Hoaxes and Hidden agendas: A Twitter Conspiracy Theory Dataset

Data Paper

Samantha C. Phillips
samanthp@andrew.cmu.edu
Institute for Software Research,
Carnegie Mellon University
Pittsburgh, U.S.A.

Lynnette Hui Xian Ng
huixiann@andrew.cmu.edu
Institute for Software Research,
Carnegie Mellon University
Pittsburgh, U.S.A.

Kathleen M. Carley
kathleen.carley@cs.cmu.edu
Institute for Software Research,
Carnegie Mellon University
Pittsburgh, U.S.A.

ABSTRACT

Hoaxes and hidden agendas make for compelling conspiracy theories. While many of these theories are ultimately innocuous, others have the potential to do real harm, instigating real-world support or disapproval of the theories. This is further fueled by social media which provides a platform for conspiracy theories to spread at unprecedented rates. Thus, there is a need for the development of automated models to detect conspiracy theories from the social media space in order to quickly and effectively identify the topics of the season and the prevailing stance. To support this development, we create ground truth data through human annotation. In this work, we collect and manually annotate a dataset from Twitter, comprising of four conspiracy theories. Each Tweet is annotated with one of the four topics {climate change, COVID-19 origin, COVID-19 vaccine, Epstein-Maxwell trial}, and its stance towards the conspiracy theory {support, neutral, against}. We perform experiments on this multi-topic dataset to demonstrate its usage in conspiracy-detection, stance-detection and topic-detection.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**.

KEYWORDS

conspiracy theory, dataset, Twitter, stance detection, classification

ACM Reference Format:
Samantha C. Phillips, Lynnette Hui Xian Ng, and Kathleen M. Carley. 2022. Hoaxes and Hidden agendas: A Twitter Conspiracy Theory Dataset: Data Paper. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487553.3524665

1 INTRODUCTION

Conspiracy theories are unsubstantiated narratives designed to explain significant social or political events with secret plots by malicious and powerful actors [7, 8]. People who believe in conspiracy theories typically either claim such an event is a hoax or a real and major threat as a part of a broader hidden agenda. While a large part of these theories are treated as stories, some of them have instigated real-world consequences. A well-known example is the Pizzagate conspiracy theory that went viral during the 2016 United States presidential election, purporting a child sex ring at the basement of a pizza shop. This resulted in harassment of the owner and employees of the pizza shop, and shots fired in the restaurant by an individual attempting "save the children" [4, 15]. Social media provides the platform for the chaos of conspiracy theories to thrive and spread as digital wildfire [7]. Moreover, these theories offer avenues for nefarious actors to influence or manipulate vulnerable communities with potentially large consequences for social and political processes [5, 7].

As such, it is of paramount importance to identify conspiracy theories to aid governmental entities, journalists and social media companies with mitigating the spread of these claims. Another vital research area is the development of models to classify the stance of social media texts as supportive, against or neutral towards any conspiracy theory for further exploration of the dynamics of users engaged in conspiracy theories online.

In this study, we collect and analyze Twitter data surrounding four conspiracy theories. This work lies at the intersection of natural language processing, stance detection and mis/disinformation mitigation for social media. Our contributions are as follows:

(1) We collect, manually annotate, and publicly share a conspiracy theory dataset with 3100 tweets. This dataset consists of conspiracy theories across more than one domain, from climate change to public health to discussion around criminal trials.[1]
(2) We fine-tune text classification models to detect whether a tweet (1) contains a conspiracy or not; (2) the stance of the tweet; (3) the topic of the tweet.

2 RELATED WORK

There has been substantial work exploring automated text analysis of conspiracy theories. The COVID-19 pandemic provided a ripe breeding ground for conspiracy theories to thrive, resulting in a series of COVID-19 conspiracy theory studies [14, 27]. Corpora of COVID-19 tweets have been collected and prior work performed conspiracy-detection classification using random forest [10], support vector machine methods [23] and neural network models [21]. Another study uses word embeddings to identify semantic

[1] https://github.com/samanthph/twitter-consp-theory-data.git

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524665

properties of conspiracy theory tweets associated with epistemic, existential, and social motives for believing in such theories [2].

One common part of conspiracy theory analysis is stance detection, the identification of support or disapproval towards a conspiracy theory. Prior work built stance detection models for conspiracy theories towards climate change [26] and vaccines [13]. A recent study annotates 10,000 tweets related to 5G and COVID-19 conspiracy theories with labels indicating support or otherwise [24]. More broadly, several studies use stance detection to detect support or disapproval towards misinformation for fact-checking purposes [3, 9].

Conspiracy theories can take many forms of misinformation, perhaps spread maliciously to sew discourse or intended to inform or protect others. Within this social media landscape, Alam et al. [1] instituted a "call to arms", encouraging the research community to publish their data to contribute to the fight against mis/disinformation online and aid in debunking these proposed hoaxes and hidden agendas. We hope to contribute to the effort to build effective conspiracy-detection and stance-detection models by constructing a publicly available dataset containing multiple topics and annotated with the stance towards a topic.

3 DATASET CONSTRUCTION

3.1 Data Curation

We collected data on four topics using Twitter V1 API. We collected only public tweets and no personal identification information was used in the data collection nor experimentation. In order to construct a generic conspiracy theory dataset, we collected data on a mixture of topics, each containing its own hoaxes and hidden agendas as conspiracy theories. Table 1 contains the collection parameters and examples of related conspiracy theories for the four topics in this dataset.

3.2 Data Annotation Procedure

Three graduate students from computer science whose first language was English performed the data annotation. Two annotators worked independently and the third broke ties in the event of disagreement between the initial two labels. The annotation was done in two passes. The annotators first performed an initial filter of tweets that were not in the English language (299 instances) or had insufficient text (17 instances). These tweets were then removed from the dataset, leaving us with 3100 tweets.

The second annotation pass involved labelling each tweet for whether it contains a conspiracy theory or not. For tweets that contain conspiracy theories, the annotators labelled its stance, i.e. whether the tweet was supportive, against or neutral towards the conspiracy theory. This annotation process produced a Cohen's Kappa of 0.478, a moderate inter-annotator agreement between the two independent annotators [18]. Each tweet in the dataset is annotated with three values: contains a conspiracy theory or not, stance towards the conspiracy theory and topic of conspiracy theory. The topic of the tweet was obtained from the data collection phase. Each instance in the dataset contains the tweet ID and the three corresponding labels. We present the annotation scheme and the percentage of each label in the dataset in Table 2.

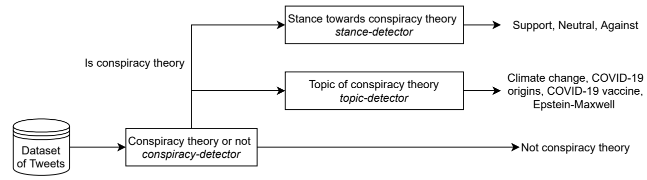

Figure 1: Illustration of performed experiments.

4 EXPERIMENTS

We analyze the dataset through three experiments, each building language models on a different aspect of the annotation, to demonstrate how to use this dataset. Figure 1 illustrates the experimental set up. Specifically, we run the following three experiments:

- **Conspiracy-detection**: Given a tweet, classify whether it contains a conspiracy theory or not.
- **Stance-detection**: Given a tweet that contains a conspiracy theory, classify its stance towards the conspiracy theory, i.e. {support, neutral, against}.
- **Topic-detection**: Given a tweet that contains a conspiracy theory, classify which of the four conspiracy theory topics it addresses.

For each experiment, we built five classifiers: a majority classifier as a baseline classifier which assigns the majority class as a class value for all the instances; and four neural network classifiers, BERT [6], ALBERT [17], RoBERTa [19] and DistilBERT [25]. The four neural network classifiers use the transformers architecture and are implemented using the *simple-transformers* Python library[2]. Prior to classification, the words in the tweets were tokenized with the default tokenizer for the model. These tokenizers make use of contextualized word embeddings, meaning they create vectors representing the words in relation to the surrounding words in the tweet. The classifiers were trained with the default classifier settings of the library: a learning rate of 4e-5, Adam parameter of 1e-8 and a batch size of 8. The proportion of each class label was passed as a parameter to assign weights for loss calculation, thereby accounting for unequal instances of each class. We ran each training run for 10 epochs with an early stopping criterion, and we observed the loss values typically converged after 2 epochs. Each classifier was ran with a five-fold cross validation using a 80:20 train:test split and the average performance metrics were reported. We used the macro-F1 performance metrics to adjust for the imbalance in the proportion of class labels in the dataset [12].

5 RESULTS

We present the results of our three experiments in Tables 3, 5 and 7. For the best classifier of the each experiment, we present the performance metrics by class in Tables 4, 6 and 8. The best classifier was determined by the highest macro-F1 score.

For all three experiments, the neural network classifiers outperform the baseline majority classifier, highlighting the need for contextualized tweet embeddings in aiding a classification model to differentiate labels. The best performing classifier for differentiating between conspiracy theory and non-conspiracy theory tweets is

[2]https://github.com/ThilinaRajapakse/simpletransformers

Table 1: Collection parameters and conspiratorial themes of the four topics.

Topic (%)	Collection parameters	Conspiratorial Themes
Climate change (25.5%)	Collection terms: Climate Change, #ActOnClimate, #Climate-Change. Collection dates: 26 Aug 2017 - 14 Sept 2019 [26]	Climate change is a hoax designed to control people, is a hidden agenda by the Deep State, or can be caused by chemical trails from aircrafts and geoengineering. [26]
COVID-19 origins (32.6%)	Collection terms: bleach, vaccine, acetic acid, steroids, essential oil, saltwater, ethanol, children, kids, garlic, alcohol, chlorine, sesame oil, conspiracy, 5G, cure, colloidal silver, dryer, bioweapon, co- caine, hydroxychloroquine, chloroquine, gates, immune, poison, fake, treat, doctor, senna makki, senna tea, #nCoV20199, #CoronaOutbreak, #CoronaVirus, #Coronavirus-Coverup, #CoronavirusOutbreak, #COVID19, #Coronavirus, #WuhanCoronavirus, #coronaviris, #Wuhan. Collection dates: 29 Mar, 15 Jun, Jun 2020 [20]	COVID-19 virus symptoms are due to 5G communication networks, or Bill Gates and other nefarious actors artificially created COVID-19 in an effort to control people. More broadly, COVID-19 is a bioweapon and/or was synthesized by humans. [14, 22]
COVID-19 vaccine (14.7%)	Collection terms: coronaravirus, coronavirus, wuhan virus, wuhanvirus, 2019nCoV, NCoV, NCoV2019, covid-19, covid19, covid 19. Collection dates: 21 Jan - 9 Jun 2020	Bill Gates funded vaccines to place devices like microchips into people to control them, 5G communication networks aid in controlling people through vaccines, or the vaccine changes a person's DNA. [11, 13, 14]
Epstein-Maxwell (27.2%)	Collection terms: epsteincoverup, GhislaineMaxwellTrial, JeffreyEpstein, LolitaExpress, PedophileIsland, epsteinDidntKill-Himself. Collection dates: 7-12 Dec 2021	Jeffrey Epstein did not kill himself in jail in 2019. There is an international scheme involving elites to cover up crimes of Epstein and powerful associates like Hillary and Bill Clinton, or there is an elaborate cover-up by the US government because they funded the crimes. [4]

Table 2: Annotation Scheme (usernames are redacted to maintain user privacy).

Annotation Label (%)	Description	Example Tweet
Initial Filters		
A	not in English language	El CUHC se tramita en [...]
B	insufficient text	https://t.co/SNru2RHs5g
Conspiracy Theory or Not		
0 (24.7%)	does not contain a conspiracy theory	This is a good article about what inspired the main reporter to bring Epstein down, which lead to the arrest of his accomplice, #GhislaineMaxwell #GhislaineMaxwellTrial #Ghislaine #Epstein
1 (75.3%)	contains a conspiracy theory	climate change = hoax created by china directed energy weapons attacking most of us
Stance towards conspiracy theory		
1 (11.7%)	against a conspiracy theory	@ow***: voter fraud isnt real nor is the deep state or qanon. climate change is real. the wage gap is real. gun violence is all too real
2 (29.4%)	neutral towards a conspiracy theory	Study finds over 40% of Republicans think Bill Gates will use a COVID-19 vaccine to implant a tracking microchip
3 (58.8%)	support towards a conspiracy theory	Because the #deepstate would not be exposed and there would be no need for a #plandemic or a murder to instigate riots. The sheeple have no idea what #hillaryclinton and the #deepstate are capable of.

RoBERTa (macro-F1=0.813). RoBERTa had the best performance for the *conspiracy theory* class (macro-F1=0.908). The best performing classifier for differentiating the stance of tweets towards a conspiracy theory is BERT (macro-F1=0.722). This BERT classifier for stance-detection performs best on the class label *Support* (macro-F1=0.822). For differentiating conspiracy theory topics, RoBERTa fared the best (macro-F1=0.944), and had the best performance for the *Epstein-Maxwell* class label (macro-F1=0.997).

Table 3: Conspiracy-detection results

Model	Precision	Recall	Macro-F1
Majority classifier	0.377	0.500	0.430
BERT	0.812	0.802	0.807
ALBERT	0.734	0.783	0.752
RoBERTa	0.825	0.803	**0.813**
DistilBERT	0.801	0.807	0.804

Table 4: Performance by class of best-model (RoBERTa) for conspiracy-detection experiment

Class	Precision	Recall	Macro-F1
Conspiracy theory	0.893	0.924	**0.908**
Not conspiracy theory	0.757	0.681	0.717

Table 5: Stance-detection results

Model	Precision	Recall	Macro-F1
Majority classifier	0.190	0.333	0.242
BERT	0.731	0.714	**0.722**
ALBERT	0.550	0.607	0.563
RoBERTa	0.698	0.707	0.702
DistilBERT	0.636	0.643	0.639

Table 6: Performance by class of best-model (BERT) for stance-detection experiment

Class	Precision	Recall	Macro-F1
Against	0.697	0.676	0.686
Neutral	0.701	0.616	0.656
Support	0.796	0.850	**0.822**

Table 7: Topic-detection results

Model	Precision	Recall	Macro-F1
Majority classifier	0.080	0.250	0.122
BERT	0.922	0.878	0.893
ALBERT	0.884	0.869	0.874
RoBERTa	0.966	0.929	**0.944**
DistilBERT	0.915	0.856	0.873

Table 8: Performance by class of best-model (RoBERTa) for topic-detection

Class	Precision	Recall	Macro-F1
Climate change	1.00	0.972	0.986
COVID-19 origins	0.925	1.00	0.961
COVID-19 vaccine	0.946	0.745	0.833
Epstein-Maxwell	0.994	1.0	**0.997**

6 DISCUSSION

The dataset annotation achieved a moderate agreement between the annotators. An inherent attribute of conspiracy theories is how quickly they evolve, often contradicting each other. For instance, theories simultaneously claim Princess Diana faked her own death and was murdered [28]. Despite training each annotator on current conspiracy theories, there is room for interpretation. Other annotation studies involving hate speech or personal attack detection also have moderate or fair inter-annotator agreement [16, 29]. While the use of multiple annotators helps mitigate this issue, subjective labelling naturally results in disagreements between annotators with different backgrounds and experiences.

The subset of the data used in the stance-detection and topic-detection models only contains tweets with conspiracy theories, hence having 764 fewer data points than the conspiracy-detection model. Yet the topic-detection model yields the best performance while the stance-detection model returns the worst. The stance-detection model might have struggled with performance due to an imbalance of class labels (see Table 2 for statistics). Conversely, topic-detection models fared well due to the uniqueness of the tweet text of each topic and ability of large-scale language models to memorize key language characteristics. There is also a significant imbalance of class labels for the conspiracy-detection task. Of the 3100 instances in our dataset, 75.3% contains a conspiracy theory and 24.7% do not. We note that the difference between tweets identified as containing a conspiracy theory is due to the collection terms used and does not reflect the true ratio on Twitter. Neural network classifiers account for class imbalance with defined class weights, setting a higher class weight for the minority class and lower weight for the majority.

Given the performance of the neural network classifiers across all three experiments is macro-F1=0.753±0.190, we infer that one does not require a huge dataset of annotated data to identify conspiracy theories, stances or topics. This observation may be due to common text structures among conspiracy theories; it is comforting in the development of subsequent models to keep pace with the changing information landscape.

Furthermore, for the conspiracy-detection task, DistilBERT took a third of the time RobERTa took to run with only a 0.009 drop in macro-F1. For the stance-detection task, DistilBERT ran in approximately half of the time BERT took with a decrease in macro-F1 values of 0.083. Finally, in the topic-detection task, DistilBERT took a third of the time RoBERTa took with a macro-F1 drop of 0.071. These results indicate the additional parameters and time required by more complex text analysis models like BERT and RoBERTa may not be necessary for this dataset, depending on the performance desired.

We posit that users with extreme opinions are typically more vocal on social media, suggesting caution in extrapolating findings. Moreover, homophily on social media interactions has been well-studied. As one may suspect, there is evidence that people tend to interact with other users similar to them in beliefs or other attributes [30]. Users who do not believe in conspiracy theories may not seek out or interact with users spreading conspiracy theories. The ratio of conspiracy supporters and deniers in the dataset is roughly 4:1. In this sample of tweets, users were more outspoken in promoting these theories than shutting them down. A future research path is to use these models to detect stance towards conspiracy theories on a much larger scale, which could allow researchers to conclude if supporters or deniers of conspiracy theories discuss them more online.

While our classifiers obtained acceptable performance, in the current information landscape, conspiracy theories evolve rapidly and can often contradict each other. The list of conspiracies collected in this work is non-exhaustive. As such, we should employ caution when generalizing the model and the model should be continually updated with new data and multi-lingual data. Future work calls for more refined language models and for studying the dynamics and evolution of conspiracy theories.

7 CONCLUSION

We present a new manually-labelled conspiracy theory Twitter dataset connected to four events, {climate change, COVID-19 origins, COVID-19 vaccine, Epstein-Maxwell}, annotated for the presence of a conspiracy theory, stance towards a conspiracy theory and topic. We also demonstrate using our dataset with five standard classifiers across three tasks: conspiracy-detection, stance-detection and topic-detection. The best-performing model performance ranges from macro-F1=0.722 to macro-F1=0.944 across experiments. This study demonstrates that identifying and classifying conspiracy theories within tweets is possible, even across domains that range from climate change to public health conspiracy theories. We hope to contribute to the development of tools for identifying conspiracy theories before they become widespread, enabling effective public messaging and mitigation.

ACKNOWLEDGMENTS

The research for this paper was supported in part by the Knight Foundation and the Office of Naval Research grant N000141812106 and by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the US Government.

REFERENCES

[1] Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. 2020. Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms. *arXiv preprint arXiv:2007.07996* (2020).

[2] Veronika Batzdorfer, Holger Steinmetz, Marco Biella, and Meysam Alizadeh. 2021. Conspiracy theories on Twitter: emerging motifs and temporal dynamics during the COVID-19 pandemic. *International Journal of Data Science and Analytics* (2021), 1–19.

[3] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Companion Proceedings of the The Web Conference 2018*. 1353–1357.

[4] Paul Bleakley. 2021. Panic, pizza and mainstreaming the alt-right: A social media analysis of Pizzagate and the rise of the QAnon conspiracy. *Current Sociology* (2021), 00113921211034896.

[5] Kathleen M Carley. 2020. Social cybersecurity: an emerging science. *Computational and mathematical organization theory* 26, 4 (2020), 365–381.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political Psychology* 40 (2019), 3–35.

[8] Adam M Enders, Joseph E Uscinski, Michelle I Seelig, Casey A Klofstad, Stefan Wuchty, John R Funchion, Manohar N Murthi, Kamal Premaratne, and Justin Stoler. 2021. The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political behavior* (2021), 1–24.

[9] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 1163–1168.

[10] Dax Gerts, Courtney D Shelley, Nidhi Parikh, Travis Pitts, Chrysm Watson Ross, Geoffrey Fairchild, Nidia Yadria Vaquera Chavez, and Ashlynn R Daughton. 2021. "Thought I'd Share First" and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study. *JMIR Public Health Surveill* 7, 4 (14 Apr 2021), e26527. https://doi.org/10.2196/26527

[11] Jack Goodman and Flora Carmichael. 2020. Coronavirus: Bill Gates 'microchip'conspiracy theory and other vaccine claims fact-checked. *BBC Reality Check. May* (2020).

[12] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1859–1874. https://aclanthology.org/C18-1158

[13] Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* 203 (2022), 23–30.

[14] Daniel Jolley and Jenny L Paterson. 2020. Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British journal of social psychology* 59, 3 (2020), 628–640.

[15] Cecilia Kang and Adam Goldman. 2016. In Washington Pizzeria attack, fake news brought real guns. https://www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

[16] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[18] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791* (2020).

[21] JD Moffitt, Catherine King, and Kathleen M Carley. 2021. Hunting Conspiracy Theories During the COVID-19 Pandemic. *Social Media+ Society* 7, 3 (2021), 20563051211043212.

[22] Tomasz Oleksy, Anna Wnuk, Dominika Maison, and Agnieszka Łyś. 2021. Content matters. Different predictors and social consequences of general and government-related conspiracy theories on COVID-19. *Personality and individual differences* 168 (2021), 110289.

[23] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. *WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets*. Association for Computing Machinery, New York, NY, USA, 21–25. https://doi.org/10.1145/3472720.3483617

[24] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*. 21–25.

[25] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[26] Aman Tyagi and Kathleen M Carley. 2021. Climate Change Conspiracy Theories on Social Media. *arXiv preprint arXiv:2107.03318* (2021).

[27] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nature human behaviour* 4, 5 (2020), 460–471.

[28] Michael J Wood, Karen M Douglas, and Robbie M Sutton. 2012. Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science* 3, 6 (2012), 767–773.

[29] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.

[30] Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. 2021. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication* 38, 1-2 (2021), 98–139.