. Mitchell. Real-time detection of content polluters in partially observable twitter networks. In *WWW (Companion Volume)*, pages 1331–1339. ACM, 2018.
22. M.-A. Rizoiu, T. Graham, R. Zhang, Y. Zhang, R. Ackland, and L. Xie. #DebateNight: The role and influence of socialbots on Twitter during the 1st 2016 U.S. Presidential debate. In *ICWSM*, pages 300–309. AAAI Press, 2018.
23. C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *WWW (Companion Volume)*, pages 745–750. ACM, 2016.
24. K. Starbird and T. Wilson. Cross-Platform Disinformation Campaigns: Lessons Learned and Next Steps. *Harvard Kennedy School Misinformation Review*, Jan. 2020.
25. Stilgherrian. Twitter bots and trolls promote conspiracy theories about Australian bushfires — ZDNet. https://www.zdnet.com/article/twitter-bots-and-trolls-promote-conspiracy-theories-about-australian-bushfires/, Jan. 2020. (Accessed on 2020-01-28).
26. J. Tuke, A. Nguyen, M. Nasim, D. Mellor, A. Wickramasinghe, N. Bean, and L. Mitchell. Pachinko prediction: A bayesian method for event prediction from social media data. *Information Processing & Management*, 57(2):102147, 2020.
27. D. Weber, M. Nasim, L. Mitchell, and L. Falzon. Reliability of near real time social media data for social network analysis. In *HT*. ACM, 2020. Submitted.