2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Do Bots Have Moral Judgement? The Difference Between Bots and Humans in Moral Rhetoric

Ece Çiğdem Mutlu, Toktam Oghaz, Ege Tütüncüler, Ivan Garibay*
University of Central Florida
{ece.mutlu, toktam.oghaz, ege.tutunculer, igaribay}@ucf.edu

***Abstract*—Understanding moral foundations can yield powerful results in terms of perceiving the intended meaning of the text data, as the concept of morality provides additional information on the unobservable characteristics of information processing and non-conscious cognitive processes. Considering that moral values vary significantly across cultures and yet many recurrent themes are observed and that each culture builds its societal and ideological narratives on top of its moral virtues, an enhanced understanding of morality can prove to be a valuable tool in deterring disinformation narratives by adversaries. Therefore, we investigate the evolution of latent moral loadings over time and across different sub-narratives on human and bot-generated tweets. For this purpose, we analyze the Syrian White Helmets-related tweets from April 1st, 2018 to April 30th, 2019. For the operationalization and quantification of moral rhetoric in tweets, we use Moral Foundations Dictionary in which five psychological dimensions (Harm/Care, Subversion/Authority, Cheating/Fairness, Betrayal/Loyalty and Degradation/Purity) are considered. Our results present the significant differences between the strength and patterns of moral rhetoric for human and bot-generated content on Twitter.**

***Index Terms*—Bot, latent semantic analysis, moral judgement, MFT, moral foundations, Twitter, White Helmets**

I. Introduction

It is essential to understand the differences in the notion of morality at both the cultural and individual levels, as morality guides human social interactions and can potentially lead to polarity, violence, and hostility when there is a clash of moral values within a society [1]. Differences in morality can result in polarity within a social group and can fuel societal tensions caused by disinformation efforts of malicious agents. Thus, identifying the bits-and-pieces of morality components in social media content, such as tweets revolving around certain narratives in Twitter, can help combat the spread of misinformation and the social engineering efforts of adversaries. Furthermore, gaining more in-depth insight into morality dimensions can potentially provide superior results compared to what the extant literature achieves by using sentiment analyses in combating disinformation [2].

Sharing information, interests, and opinions of individuals in online social media platforms has led scientists from various fields to focus on user-generated text data [3]. The availability of large-scale time-series text data from social media platforms has made the application of both unsupervised and supervised methods more rigorous and diverse such as classification and misinformation detection, bot detection, etc. To alleviate the difficulty of detecting malicious bot activities from textual content, the application of deep learning to extract context features have been considered in literature [4], [5]. However, supervised methods require annotated data for human utterances. To address the challenges for supervised techniques, many theory-driven approaches have been developed which benefit from the psychological dictionaries coupled with data-driven natural language processing methods. Among the developed dictionaries for this purpose, many studies have reported on the high overlap in the classification and identification of morality dimensions in Moral Foundations Theory (MFT) for human coders and lexicon-based methods [6], [7].

Literature have generally focused on the diversity of moral values of i) different groups on the same topic or, ii) a single population on various topics. All these studies show the differences in morality as an average rate or amount over a period of time. However, moral values and moral judgements may change as a result of social interactions or other external influences [8]. Therefore, we aim to elucidate the evolution of the moral values over time and across different sub-narratives via analyzing a Twitter data set on White Helmets of Syria. Since we aim to demonstrate the importance of moral foundations in combating misinformation, and it is established that software-controlled accounts (bots) spread more negative information to polarize societies [9], this work presents the differences between the moral rhetoric in human and bot-generated contents to investigate the potential use of moral foundations in the detection of malicious bot accounts.

II. Method

A. Quantifying Moral Foundations in a Text

In order to operationalize and capture dimensions of morality in our Twitter data set, we draw from social psychology literature and use the Moral Foundations Theory (MFT) [8], [10]. The moral rhetoric that provides a basis for our analyses can be defined as the linguistic component for expressing various moral concerns by taking a moral stance towards an issue [11]. MFT contends that five psychological subsystems constitute moral cognition, which manifests themselves as moral concerns or intuitions. Each of these five morality-related psychological components includes these dimensions of virtues and vices: i) Harm/Care concern is associated with the protection of self and others from the harm's way; ii) Subversion/Authority component expresses concerns related to subordination and respect; iii) Cheating/Fairness concern is

* Corresponding author.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

related to justice in cooperative acts, prevention of dishonesty, and reciprocity in social interactions; iv) Betrayal/Loyalty dimension is based on the expressions of self-sacrifice for both ends of the virtue-vice spectrum, such as patriotism-betrayal, faithfulness-unfaithfulness; and v) Degradation/Purity is associated with sanctity in the virtue dimension and degradation and pollution in the vice dimension.

To quantify the moral foundations in Twitter text data, we compared the efficiencies of the three most common dictionaries. For this purpose, we first annotated 400 tweets in the data set and compared the human-annotation results with the labels obtained via original Moral Foundations Dictionary (MFD)[1], its enlarged (MFD 2.0) [12] and extended (eMFD) [13] versions. We found 78.75, 93.75 and 84.00% similarity between the human-annotated tweet labels and the dictionary results, respectively; therefore, we used MFD 2.0 to quantify the latent moral loadings in the further analyses.

B. Semantic Similarity to Moral Foundations

To compute the proximity between the moral words in MFD 2.0 and the tweets in our data set, we used two vector space models to generate feature vector representations: i) term frequency-inverse document frequency (tf-idf) [14], and ii) distributed bag-of-words paragraph vectors model based on Doc2vec [15]. Suppose that $T^{i,t}$ is the i^{th} tweet in data set published at time t. We converted each tweet ($T^{i,t}$) to a vector in the semantic space using the Doc2vec model by forming a vector $r^{i,t} = (r_1^{i,t}, \ldots, r_v^{i,t})'$.

$$M_d^{i,t} = \frac{r^{i,t} * f_d^i}{|r^{i,t}| * |f_d^i|} \tag{1}$$

where $M_d^{i,t}$ is the moral score of i^{th} tweet created at time t in d moral foundations dimension. The algorithm gives ten results for each tweet input; these are the scores of the vice and virtue corresponding to the five dimensions of moral foundations. Each $T^{i,t}$ may either have a vice rhetoric in the first moral dimension (Care) of Care/Harm $M_{+1}^{i,t}$ or a virtue rhetoric (Harm) $M_{-1}^{i_t}$ according to whether the tweet includes similar words in the lexicons of corresponding dimensions.

C. Dataset Description

We investigated a Twitter data set which was provided by Leidos Inc[2] as part of the "Computational Simulation of Online Social Behavior (SocialSim)" DARPA program.This data consists of 1,052,821 tweets related to the disinformation campaigns carried against the White Helmets from April 1st, 2018 to April 30th, 2019. White Helmets belong to a civilian volunteer organization that operates in war-torn areas and provides services, such as the evacuation of people from bombed areas, search and rescue, delivering medical supplies, etc. The narratives within the content of these tweets are mostly attacks against the integrity of the White Helmets' work and mission statement, conspiracy theories, the accusation of the organization for being involved with foreign agents, and nullifying the narrative of the chemical attack by censuring the organization of staging the event.

D. Sub-narrative Identification with Topic Analysis

Topic discovery for social media content has been widely studied in literature [16], [17]. In this paper, we employed Latent Dirichlet Allocation (LDA) [18] to capture the mixture of topics over a collection of tweets. LDA is a generative probabilistic model with three layer Hierarchical Bayesian architecture, which assumes a finite set of vocabulary for each topic, and a set of topic probabilities for each term. As one of the most challenging steps of applying LDA is tuning parameters, specially, defining the number of topic clusters (K) and their probable initial proportions, we utilized perplexity and coherence scores besides human judgment for this purpose. We repeated LDA with the number of topics as $K \in \{2, \ldots, 20\}$ with increments of size 1. Additionally, different solver techniques for optimization of LDA algorithm have been tried and best perplexity and coherence values are obtained with the use of stochastic variational Bayes solver [19]. We also observed the highest coherence score and the lowest perplexity score for $K = 4$.

III. Results

This study aims to investigate the latent moral loadings of White Helmets-related posts of Twitter users and to understand the differences in the dynamics of moral rhetoric of human and bot-generated contents. To question the necessity of moral foundations analysis and the capability of the MFT 2.0 dictionary, we measured the ratio of moral words to non-moral words and recognized that almost 57% of the tweets include more moral than non-moral words, unsurprisingly, and only 1.7% of whole tweets do not include any moral rhetoric. Additionally, to identify the bot accounts in the data set, we used the Botometer scores[3] in the range from 0 to 5, in which a high score represent a high likelihood of being bot. Despite the appearance of different arbitrary thresholds for this variable on literature, considering a middle Botometer score as the threshold have been identified to be more appropriate for classification. A high threshold results in the misclassification of many bot accounts as humans, while a low Botometer score as threshold increases the probability of assigning human accounts as bots [20]. Accordingly, we accepted our threshold within the same range as in [21], such that the accounts whose Botometer scores are higher than 3.5 are classified as bot accounts. As a result, 5.02 % of the accounts are classified as bot accounts and further analyses are conducted based on this classification.

RQ1: *Is there a significant difference between the moral rhetoric of human-generated tweets and bot-generated tweets?*

First, we calculated the average moral loading scores of contents generated by human and bots in each dimension, to compare the differences between moral rhetoric of human and bot-generated tweets. Results show that humans tend to share

[1]https://moralfoundations.org

[2]https://www.leidos.com

[3]https://botometer.iuni.iu.edu

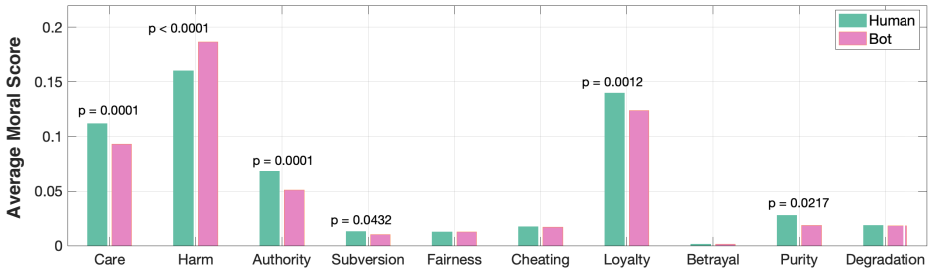
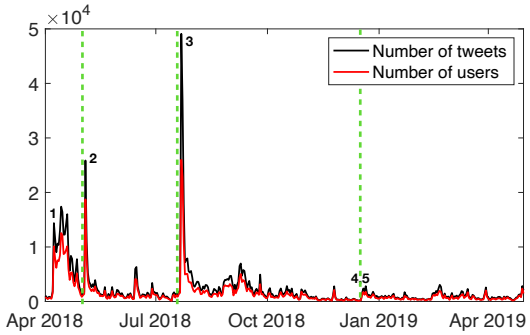


Fig. 1. Average moral scores of tweets generated by bot (pink) and human (green) users in each MFT dimensions (The p-values of Mann-Whitney U test results are written above to the bars in case of significant difference between scores the significant differences).

tweets that involve a stronger virtue moral rhetoric comparing to bots, while the strength of the moral rhetoric is higher in each vice dimension for bots compared to humans. This result demonstrates that bots spread negative information not only sentimentally but also in terms of moral rhetoric. Even though there is no significant difference between the use of Fairness, Cheating Betrayal and Degradation rhetoric, the significant differences (p values of the Mann-Whitney U tests are written on the top for each pairwise comparison) are observed in all commonly observed moral dimensions (Figure 1).

Furthermore, we examined the daily number of user activities and the number of unique users who are involved in these activities to obtain a better understanding of the data (Figure 2). Unsurprisingly, the long time range of the data set and the diversity of the events brought instability in daily user activities, and some essential events triggered the burstiness of the specific Twitter cascades at specific times. Users' stances might change over time, or the changes in the ongoing set of events might affect people's attitudes even if there is no external effect from another source. Therefore, we examined how moral foundations change across time on different sub-narratives.



1 - Douma chemical attack (source: *New York Times 04/08/18*)
2 - Trump's Syria aid freeze hits 'White Helmets' rescue workers (source: *ABC News 05/04/18*)
3 - Syrian White Helmets evacuated to Jordan through Israel (source: *CBC News 07/22/18*)
4 - The Russia-backed campaign to link the volunteer rescuers with al-Qaida exposes how conspiracy theories take root: 'It's like a factory' (source: *The Guardian 12/18/18*)
5 - Air strikes kill at least 9 civilians in Syria's Idlib - White Helmets (source: *TRT 12/23/18*)

Fig. 2. Daily number of user activities (black line) and unique user involved (red line). (Titles of the news possibly related to the bursts in Twitter cascades are given below and shown with green dashed line.

RQ2: *Does the moral rhetoric of user-generated content of bots and humans on Twitter change over time?*

To understand the change in the moral foundations across time, we determined each time range by considering the major events that caused bursts in Twitter activities. The first time range t_1 covers tweets between the time of Douma chemical attack and Trump's Syria aid freeze (April 1st, 2018 - May 04th, 2018). The second time range t_2 starts with the end of the first time period and covers the events until the evacuation of White Helmets to Jordan through Israel (May 04th, 2018 - Jul 22th, 2018). Despite the relative un-burstiness of the rest of the data, third t_3 and fourth time ranges t_4 are divided before and after becoming the target of a disinformation campaign that positions White Helmets as an al-Qaida-linked terrorist organization (Dec 18th, 2018).

Figure 3 shows how the use of vice (Harm, Subversion, Cheating, Betrayal, Degradation) and virtue (Care, Authority, Fairness, Loyalty, Purity) moral rhetoric has varied across time in our data set. Figure 3.a1-b1 shows the time series of the average moral loading scores of vice dimensions in human and bot-generated tweets respectively. The strength of moral rhetoric of the tweets show stationary behavior with small deviations across time in human-generated tweets. In bot-generated tweets, on the other hand, these fluctuations show more chaotic and random-walk pattern despite preserving the stationary structure ($p < 0.005$). We did not present the time-series of the average moral loading scores of virtue dimensions due to the similarity of the results.

Moral loadings in the five dimensions of the vice moral foundations by four different time ranges for human-generated tweets (Figure 3.a2) and bot-generated tweets (Figure 3.b2) are also presented. These sub-figures explain how language around certain words and concepts evolves over time in terms of the strength in their rhetoric. In Figure 3, each moral dimension j between t_1 and t_2 time periods is obtained as follows:

$$\frac{\sum_i \sum_{t_1<t^*<t_2} M_{+j}^{i,t^*}}{N\left(\sum_i M_{+j}^{i,t_1<t^*<t_2}\right)}. \quad (2)$$

where, the denominator calculates the number of tweets that are classified as having a specific moral loading score, i.e. Care, and are published between t_1 and t_2.

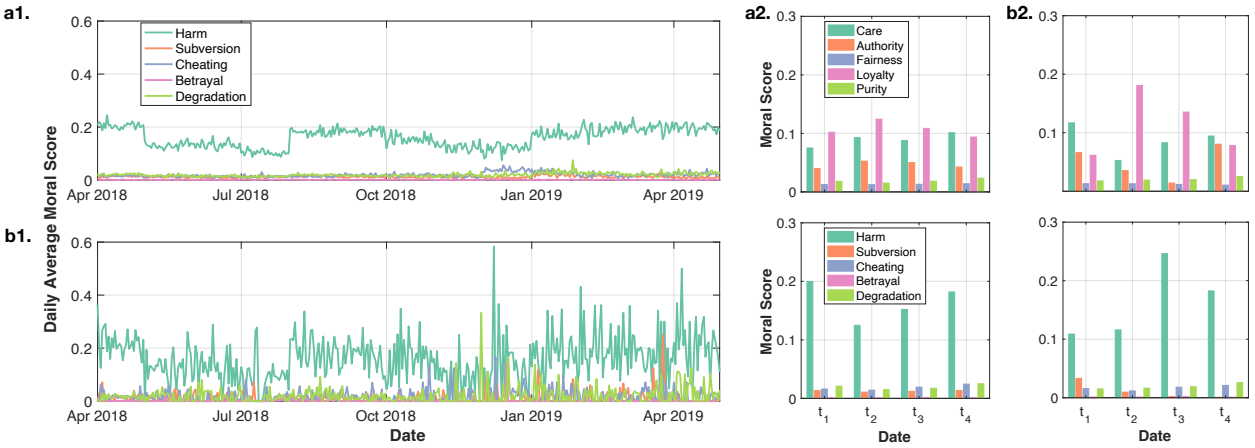


Fig. 3. The daily average moral loading scores of vice dimensions as a time-series in a1. human b1. bot-generated tweets. Bar plots show that the average moral scores for virtue (top) and vice (bottom) dimensions for t_1 (April 1st, 2018 to May 04th, 2018), t_2 (May 04th, 2018 to Jul 22th, 2018), t_3 (Jul 22th, 2018 to Dec 18th, 2018) and t_4 (Dec 18th, 2018 to Apr 30th, 2019) in a2. human b2. bot-generated tweets.

TABLE I
KEYWORDS USED IN THE TOPIC LABELING OF TWEETS

Topic	Keywords
$topic_1$	'chemical','attack','douma'
$topic_2$	'trump','freeze','fund'
$topic_3$	'evacuate','Jordan','Israel','rescue'
$topic_4$	'airstrikes','idlib','civilian','kill','russia'

Results mainly show that the pattern of the moral rhetoric of the tweets among five dimensions are very similar while its amount varies over time, i.e. tweets including Harm, Care and Loyalty moral rhetoric are generally higher in moral loading score. The least strong moral foundation in terms of moral loading scores in our data set is Betrayal. Here, we can conclude that people are most likely polarized in their moral judgement in Loyalty/Betrayal dimension of MFT. Although the dominance of the Care, Loyalty and Harm rhetoric is obvious in both human and bot tweets, human-generated tweets show more stable moral pattern across time. This can be explained with the steadiness of human moral judgement across an issue with the passing time. This phenomena is less observed in bot-generated tweets since they are automatically generated without any moral judgement.

RQ3: *Does the moral rhetoric of user-generated content on Twitter change across different sub-narratives of the same domain?*

For further analysis, we investigated the diversity in the MFT dimensions across the four main identified topics. We observed that the LDA results fairly matched with the main sub-narratives of identified events for the same time period. We manually labeled tweets according to their involvement of specific keywords as $topic_k, k \in \{1, \ldots, 4\}$ (Table I).

Figure 4 demonstrates how positive/virtue (left) and negative/vice (right) emotions/stances vary across four major topics among human and bot-generated tweets. To obtain these results, for each moral dimension j and topic k, the average value of the moral scores are computed as follows:

$$\frac{\sum_t \sum_{i \in topic_k} M_{+j}^{i,t}}{N\left(\sum_t M_{+j}^{i \in topic_2, t}\right)}, \tag{3}$$

where, the denominator calculates the number of tweets that are classified as having a moral loading score of dimension j, include one of the keyword set of topic k, and not include other keywords given in Table 1.

Figure 4.a and 4.b show the strength of the five virtue and vice moral dimensions for human and bot-generated tweets across four topics. We observed that the abundance of Harm rhetoric across all topics is noteworthy and it is more significant for bot-generated tweets. Furthermore, the dominance of the Loyalty and Care dimensions vary across different topics. Additionally, we observed that average virtue moral loading values are very close to each other in Topics 1 and 4 for the human-generated tweets. It means that different kind of attacks to the Syrian White-Helmets trigger each dimension of the moral values of online audiences. For Topic 2, the funding freeze event caused tweets with a higher Loyalty rhetoric. However, both Loyalty and Care have been observed with higher values for the tweets related to the evacuation event. The results for vice rhetoric of human-generated tweets are more distinctive. While the attacks in the first and the fourth topics caused more Harm tweets to emerge, the tweets belonging to the second and third topics did not show strong rhetoric in any vice dimension. The most important finding here is that, bot-generated tweets show almost completely-similar behavior across all sub-narratives while human-generated tweets have more sensible and predictable explanations.

IV. CONCLUSIONS

Although literature have used sentiment analysis widely, gaining more in-depth insight into moral dimensions coupled with the extant NLP techniques, can potentially provide a

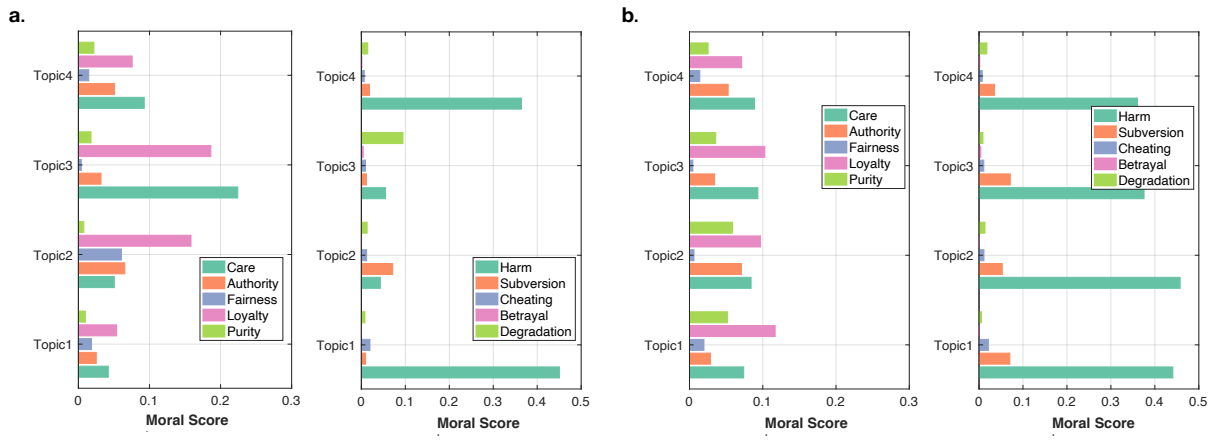


Fig. 4. The average moral loading scores of a. human, b. bot-generated tweets in virtue (left) and vice (right) dimensions across the four identified topics.

better understanding to the reasoning behind the choices and feelings of individuals and the intended meaning of textual content. In this regard, we investigated the evolution of moral foundations in the human and bot-generated tweets by time across different sub-narratives of a single domain (Syrian White Helmets). We observed that: i) bot accounts use stronger vice rhetoric than humans; ii) the moral scores of human–generated contents show more stationary behavior over time comparing to chaotic patterns for bots; and iii) human-generated content show predictable behavior over time comparing to bots. To best of our knowledge, this work is the first study that provides a comparison of moral rhetoric for human and bot-generated tweets. These results might be further supported by varying data sets to test the scalability on robustness of the results. These finding also show that moral rhetoric can be an important attribute for malicious bot activity detection.

Acknowledgment

This work was partially supported by grant FA8650-18-C-7823 from the Defense Advanced Research Projects Agency.

References

[1] J. Graham, "Morality beyond the lab," *Science*, vol. 345, no. 6202, pp. 1242–1242, 2014.

[2] R. Rezapour, S. H. Shah, and J. Diesner, "Enhancing the measurement of social effects by capturing morality," in *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2019, pp. 35–45.

[3] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network to recommend airline tickets by using online reviews," in *IEEE 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020*, 2020.

[4] M. Heidari, J. H. J. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[5] M. Heidari and J. H. J. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *IEEE 2020 11th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, UEMCON 2020*, 2020.

[6] R. Weber, J. M. Mangus, R. Huskey, F. R. Hopp, O. Amir, R. Swanson, A. Gordon, P. Khooshabeh, L. Hahn, and R. Tamborini, "Extracting latent moral information from text narratives: Relevance, challenges, and solutions," *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 119–139, 2018.

[7] W. Hofmann, D. C. Wisneski, M. J. Brandt, and L. J. Skitka, "Morality in everyday life," *Science*, vol. 345, no. 6202, pp. 1340–1343, 2014.

[8] J. Haidt, "The new synthesis in moral psychology," *science*, vol. 316, no. 5827, pp. 998–1002, 2007.

[9] M. Stella, E. Ferrara, and M. De Domenico, "Bots increase exposure to negative and inflammatory content in online social systems," *Proceedings of the National Academy of Sciences*, vol. 115, no. 49, pp. 12 435–12 440, 2018.

[10] J. Haidt and C. Joseph, "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues," *Daedalus*, vol. 133, no. 4, pp. 55–66, 2004.

[11] E. Sagi and M. Dehghani, "Measuring moral rhetoric in text," *Social science computer review*, vol. 32, no. 2, pp. 132–144, 2014.

[12] J. Frimer, R. Boghrati, J. Haidt, J. Graham, and M. Dehgani, "Moral foundations dictionary for linguistic analyses 2.0," *Unpublished manuscript*, 2019.

[13] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text," 2020.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[15] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Intern. conf. on machine learning*, 2014, pp. 1188–1196.

[16] T. A. Oghaz, E. C. Mutlu, J. Jasser, N. Yousefi, and I. Garibay, "Probabilistic model of narratives over topical trends in social media: A discrete time model," in *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, ser. HT '20. Association for Computing Machinery, 2020, p. 281–290.

[17] F. Jafariakinabad and K. A. Hua, "Maximal sequence mining approach for topic detection from microblog streams," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–8.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[20] A. Rauchfleisch and J. Kaiser, "The false positive problem of automatic bot detection in social science research," *Berkman Klein Center Research Publication*, no. 2020-3, 2020.

[21] T. R. Keller and U. Klinger, "Social bots in election campaigns: Theoretical, empirical, and methodological implications," *Political Communication*, vol. 36, no. 1, pp. 171–189, 2019.