Investigating Language Use by Polarised Groups on Twitter: A Case Study of the Bushfires

Mehwish Nasim
The University of Western Australia
ARC Centre of Excellence for
Mathematical and Statistical Frontiers
Australia
mehwish.nasim@uwa.edu.au

Naeha Sharif
The University of Western Australia
Australia
naeha.sharif@uwa.edu.au

Pranav Bhandari
Flinders University
Australia
bhan0098@flinders.edu.au

Derek Weber
Defence Science and Technology
Group
Australia
derek.weber@defence.gov.au

Martin Wood
Defence Science and Technology
Group
Australia
martin.wood@defence.gov.au

Lucia Falzon
The University of Adelaide
Australia
falzonlucia@gmail.com

Yoshihisa Kashima
The University of Melbourne
Australia
ykashima@unimelb.edu.au

ABSTRACT

Online social media platforms have become an important forum for public discourse, and have often been implicated in exacerbating polarisation in public sphere. Yet the precise mechanisms by which polarisation is driven are not fully understood. The study of linguistic style and features has been shown to be useful in exploring various aspects of online group discussions and, in turn, the processes which could contribute to polarisation. We present a case study around the hashtag #ArsonEmergency, collected from Australian Twittersphere during the unprecedented bushfires of 2019/2020. The dataset consists of two polarised groups and one unaffiliated group. We examine the linguistic style, moral language, and happiness profiles of 1786 users active during this catastrophic event. Our results suggest that polarised groups pushed 'affective polarisation' on Twitter while discussing the Australian Bushfires.

CCS CONCEPTS

• **Information systems → Social networking sites**; **Content analysis and feature selection**.

KEYWORDS

Moral language, Polarisation, Twitter, Bushfires

ACM Reference Format:
Mehwish Nasim, Naeha Sharif, Pranav Bhandari, Derek Weber, Martin Wood, Lucia Falzon, and Yoshihisa Kashima. 2022. Investigating Language Use by Polarised Groups on Twitter: A Case Study of the Bushfires. In *Australasian Document Computing Symposium (ADCS '22), December 15–16, 2022, Adelaide, Australia; Author Preprint.* ACM, New York, NY, USA, 7 pages. https://doi.org/TBA

1 INTRODUCTION AND BACKGROUND

Despite the potential for decentralized information sharing and public participation in online social networks, they have often been implicated in exacerbating polarisation in recent times. Indeed, online polarisation is a broad and well-studied topic [9], including in the context of politics [1]. As social media has increasingly been used for political communications, election campaigns have become rich sources of study of polarisation [e.g., 26].

However, the precise mechanisms that drive polarisation on social media are not fully understood. For instance, according to a cross-platform comparison in Israel [14], Facebook, a social media platform that supports a stable social network, was *less* polarising than Twitter, a platform that is not designed for a stable social network, but is better characterized as a dynamically evolving network of topics [16]. This suggests that polarisation may not necessarily be driven by an echo chamber - like-minded people mutually reinforcing their opinions within a stable social network.

In this context, language use amongst polarised groups may shed further light on the mechanisms of polarisation in American politics [10, 17], particularly in the online realm [2, 20]. Recent work from [8] focused on content polarity based on ground truth data. They argued that n-gram features were useful for identifying partisans in their data, as they indicate a distinctive writing style for such users; they also argued in favour of other powerful natural language processing (NLP) techniques, suggesting that they might be more helpful in identifying polarised users. Researchers have also analysed how language is used in the 2016 election cycle, including in Presidential debates (transcripts) and amongst followers [17].

ADCS '22, December 15–16, 2022, Adelaide, Australia; Author Preprint
© 2022 Copyright held by the owner/author(s).
ACM ISBN TBA.
https://doi.org/TBA

One potential mechanism of polarisation is moral language use. This is because moral language tends to justify opinions by referring to uncompromising moral foundations [11]. This can further strengthen emotional commitment and exacerbate 'affective polarisation' [14]. Recent research [22] studied the Twitter messages of members of Congress from 2016-2018 and speeches given on the floor from 1981 to 2017, and found higher use of moral language by members of a party with *lower political power*. For example, Democrats used more moral language in the period after the Republicans won the 2016 presidential election. Such moral language use may have further fueled emotional reactions to their election loss and entrenched further polarisation.

Contribution: We examine a non-political online activity during Australia's "Black Summer" bushfires of 2019/2020. It was a major natural disaster, which burnt more than 16 million hectares of south-eastern Australia[1]. Although tweets about Australian bushfires may seem to be a far cry from political polarisation in the USA, a clear polarisation was observed in Twitter [23]. We explore whether moral and emotional language use observed in American politics can also be found in the seemingly non-political Australian Twittersphere. To this end, we analyse the use of moral language of two polarised groups and a neutral group in a dataset about Australian bushfires [23]. We look at the use and emotional valance of moral language, as well as correspondence between the linguistic style and content of the tweets by the three groups.

2 METHODOLOGY

2.1 Data

In December 2019/January 2020, Australia suffered unprecedented bushfires. Although the consensus in the scientific community is that climate change aggravated the bushfires, a significant minority of the broader community disagrees. During Dec'19 - Jan'20, the hashtag #ArsonEmergency trended on Twitter. Weber et al. [23]'s Twitter dataset focuses on this topic, and was obtained using period historical searches for the term #ArsonEmergency. The authors examined a variety of interaction networks, botness scores, and trolling behavior in the dataset. They also provided a labeled dataset of polarised groups identified with community detection on the dataset's retweet network and labeled via manual inspection of a selection of the accounts. Three labeled groups are included: *Supporters* – the community using #ArsonEmergency to promote the idea that arson was the primary cause of the bushfires; *Opposers* – the community that countered the arson narrative; and, *Unaffiliated* – discussion participants who were not members of either polarised community. By October 2020, *Supporters* had lost a sixth of their members to deletion or suspension, while *Opposers* had lost around 3% [24].

We considered the language use of all the *Supporters* and *Opposers* and a random 7% selection of *Unaffiliated* profiles, a total of 1,786 users. The dataset has 399 *Supporter* accounts, 553 *Opposer* accounts and 834 *Unaffiliated accounts*.

2.2 Hypotheses

Studies in the political polarisation context have shown that people with extreme opinions, regardless of ideology, tend to share similar characteristics in language use. Frimer [6] compared the language used by liberals and conservatives and found that conservatives used more moral words that fall into the category of authority and purity, fewer in the loyalty category, and no more or less in the category of harm or fairness. To test the use of moral language and linguistic style on a contentious topic of bushfires, we state the following null hypotheses:

$H1a$: The difference in mean for the percentage of moral words used by the three groups (supporters, opposers and unaffiliated) for each aspect of moral foundation [11] is 'equal to zero' and the alternative hypothesis is that the difference in mean is not equal to zero.

$H1b$: The difference in mean for the percentage of moral words used by a pair of groups, irrespective of the category of moral foundation, is 'equal to zero', and the alternative hypothesis is that the difference in mean is not equal to zero.

$H2$: The difference in mean of the emotional valence of the language between the three groups is zero, and the alternative hypothesis is that the difference in mean is not equal to zero.

In addition to the above hypotheses we also tested the similarity in linguistic style and content between the three groups.

2.3 Analysis

To test our hypothesis we conducted the following analysis:

Moral Scores: The *moral-foundation dictionary* [11] contains approximately 350 words based on five moral foundations: *authority, care, fairness, loyalty,* and *sanctity*. Each foundation has two aspects: *vice* (negative words) and *virtue* (positive words). In other words, *Moral Foundation Dictionary* is a dictionary-based tool that best epitomizes the use or misuse of the five moral foundations. Studies of mutual use of moral language by polarised political elites by Frimer et al. [7] and Wang and Inbar [22] provided motivation to apply the technique in this work. We computed the scores for each moral category on per profile basis within each group, to test the first hypothesis.

Hopp et al. [13] presented the *extended version of Moral Foundations Dictionary (called eMFD)*, where language representations were incorporated by demographic embedding. The extended version of MFD outperforms its previous versions of MFD because it relies on crowd-sourced and content-driven data. Therefore, this can be used consistently against various domains to study moral intuitions of political, social, and communicative effects.

Happiness Scores: We consider the emotional valence of each word in terms of 'positivity' and 'negativity'. Natural human language has a universal positivity bias and the emotional content of words is consistent between languages under translation [4]. We measured the happiness scores on a *per group basis and per users basis* using a happiness dictionary [5], which contains about 10,000 words in eight languages, each with a happiness score between 1 and 9. The higher the score the happier (or more positive) the word. In order to give a lower weight to the common words in each group, we used the term frequency and inverse document frequency (tf-idf) metric to compute the happiness scores for each group, using the top 10% most frequent words in the tweets. We use this metric to test the second hypothesis.

[1] From July 2019 to February 2020: https://www.abc.net.au/news/2020-02-19/australia-bushfires-how-heat-and-drought-created-a-tinderbox/11976134

	Supporters	**Opposers**	**Unaffiliated**
Supporters	N/A	-	Authority
Opposers	Care Loyalty Sanctity	N/A	Fairness Loyalty Sanctity
Unaffiliated	Care Fairness Sanctity	Loyalty Sanctity	N/A

Table 1: Pairwise comparison for each aspect of moral foundation between the three groups. Those mentioned had a zero difference between the means at $\alpha = 0.01$. The lower triangle represents virtue and upper triangle represents vice aspects.

Linguistic Content and Style: The choice of words in communication often reflects the personality and psychological states. They can be divided into two broad categories [21]: 1) *content words* and 2) *style words*. While *content words* (e.g., nouns, verbs, adjectives, and adverbs) provide the content of communication, *style words* (e.g., pronouns, interjections, and conjunctions) express social and psychological properties e.g., emotions. We used the *CMU Twitter Part-of-Speech (POS) Tagger* [18] to obtain POS Tags for each word in the tweets. For our experiments, we only categorized nouns, verbs, adjectives, and adverbs as content words and pronouns (first, second and third person), interjections, and conjunctions as style words. We computed the content and style embeddings for *Supporter*, *Opposer* and *Unaffiliated* tweets using the tf-idf values of the words corresponding to the content or style category, respectively. Finally, we analysed the *cosine similarity* between the three groups of users in terms of their style and content embedding.

3 RESULTS

Moral Scores: Few of the *moral-foundation dictionary*'s terms could be found in the tweets, and the mean percentage of the use of moral words was very small for each group. Since we computed the percentage on per user basis, we performed a series of t-tests to compare the distributions of the mean usage of moral words in %age (per user), for each aspect of the moral foundations and then for the total percentage of moral words, irrespective of which moral foundation category they belong to (again on per user basis). We have summarised the pairwise results for each aspect of the moral foundation in Table 1. We have only mentioned the foundations for which we failed to reject the null hypothesis (H1a), implying that the difference in the mean is equal to zero. All three groups have similar percentages for the virtue aspect of sanctity-related words. *Supporters* and *Opposers* also align on the virtue aspect of the loyalty foundation but not on anything else. The distributions for *Opposers* and *Unaffiliated* align on the vice aspects of fairness, loyalty and sanctity as well as on the virtue aspect of loyalty and sanctity.

For the total use of moral words, we calculated the mean for each user who had used at least one moral foundation word Table 3. Table 4 shows the frequency of moral words. *Unaffiliated* users used many more moral words than the others. A t-test found the difference between the *Supporter* and *Opposer* means was zero at $\alpha = 0.01$, hence we fail to reject the null hypothesis (H1b) for the two polarised

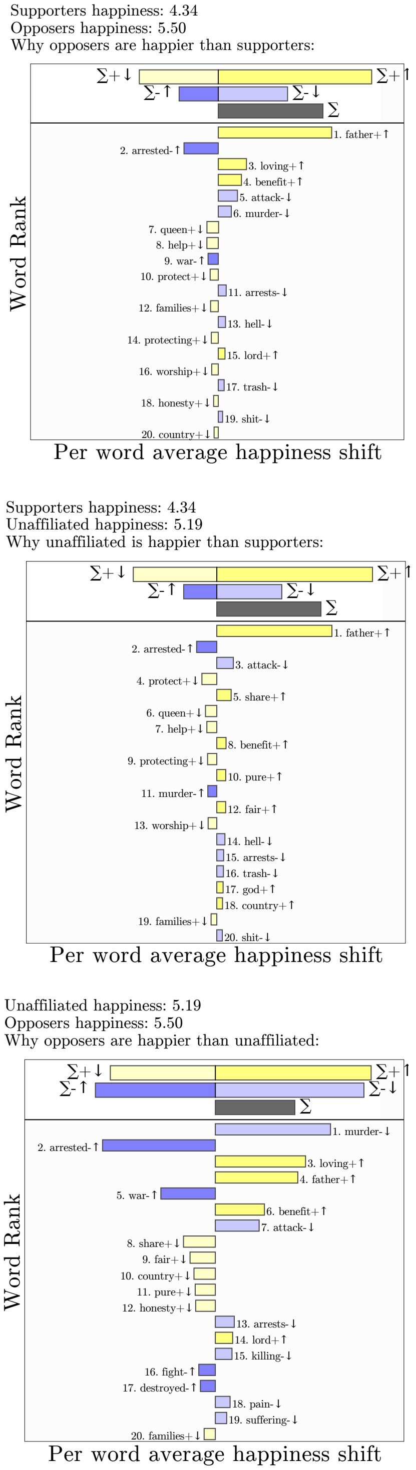

Figure 1: Word shift graphs of the per-word happiness shift for the *Opposer*, *Supporter* and *Unaffiliated* groups. Please note, we only used the words present in the moral foundations.

groups. The boxplot for each group is shown in Figure 4. The *Unaffiliated* group used the highest percentage of moral words (6.18%), followed by *Opposers* (2.73%) and then *Supporters* (2.70%). Research on moral psychology shows that people with lower power use more moral words [22], indicating the *Unaffiliated* resorted to presenting moral judgments in the discussion [3]. *Opposers* used a higher percentage of authority-related moral words (1.3%) than *Supporters* (1.06%); possibly as a way to attract attention to their point of view . An obvious limitation of this dictionary is that it assigns one foundation to each word. We then used another method [15] that uses bias and intensity. *Bias and Intensity* is calculated for every dictionary in the corpus towards the moral foundation axis following the FrameAxis method presented by Kwak et al. [15]. The relevance of a document to the moral foundation category is given by the term *bias*. The negative bias value depicts vice dimensions more, whereas the virtue dimension is shown if the bias value is positive. The relevance of the tweet to each moral foundation is identified by the term *intensity*. The bar charts representing the mean activation scores are presented in Figure 2.

Figure 2 shows the average moral scores for the data set. This shows that on average all the moral foundations care, authority, fairness, loyalty, and sanctity have a proportional contribution to the corpus. The vice domain is dominant on all the moral foundation axis, referring to the words present that do not represent high moral standards in conversations. This could be the result of a discrepancy in the thinking of the *Opposers* and *Supporters* community. It can also be observed that people do not necessarily make use of or do not consider the obligation to use moral words while convincing or imposing their thoughts on others. Although the virtue score for the care group is higher than others, it is still significantly behind compared to the vice score in the same group. Words such as 'damage', 'destroy', 'disease', 'lie', 'trait', 'disagree', and many other vice words are commonly used in the corpus, which justifies these scores. In the individual analysis of each of the categories, the vice domain is dominant in both cases. Both categories express care in the virtue domain, but the scores are relatively low. This suggests that these tweets and conversations tend to promote a negative attitude, regardless of whether the group was in favor or against the theory of arson.

Box plots depicted in Figure 3 present model coefficients (bias and intensity) with 0.95 confidence interval. Negative bias in each of the moral domains. Negative bias values support the outcome of the bar graphs where the vice domain is prominent. The intensity values are comparable for each of the *Supporter* and *Opposer* categories, determining that words from the moral dictionaries are present in both, but there is not enough evidence to differentiate between specific behaviors between the two parties. The lower values of intensity suggest that both the *Supporter* and *Opposer* categories do not use many words present in the eMFD.

Happiness Scores: The *Opposers* were happier (tweeted more positive words) compared to the *Supporter* and *Unaffiliated* groups. The mean scores and standard deviation of the top 10% most frequent words for each group (using *tf-idf*) shows a similar trend: *Supporters*= 5.20 (1.12), *Opposers*= **5.64** (0.77) and *Unaffiliated*= 5.18 (0.91). There was no difference between the means of the *Supporters* and the *Unaffiliated*, hence we fail to reject $H2$. However, for the *Supporters* and *Opposers* as well as for *Opposers* and *Unaffiliated* groups at $\alpha = 0.05$ we reject $H2$, as there is a significant difference in their means. Research has shown that fake news and misinformation contain words that evoke emotions, particularly fear and anger. Such words score low in "happiness". Having a lower proportion of positive (happy) words indicates that *Supporters* used a more negative speech in the dataset. We further cross-computed the happiness scores on the moral words used by each group. In terms of moral foundations, the *Opposers* used the most positive words, followed by the *Unaffiliated* group. A word-shift graph of the results is shown in Figure 1.

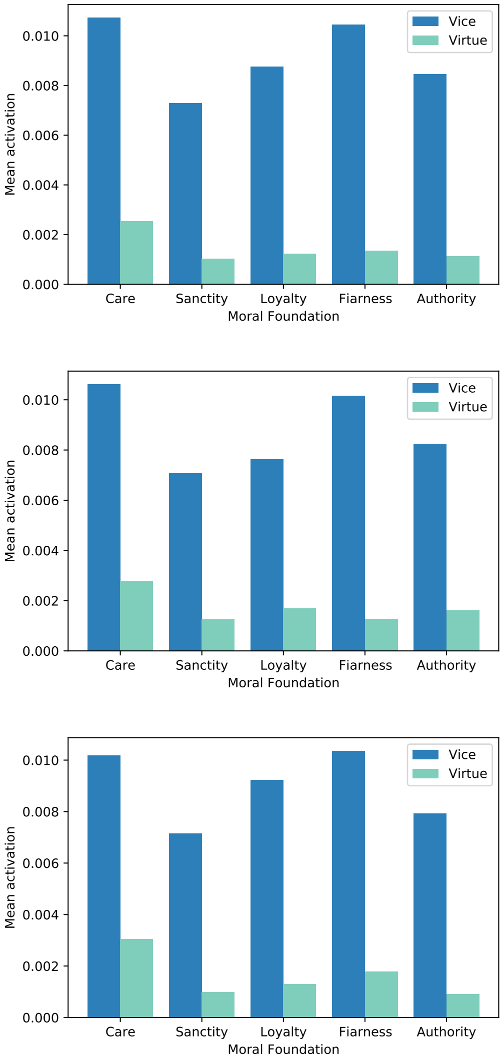

Figure 2: Figures representing the mean activation scores for the dataset. The first figure represents the mean activation scores for the overall document, second for the supporter category and third for the opposer category.

Linguistic Style and Content Analysis: We computed cosine similarity on style and content embeddings for each group. There is

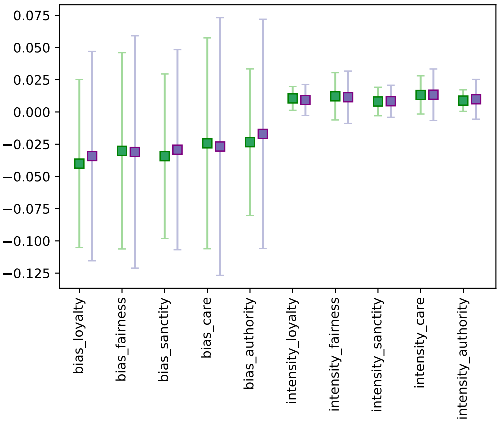

Figure 3: Bias and Intensity calculations for the Weber et al. [23] data set. The green plots determine the bias and intensity for the *Opposer* category whereas the purple plot determines the bias and intensity for the *Supporter* category.

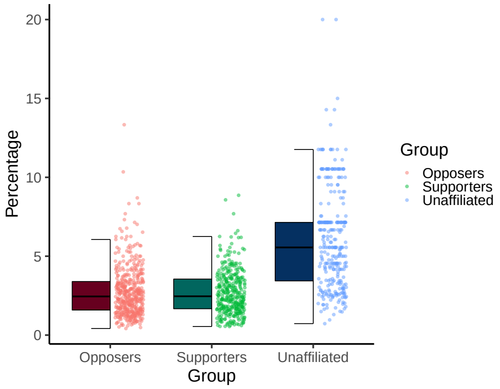

Figure 4: Boxplots depicting the percentage of use of moral words by each user, regardless of the word's moral foundation.

a very high content similarity between each group (almost close to 1). This could be attributed to the fact that the dataset was collected on a very specific topic. We also see a high similarity (0.791) in style between *Supporter* and *Unaffiliated* groups, while the similarity in style for the *Opposer* and the *Unaffiliated* groups was only 0.039, and for *Opposers* and *Supporters* was 0.004, possibly indicating that *Opposers* were not emotionally similar to the other two groups.

Attitude markers express the writer's affect or emotions. They are usually expressed by nouns (e.g. infidel, terrorist, coward, fool, etc.), adjectives (e.g. good, happy, foolish, ludicrous, etc.), verbs (e.g. love, win, plead, massacre, etc.) and adverbs (e.g. unfortunately, absolutely, practically. etc; see Biber et al. 1999; Gales 2010 (Table 2).

We used the profanity dictionary provided by Luis von Ahn's research group[2]. The dictionary contains more than 1300 words that can be considered offensive. However, examining the dictionary revealed that not all words could be considered under the umbrella of profanity. We removed a few such words that were not offensive.

In the bushfires dataset, though we have seen that supporters used the fewest positive words during bushfires, the supporters have still avoided a profuse use of offensive language; some of the supporters have used subtle stance markers, which appeal to the emotions of a reader, while the percentage of profane words in *Opposers*' tweets was higher.

4 DISCUSSION

We also tested the moral valence of tweets in two additional datasets. The FrameAxis method was also used to map the moral valence of tweets that were published after the killing of George Floyd in the USA on May 25^th, 2020 [19]. This incident triggered a wave of protest against injustice and discrimination relating to the race and ethnicity of the people on Twitter where the narrative "#BlackLivesMatter" was trending. Since the tweets demonstrated anger and raised a voice against injustice, the vice domain was prevalent for the Authority, Care, and Fairness category, but the dominance of virtue in the categories of loyalty and sanctity was striking. Priniski et al. [19] interpret this result by noting that the tweets supported *catalyst activism* where people collectively stood against the topic. This result contrasts the results we obtained from our dataset where the vice nature is dominant across all the five moral foundation axes.

We also used the FrameAxis method to compare the moral valence of documents against the THREAT dataset presented in Hammer et al. [12], Wester et al. [25] which is a collection of YouTube comments containing various discussions related to different political, religious, and cultural conflicts to contrast the result obtained in our dataset. The comments were manually annotated into two categories as threat and non-threat. The threat categories contained the use of violent and fatal words and were highly negative in emotions. Even though the non-threat category did not promote violence but was still dominated by negative sentiments and expressions. The moral scores were highly vice-dominant for each of the moral foundation axes. However, the presence of virtuous nature was significant as well and the intensity of the use of moral words was relatively large compared to our dataset. The notable contrast here is that even though the THREAT dataset presents comments around highly sensitive contexts like religion, culture, and race, the presence of virtue words in the comments was significant but in the case of the bushfires dataset, the discussions were led by lower moral words.

5 CONCLUSIONS AND FUTURE WORK

We conducted an analysis of the language used in the Australian Twittersphere on the contentious topic of bushfires. At the aggregate level, both the *Opposers* and *Supporters* of the arson narrative used similar proportions of moral foundation words, a finding which agrees with studies conducted in other countries for political polarisation. Notably, the use of moral words by the *Unaffiliated* group was very high. At a deeper level, we see that the distributions of moral words only align for a few foundations between the groups. *Opposers* used more authority words, perhaps to draw attention to deductive evidence, they also used more positive words. This might indicate that *Opposers* drove assertive polarisation in the pro/anti-science debate, while *Supporters* generated expression-based polarisation

[2] https://www.cs.cmu.edu/ biglou/resources/bad-words.txt

Linguistic Markers	Supporters	Opposers	Unaffiliated
Common Noun	16.27	21.33	18.57
Adjective	4.96	4.15	4.64
Adverb	4.66	3.44	3.81
Interjection	0.48	0.22	0.28
Pre-or Postposition	9.79	10.61	10.57
Coordinating Conjunction	2.47	3.47	3.06
Hashtag	5.72	3.04	4.20
@-mention	4.54	1.21	2.41
URL or Email	1.39	1.02	1.03
First-person Pronoun	1.38	1.29	1.12
Second Person Pronoun	0.88	0.45	0.61
Third Person Pronoun	2.76	1.90	2.36
Profane Words	1.43	2.04	1.74

Table 2: Percentage of various linguistic markers used by supporter, opposer and unaffiliated groups.

as *Supporters* used more care and loyalty words, possibly to invite sympathy for their viewpoint. *Supporters* also used the least positive words. Additionally, we did not find any similarity in style between *Opposers* and *Supporters*. However, we found stylistic similarity between *Supporters* and *Unaffiliated*, while *Opposers* were not similar to any of those. In the future, we want to further analyse whether such polarised groups persist through time and whether the linguistic characteristics hold for their extended Twitter timeline.

6 ACKNOWLEDGMENTS

This research was partially supported by Australian Research Council's (ARC) Centre of Excellence for Mathematical and Statistical Frontiers, Australia under project ID CE140100049, awarded to MN.

MN and PB also acknowledge financial support from the SA Defence Innovation Partnership CRF Round05 grant.

	Supporters	Opposers	Unaffiliated
Authority.vice	0.591	1.473	2.460
Authority.virtue	1.340	1.497	5.574
Care.vice	0.898	1.304	5.095
Care.virtue	0.679	0.975	4.160
Fairness.vice	0.622	1.088	3.852
Fairness.virtue	0.546	0.893	3.672
Loyalty.vice	0.603	0.963	3.886
Loyalty.virtue	0.728	1.234	2.734
Sanctity.vice	0.872	1.211	2.944
Sanctity.virtue	0.572	1.190	2.993

Table 3: Mean Percentage of various categories of moral words used by the three groups. Please note that for this table we have calculated the mean keeping the profiles that have used at least one word from the respective foundation/aspect. Hence values of means may seem a little higher that the results reported in the main paper.

Frequency of Moral Words					
Supporters	Counts	**Opposers**	Counts	**Unaffiliated**	Counts
police	140	traitor	410	traitor	92
attack	89	father	198	father	60
help	55	arrested	89	police	29
control	49	liars	70	arrested	22
arrested	48	group	44	country	13
queen	48	benefit	42	betrayal	13
spreading	48	spreading	41	murder	13
rebels	46	followers	40	liars	12
arrests	42	police	40	spreading	11
country	37	cult	38	share	10
murder	36	loving	37	attack	9
protect	36	war	22	scam	8
leaders	35	lord	16	control	7
rebellion	30	country	12	fair	6
protecting	28	fucking	10	pure	5
trash	28	god	9	help	5
hell	27	exploited	8	benefit	5

Table 4: Most frequently used moral words by each group.

REFERENCES

[1] Michael D. Conover, Jacob Ratkiewicz, Matthew R. Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *ICWSM*. The AAAI Press.

[2] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2970–3005.

[3] Peter DeScioli and Robert Kurzban. 2013. A solution to the mysteries of morality. *Psychological bulletin* 139, 2 (2013), 477.

[4] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences* 112, 8 (2015), 2389–2394.

[5] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one* 6, 12 (2011), e26752.

[6] Jeremy A Frimer. 2020. Do liberals and conservatives use different moral languages? Two replications and six extensions of Graham, Haidt, and Nosek's (2009) moral text analysis. *Journal of Research in Personality* 84 (2020), 103906.

[7] Jeremy A Frimer, Caitlin E Tell, and Matt Motyl. 2017. Sacralizing liberals and fair-minded conservatives: Ideological symmetry in the moral motives in the culture war. *Analyses of Social Issues and Public Policy* 17, 1 (2017), 33–59.

[8] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*. 913–922.

[9] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Polarization on Social Media. In *Tutorial of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. https://www2018.thewebconf.org/program/tutorials-track/tutorial-202/

[10] Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica* 87, 4 (2019), 1307–1340. https://doi.org/10.3982/ecta16566

[11] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130.

[12] Hugo L Hammer, Michael A Riegler, Lilja Øvrelid, and Erik Velldal. 2019. Threat: A large annotated corpus for detection of violent threats. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–5.

[13] Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods* 53, 1 (2021), 232–246.

[14] Neta Kligler-Vilenchik, Christian Baden, and Moran Yarchi. 2020. Interpretative polarization across platforms: How political disagreement develops over time on Facebook, Twitter, and WhatsApp. *Social Media+ Society* 6, 3 (2020), 2056305120944393.

[15] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science* 7 (2021), e644.

[16] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. 591–600.

[17] Ping Li, Benjamin Schloss, and D. Jake Follmer. 2017. Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior Research Methods* 49, 5 (jul 2017), 1668–1685. https://doi.org/10.3758/s13428-017-0931-5

[18] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*. 380–390.

[19] J Hunter Priniski, Negar Mokhberian, Bahareh Harandizadeh, Fred Morstatter, Kristina Lerman, Hongjing Lu, and P Jeffrey Brantingham. 2021. Mapping moral valence of tweets following the killing of George Floyd. *arXiv preprint arXiv:2104.09578* (2021).

[20] Karolina Sylwester and Matthew Purver. 2015. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLOS ONE* 10, 9 (sep 2015), e0137422. https://doi.org/10.1371/journal.pone.0137422

[21] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[22] Sze-Yuh Nina Wang and Yoel Inbar. 2020. Moral-Language Use by US Political Elites. *Psychological Science* (2020), 0956797620960397. https://doi.org/10.1177/0956797620960397

[23] Derek Weber, Mehwish Nasim, Lucia Falzon, and Lewis Mitchell. 2020. # Arson-Emergency and Australia's "Black Summer": Polarisation and Misinformation on Social Media. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer, 159–173.

[24] Derek Weber, Mehwish Nasim, Lucia Falzon, and Lewis Mitchell. 2020. #ArsonEmergency and Australia's "Black Summer": A study of polarisation and its broader effect on the online discussion. Talk presented at the fifth Australian Social Network Analysis Conference, ASNAC'20, 25–27 November, Adelaide, Australia.

[25] Aksel Wester, Lilja Øvrelid, Erik Velldal, and Hugo Lewi Hammer. 2016. Threat detection in online discussions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 66–71.

[26] S.C. Woolley and D.R. Guilbeault. 2018. United States: Manufacturing Consensus Online. In *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*, P.N. Howard and S.C. Woolley (Eds.). Oxford University Press, Chapter 8, 185–211. https://doi.org/10.1093/oso/9780190931407.001.0001