Supplementary Information for:

Reconsidering Evidence of Moral Contagion in Online Social Networks

Jason W. Burton[1]*, Nicole Cruz[2], and Ulrike Hahn[1]

[1] Department of Psychological Sciences, Birkbeck, University of London, London, WC1E 7HX, UK
[2] School of Psychology, University of New South Wales, Sydney, 2052, AU
* Corresponding author: Jason W. Burton (jburto03@mail.bbk.ac.uk; https://orcid.org/0000-0002-6797-2299)

Contents:

1. Supplementary Methods
2. Supplementary Results
3. Supplementary Tables 1-3
4. Supplementary Figures 1-9
5. Supplementary References

1. Supplementary Methods

1.1 Datasets

A total of six corpora were analyzed in this study, plus the re-analysis of Brady et al.'s (2017) data. Brady et al.'s[1] data was retrieved from their public project page on the Open Science Framework (OSF), and four pre-existing corpora were obtained via Google's dataset search engine, which hosts links to a wide range of open data repositories. The COVID-19 and #MuellerReport corpora were collected by the authors by connecting to the Twitter REST API with the `rtweet` package[2]. While no specific corpus or topic was initially targeted, certain criteria were employed. To be considered for this study, corpora had to contain Twitter data (i.e., tweet messages and retweet counts), contain messages written in English, and relate to a polarizing or morally-charged real-world issue, event, or social movement. Once retrieved, corpora were further narrowed by collapsing repeated messages into a single observation (to generate a composite diffusion count that combined the raw retweet count with the number of times the message appeared in the corpus) and removing non-English messages. Since the pre-existing corpora did not include language identifying metadata, the `textcat` package was employed to extract English tweets in these instances. We report descriptive statistics for each corpus in Supplementary Table 1. All analyses were done in R and scripts are available on OSF [https://osf.io/4zjk6/?view_only=a9f977a26f754b2da4c59b2fba29cd50].

Brady et al. (2017). First and foremost, the present study drew directly from the recent study of moral contagion in online social networks by Brady et al. (2017). Their data, which they shared on a public OSF project page, was thus crucial to the present study for both inspiration and corroboration. The data collected by Brady et al. (2017) focused on topical political issues in the United States: gun control (n = 48,394), same-sex marriage (n = 29,060), and climate change (235,548). Using the Twitter API and sets of topic-related filter words (e.g., guns, gun control, and NRA for the gun control topic), tweets and metadata were extracted between 30 October and 15 December 2015.

COVID-19. For this corpus we collected tweets pertaining to the (ongoing at the time of writing) COVID-19 pandemic. Using the `rtweet` package, we specified a search for English tweets including at least one of the following terms: *#COVID-19*, *COVID-19, COVID19, covid19, COVID, covid,* or *coronavirus*. Collected tweets were posted on 23-24 March 2020, a period in which nation-wide lockdowns were being put into effect across the globe. While the topic of infectious disease does not necessarily evoke feelings of morality or polarization a priori, the COVID-19 pandemic has elicited highly contentious debate in political, scientific, and public spheres. For example, Reuters reported results of a poll showing that Democrats are about twice as likely as Republicans to say COVID-19 poses an imminent threat to the US[3], and researchers identified political polarization as an important part of the social context that should be addressed in responses to COVID-19[4].

#MeToo. Our second corpus comprised of Twitter messages containing the *#metoo* hashtag was obtained from the data.world repository. The tweets were collected from the Twitter API between 29 November and 25 December 2017, little more than a month after the #metoo hashtag first appeared online in coordination with the "Me Too movement"[5]. The "Me Too movement" is a movement against sexual harassment and assault. It was ignited by Hollywood sexual abuse allegations and has since become an international phenomenon garnering widespread media attention, support, and critique.

#MuellerReport. A third corpus was collected by using the *#muellerreport* hashtag to retrieve tweets from the Twitter API created between 23 and 25 March 2019 — the weekend during which US Attorney General William Barr released his summary of Special Counsel Robert Mueller's investigation into Donald Trump's 2016 presidential campaign. This corpus was of special interest because the Mueller Report has been a major source of controversy. While originally a non-polarized issue, the public opinion divided over time[6] meaning that moral-emotion could have plausibly played a part in moralizing conversations on Twitter.

2016 US Election. Our fourth corpus containing viral tweets (those with 1,000+ retweets) from the 2016 US Presidential Election was obtained from the Zenodo repository. The set of

tweets was collected with the Twitter API and extracted messages that contained specific hashtags (*#MyVote2016*, *#ElectionDay*, and *#electionnight*) and/or user handles (*@realDonaldTrump* and *@HillaryClinton*)[7]. This corpus was of special interest as it contained many "fake news" messages as coded by the curators, which one might expect to use especially morally- and emotionally-charged language to garner extra attention given the conclusions is Brady et al.'s original study[1].

Post-Brexit. A fourth corpus containing unfiltered tweets and metadata from the morning that Brexit was announced was obtained from the Mendeley Data repository. These tweets were collected with NCapture from QSR and employed a tight temporal parameter so as to capture the public's reaction to the political event[8]. Brexit refers to the result of the 2016 EU Referendum in the United Kingdom, and this dataset includes Twitter responses from across the globe.

#WomensMarch. Our sixth and final corpus with tweets containing the *#womensmarch* hashtag was obtained from the data.world repository. Using the Twitter API, 15,000 messages were collected that referenced the pro-women's rights, and effectively anti-Trump, protest that took place in the wake of the presidential inauguration on 21 January 2017[9]. The Women's March has since become a worldwide movement with annual marches in late January to non-violently protest for women's reproductive rights, LGBTQ rights, immigration and healthcare reform, as well as racial, gender, and religious equality.

1.2 Measures

Measuring language. The same dictionaries used in Brady et al. (2017) were used to quantify distinctly emotional (n = 819; e.g., panic, fear, heartwarming), distinctly moral (n = 316; e.g., fair, racism, solidarity), and moral-emotional language (n = 72; e.g., shame, victimize, disgust). Importantly, there is no overlap in these dictionaries, meaning that each tweet could be allocated the discrete scores that formulate three independent predictor variables. To ensure our scripts were accurately counting words and word stems, we performed a check in which we re-ran the scripts with a random sample of 10 word stems and 10 words and manually checked that the correct counts were displayed on a random sample of 20 tweets from each corpus that had at least one word/stem counted. By selecting a manageable number of tweets, words, and word stems, we were able to check for both false positives and false negatives and then simply scale up our scripts. We found that our scripts were accurately counting words and word stems, and the tweets included in each corpus were relevant to their respective topics.

Measuring message diffusion. We quantified message diffusion by adding the retweet count captured in messages' metadata to the number of times that messages' exact text appeared in the dataset. By then grouping together tweets that had identical message text and collapsing into a single observation (with other relevant data taken from the earliest posting of a message), we were able to avoid penalizing retweet chains, which is an important indicator of message diffusion on Twitter, while also accounting for unconventional retweets where a user copies and pastes someone's message rather than clicking the retweet button.

1.3 Packages, preprocessing, and models

Data Preprocessing. Datasets were preprocessed with the `tm` and `dplyr` packages in R prior to applying the dictionary-based text analysis. This included converting all text to ASCII characters and removing retweet prefixes (i.e. "RT"), usernames, punctuation, and URLs. Observations in which no text remained after the preprocessing were removed from the analysis. Preprocessing script is available on OSF.

Models. Using the `MASS` package, we used the `glm.nb` function to fit various specifications of negative binomial regression models. While the specification curve analysis involved fitting many model specifications not listed here, the three most central models considered in our analysis are:

1) the main multivariate moral contagion model in which diffusion is predicted by the three independent language measures based on the distinctly moral, distinctly emotional, and moral-emotional dictionaries;
2) the univariate moral contagion model in which diffusion is predicted by only moral-emotional language; and
3) the XYZ contagion model in which diffusion is predicted by a single "XYZ count" variable that is simply the number of X's, Y's, and/or Z's present in the message text after preprocessing.

2. Supplementary Results

2.1 Evaluating Brady et al.'s dictionaries as predictors of human judgements of moral expression in the Moral Foundations Twitter Corpus (MFTC)[10]

One possible explanation for the inconsistent moral contagion effects observed in the present work is measurement error. That is, the dictionaries used by Brady et al. (2017) might not be accurately measuring expressions of moral sentiment in tweets. Identifying moral sentiments in text is difficult because different types of moral sentiment can co-occur, they might only be implicitly signaled, and because the ground truth is inherently subjective[10]. In order to investigate how well Brady et al.'s (2017) dictionaries identify expressions of moral sentiment we conducted a supplementary analysis with the MFTC, which contains 35,108 tweets from seven topics of discourse ["All lives matter" (ALM), Baltimore protests, "Black lives matter" (BLM), hate speech messages from Davidson et al. 2017[11], 2016 US Presidential Election, #MeToo, and Hurricane Sandy] that have been manually annotated by three to five human annotators for moral sentiment[10]. Note that the #MeToo and the 2016 US Presidential Election corpora included in the MFTC and those addressed in the main analyses of the present work are different, despite sharing discourse topics.

Since the present work is not concerned with individual categories of moral sentiment (e.g., purity, loyalty, authority, etc.), we collapsed the category labels such that we compared the total number of moral labels to the number of non-moral labels assigned by the annotators, in turn producing a binary classification of each tweet as moral or non-moral. We then applied four logistic regression classifiers with the moral-emotional, distinctly moral, and distinctly emotional dictionaries as predictors (one multiple logistic regression with all predictors included, and the three nested, single-variable logistic regressions) to see if the dictionaries' predicted classifications aligned with human judgements of moral expression.

Supplementary Figure 1 displays ROC curves and calculated AUC values for each logistic regression classifier as applied to each corpus included in the MFTC. Across the seven corpora the mean AUC for the multiple logistic regression classifier ranged from 51.7% in the Davidson hate speech corpus to 83.2% in the #MeToo corpus (M_{AUC} = 72.2%). In line with the analysis reported by Hoover et al.[10], we found classification performance to vary significantly by context. In addition, we calculated the logistic regression classifiers' precision, recall, and F1 metrics. Due to class imbalances in the data (Supplementary Table 2), we used repeated under-sampling whereby we randomly excluded observations from the majority class in each corpus and re-fit the classifiers and then averaged the calculations across 100 iterations (Supplementary Table 3). We found the logistic regression classifiers to have poor recall (M_{Recall} = 53.9%). This suggests that the dictionary-based approach does not effectively identify all tweets in which human annotators find moral sentiment expressed, which raises an additional methodological concern about the specific measurements made in Brady et al. (2017).

2.2 Specification curve analyses of Brady et al.'s data

In the main text, we report the results of specification curve analyses (SCA) of the COVID-19, #MeToo, and #MuellerReport corpora. To supplement these analyses, we applied SCA to Brady et al.'s[1] data (Supplementary Figures 2-5). Across model specifications that considered their chosen covariates and three arbitrary (but defensible) increments of outliers (the top 10, 100, or 1,000 most retweeted messages), we find the moral contagion effect to be particularly robust in the climate change corpus (median B = 0.14, SD = 0.06), and positive but variable in

the gun control corpus (median B = 0.08, SD = 0.10). However, the moral contagion effect appears notably unstable in the same-sex marriage corpus with a negative median regression coefficient (median B = -0.04, 0.09). Supplementary Figure 5 further shows how supposed "outliers" can influence results, which is expected to an extent given the fat-tailed distribution of retweet data.

2.3 **Bootstrap resampling**

Bootstrap resampling was also conducted as a robustness check to keep with the procedures of Brady et al.[1]. This technique involves regenerating variations of a dataset by sampling with replacement, meaning that certain datapoints may be duplicated and others may be omitted. By iteratively repeating this procedure and re-fitting each model in question (500 iterations in this case), a distribution of effect sizes is produced along with a 95% confidence interval, which is considered indicative of the reliability of an effect within a sample. Specifically, an observed effect may be deemed stable if the confidence interval does not straddle zero.

It should be noted, however, that this procedure will only ever speak to the robustness of an effect within a sample, when the critical issue of interest is whether what has been found in a sample is indicative of the population at large (e.g., is the observed moral contagion effect generalizable to political tweets or political communications?). While it is true that an effect that is not even stable within a sample provides poorer evidence vis-à-vis the wider population than one that is, the fact that an effect is stable within a sample is insufficient to determine whether it extends beyond that sample. Moreover, the concerns that correlational analyses of big data raise for spurious factors are evidently not assuaged by bootstrap resampling: the XYZ contagion passes this robustness check in the three largest datasets analyzed (Supplementary Figure 6). Only out-of-sample prediction can address this issue, as conducted in the present study.

3. Supplementary Tables

Supplementary Table 1: Descriptive statistics of each analyzed corpus. *N* refers to the total number of raw tweets included in the corpus (including duplicates, non-English tweets, and tweets with no text), and *n* refers to the number of clean, unique tweets analyzed in the paper.

	COVID-19	**#MeToo**	**#Mueller Report**	**2016 US Election**	**Post-Brexit**	**#Womens March**
N	701,925	393,135	229,046	9,001	17,998	15,000
n	172,697	151,035	39,068	8,233	5,660	3,778
Diffusion minimum	0	0	0	1,001	0	0
Diffusion maximum	368,611	56,750	25,842	100,000	31,901	170,518
Diffusion *M* (*SD*)	266.78 (4,152.64)	8.93 (222.63)	15.54 (312.87)	3,372.65 (5,222.76)	119.60 (923.57)	705.80 (4,983.11)
ME words *M* (*SD*)	0.23 (0.52)	0.20 (0.47)	0.18 (0.47)	0.16 (0.43)	0.09 (0.32)	0.16 (0.42)
Moral words *M* (*SD*)	0.49 (0.79)	0.26 (0.53)	0.47 (0.77)	0.33 (0.61)	0.22 (0.48)	0.28 (0.57)
Emotional words *M* (*SD*)	1.09 (1.24)	0.89 (0.97)	0.99 (1.17)	0.85 (1.02)	0.69 (0.90)	0.67 (0.83)
XYZ count *M* (*SD*)	2.61 (2.18)	1.84 (1.44)	2.43 (2.08)	1.68 (1.41)	2.28 (1.37)	1.53 (1.32)

Supplementary Table 2: Frequencies of moral and non-moral expression in the manually annotated twitter corpora comprising the Moral Foundations Twitter Corpus (MFTC).

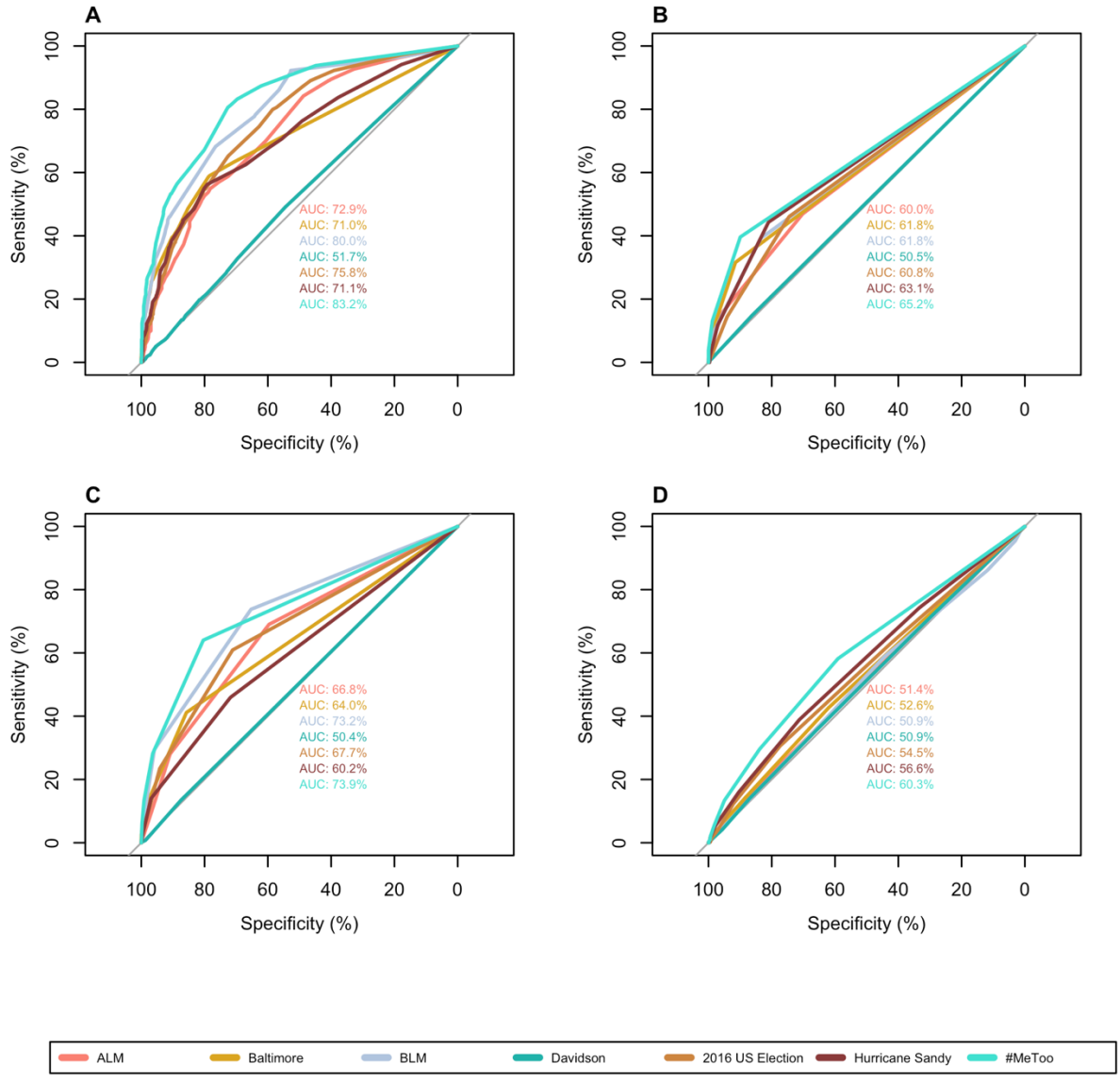
Corpus	**Non-Moral**	**Moral**	*Total*
ALM	726	3,698	4,424
Baltimore	2,869	2,724	5,593
BLM	1,133	4,124	5,257
Davidson	3,825	1,048	4,873
2016 US Election	1,877	3,481	5,358
#MeToo	914	3,977	4,891
Hurricane Sandy	585	4,006	4,591
Total	11,929	23,058	34,987

Supplementary Table 3: Performance metrics of dictionary-based logistic regression classifiers. Values indicate the mean of a given metric following 100 iterations of under-sampling, with standard deviations in parentheses. Classification threshold set to 0.5.

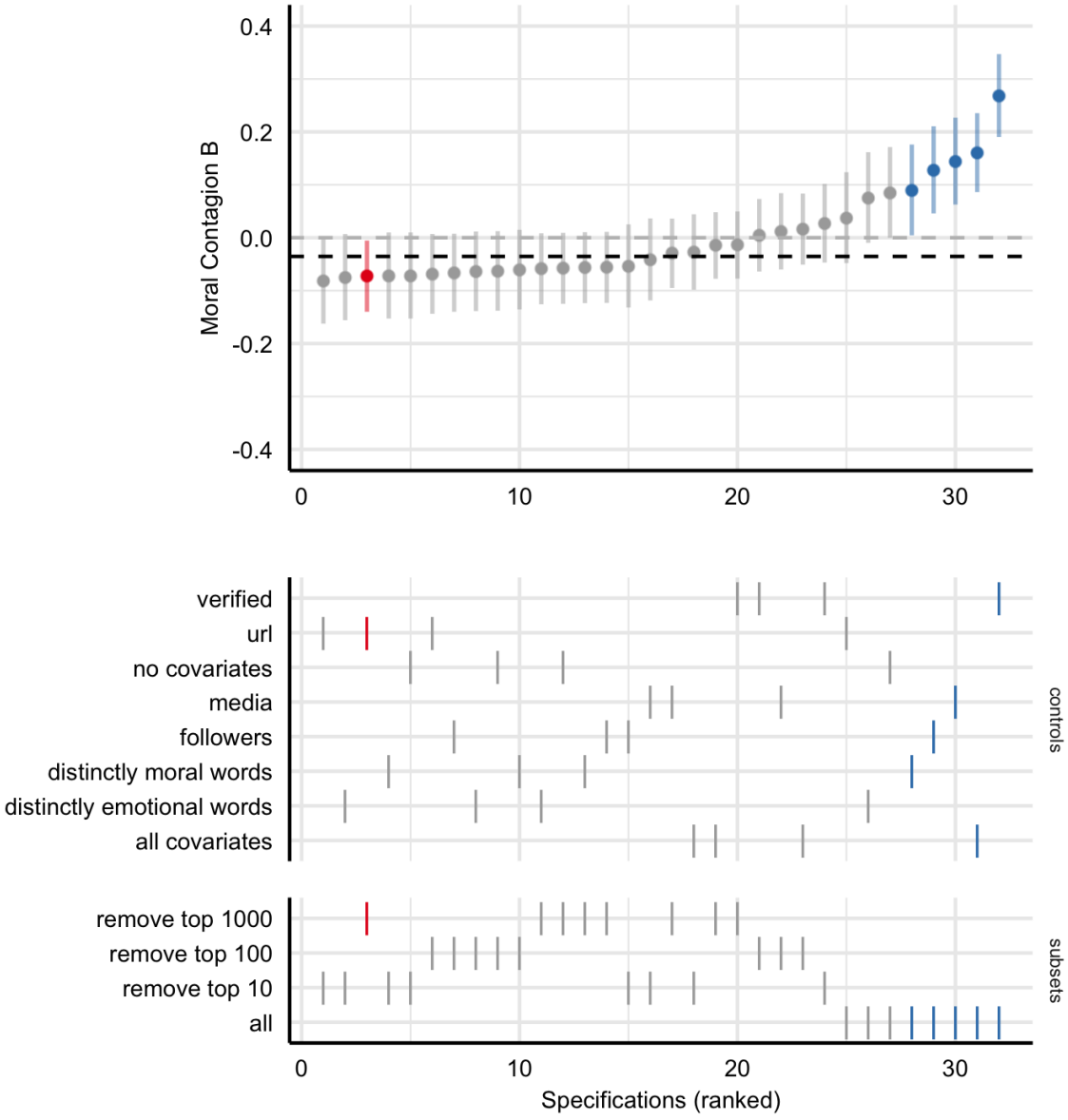
Metric	ALM	Baltimore	BLM	Davidson	2016 US Election	Hurricane Sandy	#MeToo
moral_expression ~ ME_words + moral_words + emotional_words							
AUC	0.729 (0.007)	0.710 (0.001)	0.801 (0.006)	0.519 (0.008)	0.758 (0.004)	0.712 (0.010)	0.832 (0.006)
F1	0.599 (0.009)	0.618 (0.001)	0.617 (0.006)	0.506 (0.004)	0.608 (0.004)	0.631 (0.011)	0.652 (0.007)
Precision	0.669 (0.008)	0.640 (0.001)	0.732 (0.007)	0.505 (0.004)	0.677 (0.004)	0.601 (0.010)	0.739 (0.007)
Recall	0.514 (0.008)	0.527 (0.001)	0.508 (0.011)	0.511 (0.006)	0.545 (0.005)	0.566 (0.011)	0.604 (0.007)
moral_expression ~ ME_words							
AUC	0.613 (0.013)	0.652 (0.000)	0.715 (0.017)	0.425 (0.080)	0.671 (0.008)	0.638 (0.014)	0.727 (0.023)
F1	0.529 (0.014)	0.451 (0.000)	0.501 (0.011)	0.295 (0.148)	0.538 (0.007)	0.543 (0.019)	0.530 (0.014)
Precision	0.660 (0.010)	0.530 (0.000)	0.708 (0.007)	0.425 (0.080)	0.642 (0.005)	0.527 (0.016)	0.696 (0.009)
Recall	0.521 (0.012)	0.473 (0.001)	0.513 (0.052)	0.436 (0.051)	0.399 (0.008)	0.464 (0.018)	0.586 (0.009)
moral_expression ~ moral_words							
AUC	0.722 (0.009)	0.734 (0.002)	0.737 (0.016)	0.522 (0.012)	0.708 (0.004)	0.694 (0.039)	0.770 (0.014)
F1	0.611 (0.008)	0.787 (0.002)	0.702 (0.006)	0.520 (0.017)	0.644 (0.004)	0.700 (0.010)	0.798 (0.006)
Precision	0.631 (0.005)	0.743 (0.002)	0.680 (0.003)	0.522 (0.012)	0.678 (0.002)	0.619 (0.010)	0.765 (0.003)
Recall	0.508 (0.007)	0.530 (0.001)	0.506 (0.017)	0.511 (0.008)	0.574 (0.006)	0.575 (0.012)	0.588 (0.005)
moral_expression ~ emotional_words							
AUC	0.533 (0.020)	0.587 (0.000)	0.697 (0.041)	0.373 (0.113)	0.638 (0.015)	0.593 (0.036)	0.691 (0.052)
F1	0.467 (0.016)	0.316 (0.000)	0.390 (0.011)	0.261 (0.259)	0.462 (0.008)	0.443 (0.022)	0.397 (0.014)
Precision	0.691 (0.015)	0.412 (0.000)	0.738 (0.012)	0.373 (0.113)	0.609 (0.007)	0.459 (0.019)	0.639 (0.012)
Recall	0.535 (0.017)	0.427 (0.000)	0.525 (0.089)	0.390 (0.101)	0.306 (0.008)	0.389 (0.020)	0.583 (0.013)

4. Supplementary Figures

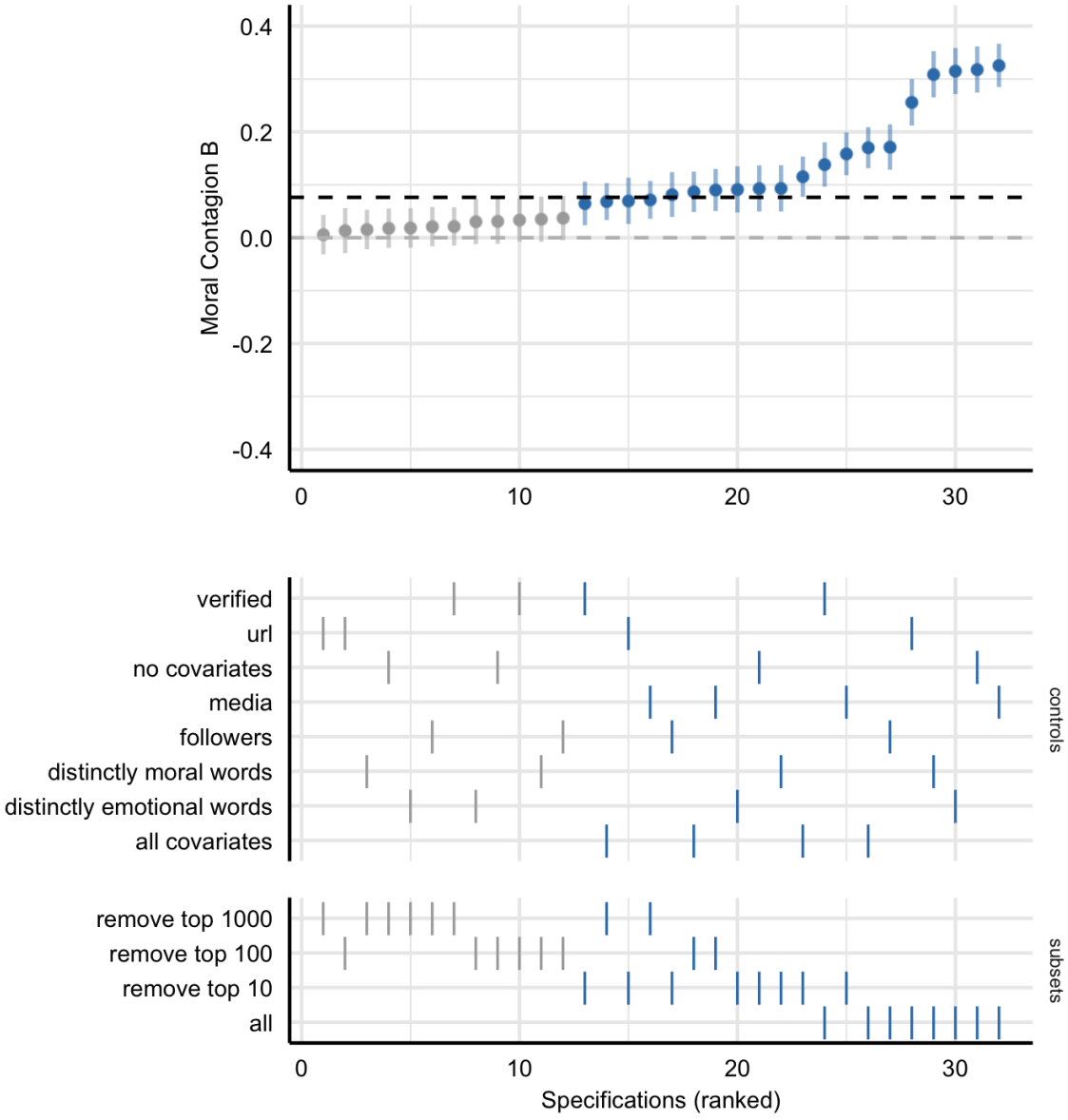
Supplementary Figure 1: ROC/AUC plots of dictionary logistic regression classifiers of moral expression when applied to the complete MFTC corpora. [A] Logistic regression classifier with all three dictionaries — moral-emotional, distinctly moral, and distinctly emotional — used as predictors of moral expression in tweets. [B] Logistic regression classifier with only the moral-emotional dictionary as a predictor. [C] Logistic regression classifier with only the distinctly moral dictionary as a predictor. [D] Logistic regression classifier with only the distinctly emotional dictionary as a predictor.



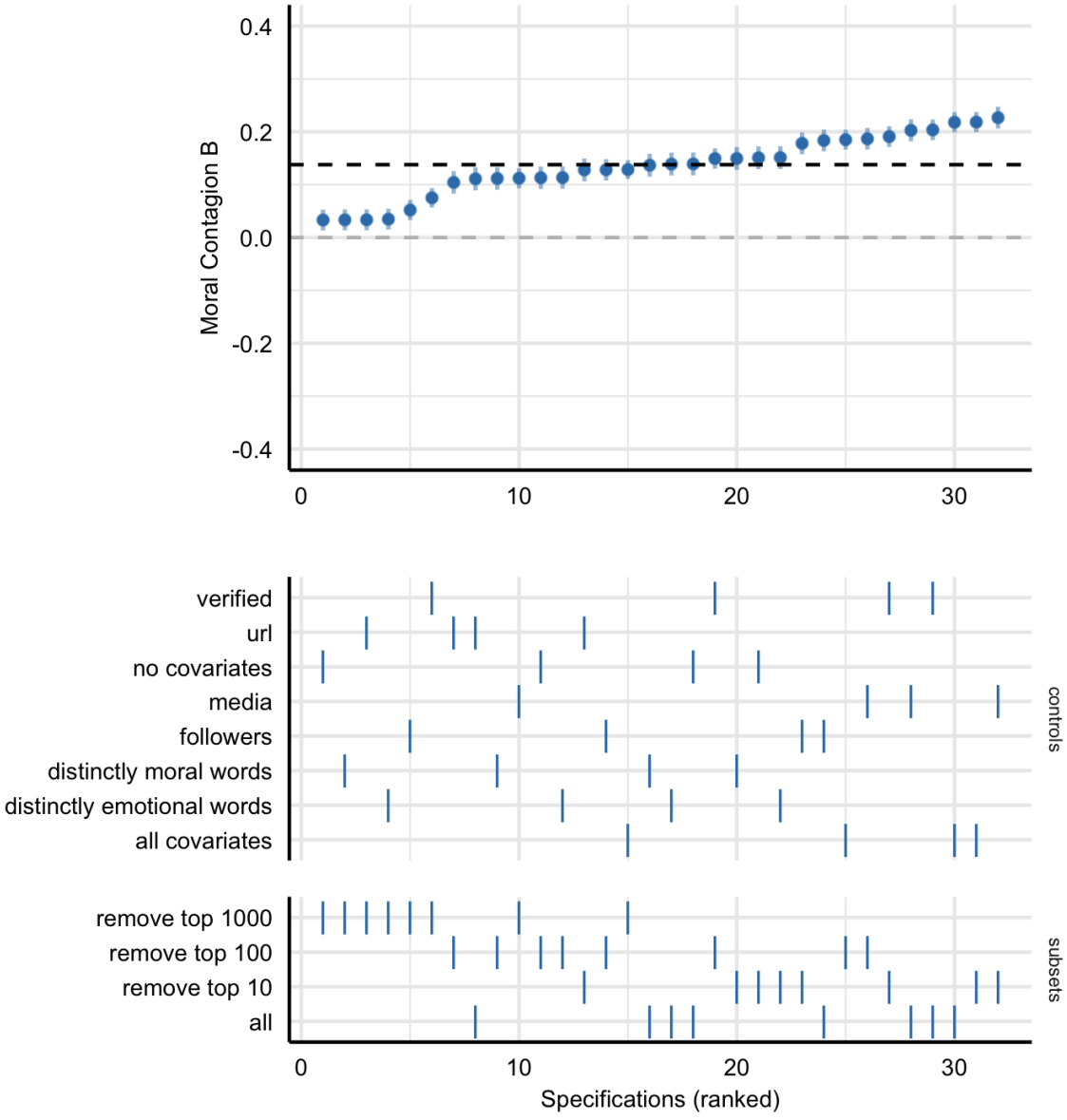
Supplementary Figure 2: Specification curve analysis of Brady et al.'s (2017) same-sex marriage corpus (median B = -0.04, 0.09). Top plot displays the unstandardized regression coefficients (y-axis) for each model specification (x-axis). Red indicates a significant (p < 0.05) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. Bottom plot displays each model specification (x-axis) as three vertically-aligned points corresponding to the outliers removed and covariates included in the negative binomial regression model (y-axis). Red indicates a significant (p < 0.05) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient.



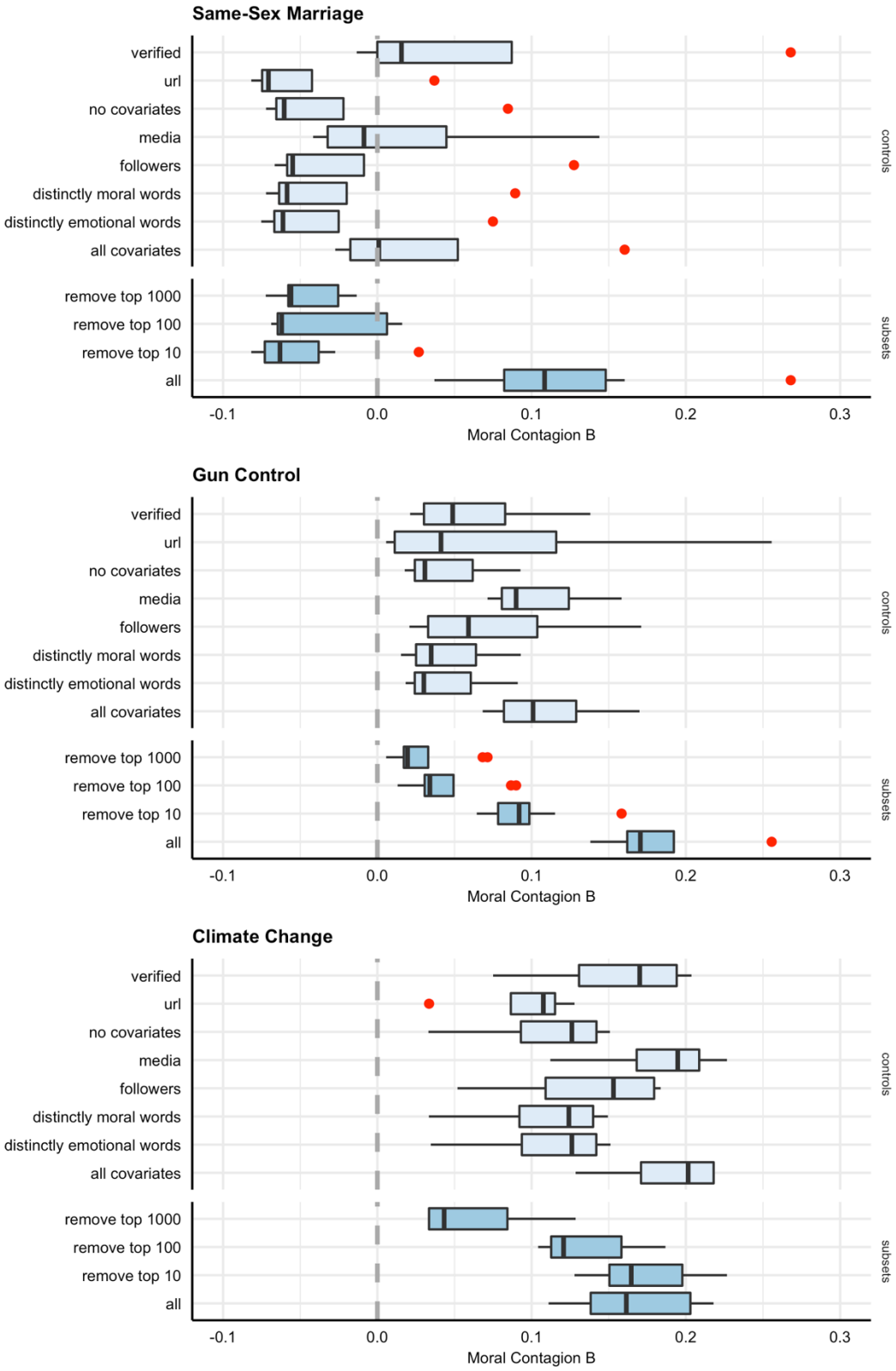
Supplementary Figure 3: Specification curve analysis of Brady et al.'s (2017) gun control corpus (median B = 0.08, SD = 0.10). Top plot displays the unstandardized regression coefficients (y-axis) for each model specification (x-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. Bottom plot displays each model specification (x-axis) as three vertically-aligned points corresponding to the outliers removed and covariates included in the negative binomial regression model (y-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient.



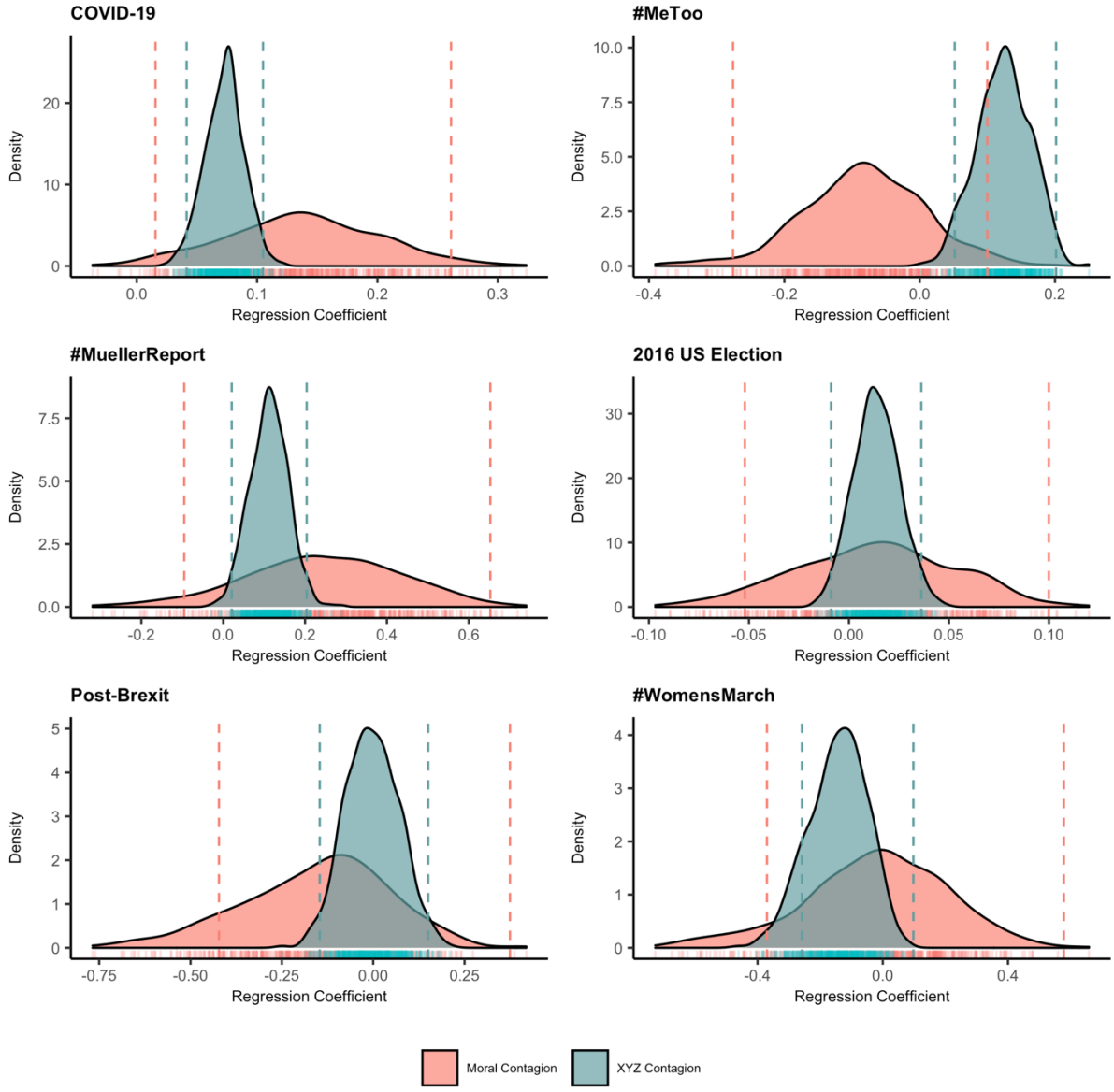
Supplementary Figure 4: Specification curve analysis of Brady et al.'s (2017) climate change corpus (median B = 0.14, SD = 0.06). Top plot displays the unstandardized regression coefficients (y-axis) for each model specification (x-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient. Bottom plot displays each model specification (x-axis) as three vertically-aligned points corresponding to the outliers removed and covariates included in the negative binomial regression model (y-axis). Red indicates a significant ($p < 0.05$) negative regression coefficient, grey indicates a non-significant coefficient, and blue indicates a significant positive coefficient.



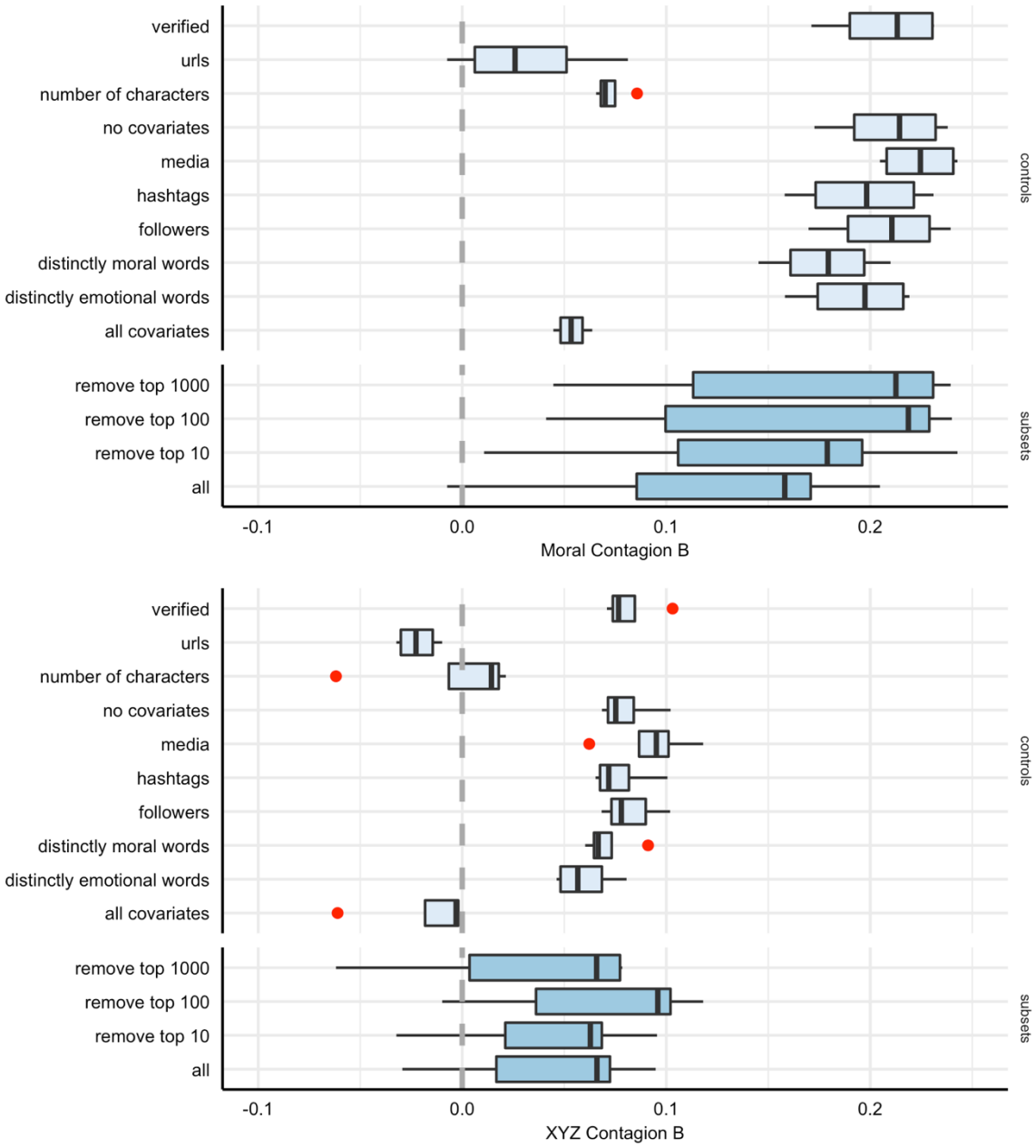
Supplementary Figure 5: Summary plot of specification curve re-analysis of Brady et al. (2017). Boxplots show the distribution of unstandardized negative binomial regression coefficients produced by model specifications accounting various covariates and outliers (y-axis).



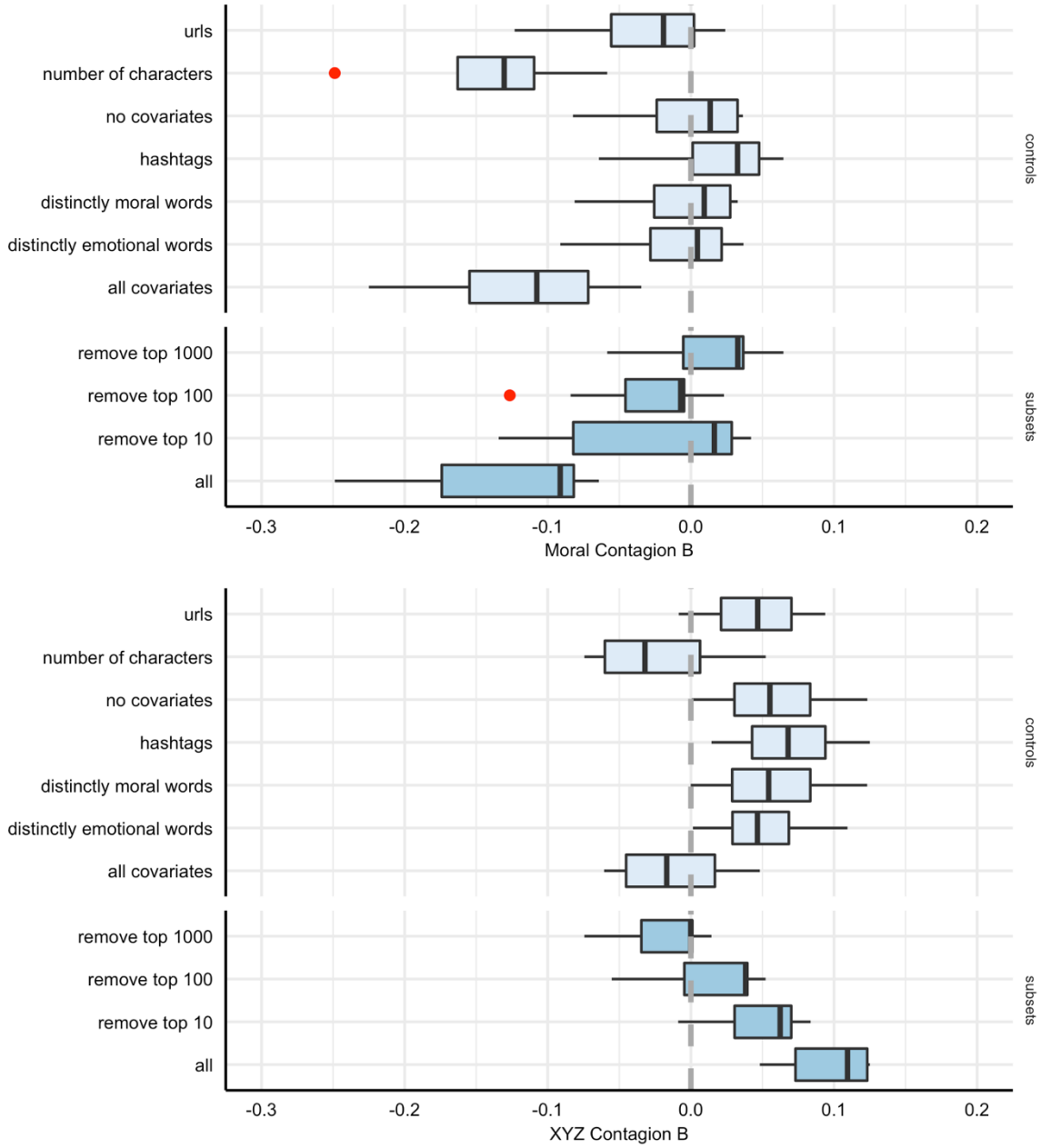
Supplementary Figure 6: Density plots of bootstrap resampling results in each corpus. Each plot displays 500 iterations (per model) of resampling. Dotted lines indicate the 95% confidence intervals for the respective effects.



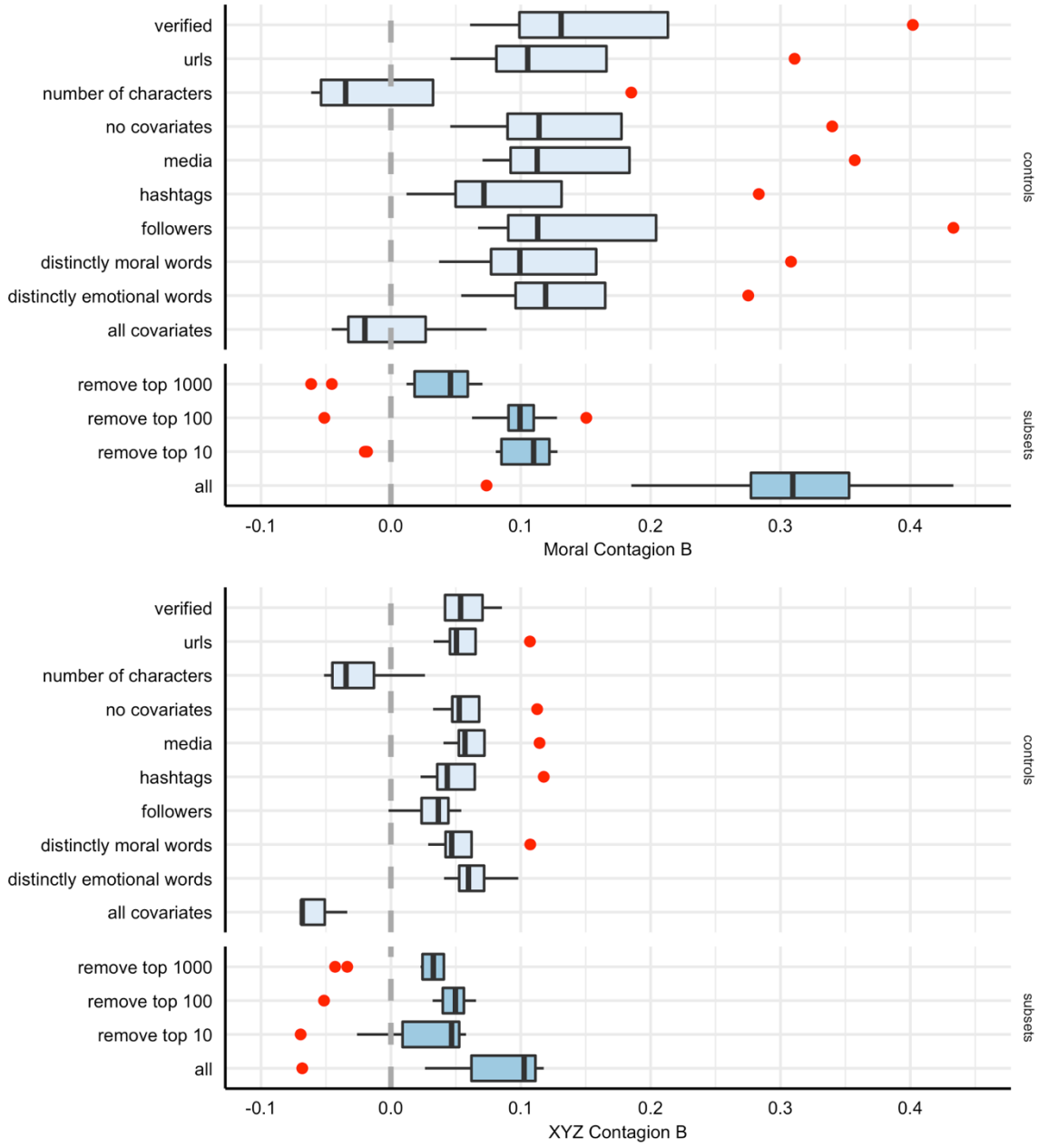
Supplementary Figure 7: Specification curve analysis summary plot of the COVID-19 corpus. Boxplots show the distribution of unstandardized negative binomial regression coefficients produced by model specifications accounting for various covariates and outliers (y-axis). The top plot displays the results of the moral contagion model and the bottom plot displays the results of the XYZ contagion model.



Supplementary Figure 8: Specification curve analysis summary plot of the #MeToo corpus. Boxplots show the distribution of unstandardized negative binomial regression coefficients produced by model specifications accounting for various covariates and outliers (y-axis). The top plot displays the results of the moral contagion model and the bottom plot displays the results of the XYZ contagion model.



Supplementary Figure 9: Specification curve analysis summary plot of the #MuellerReport corpus. Boxplots show the distribution of unstandardized negative binomial regression coefficients produced by model specifications accounting for various covariates and outliers (y-axis). The top plot displays the results of the moral contagion model and the bottom plot displays the results of the XYZ contagion model.



5. Supplementary References

1. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci.* **114**, 7313–7318 (2017).
2. Kearney, M. W. rtweet: Collecting and analyzing Twitter data. *J. Open Source Softw.* **4**, 1829 (2019).
3. Heath, B. Americans divided on party lines over risk from coronavirus: Reuters/Ipsos poll. *Reuters.* https://www.reuters.com/article/us-health-coronavirus-usa-polarization/americans-divided-on-party-lines-over-risk-from-coronavirus-reuters-ipsos-poll-idUSKBN20T2O3 (2020).
4. Van Bavel, J. J. *et al.* Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* (2020).
5. Turner, A. 390,000 #MeToo Tweets. *data.world* (2018).
6. Thomson-DeVeaux, A. The Politics Surrounding Mueller Have Changed A Lot SInce He Started. *FiveThirtyEight* (2019).
7. Amador, J., Oehmichen, A. & Molina-Solana, M. Fakenews on 2016 US elections viral tweets (November 2016 - March 2017). *Zenodo* (2017).
8. Parker, C. Brexit Tweets from the morning of it's announcement. *Mendeley Data* **v2**, (2017).
9. Adhokshaja, P. #Inauguration and #WomensMarch. *data.world* (2017).
10. Hoover, J. *et al.* Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Personal. Sci.* (2019).
11. Davidson, T., Warmsley, D., Macy, M., & Weber, I. Automated hate speech detection and the problem of offensive language. *11th AAAI CWSM*. (2017).