Understanding and Analyzing COVID-19-related Online Hate Propagation Through Hateful Memes Shared on Twitter

Nishant Vishwamitra*, Keyan Guo†, Song Liao‡, Jaden Mu§,Zheyuan Ma†,
Long Cheng ‡, Ziming Zhao†, Hongxin Hu†
*University of Texas at San Antonio, †University at Buffalo, ‡Clemson University, §East Chapel Hill High School
nishant.vishwamitra@utsa.edu, {keyanguo, zheyuanm, zimingzh, hongxinh}@buffalo.edu,
{liao5, lcheng2}@clemson.edu, jaden.mu@gmail.com

***Abstract*—Recent studies regarding the COVID-19 pandemic have revealed the widespread propagation of hateful content during this period. While significant research has focused on COVID-19-related online hate in text (*e.g.*, text-based tweets), the role of memes in propagating online hate during the pandemic has been largely overlooked. Memes are a popular mechanism used by Internet users to convey their thoughts and opinions on a variety of topics. However, memes have emerged as an important mechanism through which ideologically potent and hateful content spreads on social media platforms. In this work, we focus on investigating the role of memes in the propagation of online hate during the COVID-19 pandemic. We first collect a novel dataset of 4,001 COVID-19-related hateful memes and their replies over a 3-year period from Twitter. Then, we carry out the first large-scale investigation into the impact of these memes on Twitter users, by studying the psychological reactions of Twitter users to these memes using various text analysis methods. We find that COVID-19-related hateful memes have a significantly greater negative impact on Twitter users in comparison to text-based hateful tweets, and increasing negativity towards such memes over the 3-year period. Our new dataset of COVID-19-related hateful memes and findings from our work pave the way for studying the dissemination and moderation of COVID-19-related online hate through the medium of memes.**

Introduction

The harmful effects of the COVID-19 pandemic have produced a range of emotions in the population, including fear, anxiety, and even hostility. Notably, COVID-19-related online hate spread unabated on social media platforms, targeting people based on race/ethnicity, age, social class, immigration status, and political ideology. For instance, Asian Americans were frequent targets of online hate related to COVID-19, with derogatory terms for the disease, such as "kung flu" and "chop fluey", shared more than 10,000 times on Twitter during March 2020 alone [1]. Governments and researchers have both noted the threat posed by hateful content related to COVID-19 on society.

The recent wave of COVID-19-related online hate has engendered studies from various domains [2]. However, these studies have focused on the spread of COVID-19-related online hate through the medium of textual data, such as text-based tweets. *Memes* have emerged as a popular form of expression beyond text. Memes consist of images with superimposed text, that deliver a particular message when considered in the context of both the image and text content together [3], [4]. Traditionally used as devices to induce humor, memes have recently taken a more negative turn, by being used as mediums of spreading online hate [5], [3]. For example, memes that portray Asian people eating dog meat [6], racist memes targeting Chinese eating habits [7], and the morbid meme "Boomer-remover" [8] against older adults have been widely circulated. Since the context of memes is framed by both image and text, online hate propagated in memes is significantly different from text-only hate speech. While recent studies have provided many interesting insights into the nature of online hate in textual data during the COVID-19 pandemic, the role of memes in the propagation of online hate during the COVID-19 pandemic has been largely overlooked. Thus, there is an urgent need to study the spread of COVID-19-related online hate through the medium of memes.

COVID-19-related hateful memes pose unique challenges in studying the propagation of online hate during the pandemic. Recent studies have expounded on the negative psychological effects of COVID-19-related text-based online hate [9], [10] on Internet users. However, the psychological effects of COVID-19-related hateful memes are largely unknown. In this work, we take the first step toward studying the impact of COVID-19-related online hate spread via memes.

We make two important contributions in this work. *First*, we collected a dataset of 8,385 COVID-19-related multimodal memes (i.e., image superimposed with text) overall, with 4,001 memes categorized as hateful and 4,384 memes categorized as non-hateful, collected from over 1,448,112 tweets, along with the memes' created timestamps and 10,843 associated replies over a three-year period (*i.e.*, January 1st,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey
© 2023 Copyright is held by the owner/author(s).
ACM ISBN 979-8-4007-0409-3/23/11
http://doi.org/10.1145/3625007.3630111

Fig. 1: Samples of COVID-19-related hateful memes.

2020 - December 31st, 2020; January 1st, 2021 to December 31st, 2021 and; January 1st, 2022 to November 21st, 2022) to enable several diverse analyses. Our dataset is first-of-its-kind, and unique since it provides a new source of hateful content during the COVID-19 pandemic pertaining to the phenomenon of memes. We will make our dataset publicly available for further research. A few samples from our dataset are depicted in Figure 1, showing hateful memes depicting various harmful aspects of COVID-19, such as anti-Asian hate, mask-related hate, and hate toward older adults (*e.g.* COVID-19 called the "Boomer-remover" [8]). *Second*, we conducted an investigation into the psychological reactions of Twitter users towards COVID-19-related hateful memes using several text analysis methods, by comparing the replies to those memes to the replies to text-based COVID-19-related hateful tweets, and found that users reacted to the memes in a significantly higher negative emotional tone and hostility. Furthermore, linear trend analysis in the users' reactions to COVID-19-related hateful memes shows a longer-lasting effect of harmful emotions towards the COVID-19-related hateful memes over the 3-year period. Our findings in this work shed new light on the nature of online hate through the medium of memes during the COVID-19 pandemic, and also pave the way for future research in the area of COVID-19-related online hate via non-traditional media such as memes.

Related Work

We are interested in studying the role played by hateful memes in the propagation of online hate during the COVID-19 pandemic. Consequently, we are interested in studying the differences in COVID-19-related online hate via memes in comparison to text-based tweets. Several works have discussed traditional online hate based on textual tweets [11], [12], and recent work has also focused on traditional hateful memes [3]. Surprisingly few works [2], [13], [14] have been published to date that focuses on COVID-19-related online hate even in just text-based media, that provide annotated datasets. Our work fills the gap in the literature on online hate propagation via memes during the COVID-19 pandemic, wherein we contribute an annotated dataset of such memes along with replies to these memes, which can facilitate further research in this area.

Data Collection Methodology

We describe our methodology for collecting COVID-19-related hateful memes and replies from Twitter, and our data annotation task via online participants. Our data collection and annotation tasks have been approved by IRB.

Compilation of COVID-19-related Hashtags

To collect memes related to COVID-19, we first needed a set of hashtags to search on Twitter. To this end, we compiled a set of COVID-19-related hashtags, which we then used for collecting memes on Twitter. We started with an initial set of manually compiled hashtags that we found to be prevalent during the COVID-19 pandemic [15]. Then, we used the official Twitter API V2 [1] to collect tweets shared between January 1st, 2020 to November 21st, 2022 based on these initial hashtags. Next, we supplemented our initial set of hashtags with additional hashtags found in those collected tweets. This process enabled us to compile a total of 55 hashtags. Our final list consisted of COVID-19-related hashtags such as *Wuhanvirus*, *ChinaVirus*, *covidiots*, *TakeTheMaskOff*, and *CovidHoax*.

Collection of COVID-19-related Memes

After the compilation of hashtags, we proceeded to use these hashtags to search for memes on Twitter for the 3-year period. We only searched for tweets with image content (i.e., potential memes) based on the list of compiled hashtags. Using the compiled list of COVID-19-related hashtags, we initially collected a total of 1,448,112 potential memes. Next, we used certain criteria to exclude memes that are invalid. First, we restricted our dataset to consist of only those memes that are in English. Since we focus on multimodal memes [3] (i.e., images with superimposed text), we removed memes that did not have any text in them using an open-source tool Tesseract [16]. Then, we excluded those memes that did not have any image-based content (or Regions of Interest) in them (i.e., just plain background images) using the YOLO object detector [17]. We also removed non-static memes (*e.g.* GIFs). We eliminated duplicated memes based on the hash comparison. Finally, we were left with 8,385 valid memes in our dataset.

Memes Annotation

Few publicly available labeled datasets regarding COVID-19-related online hate exist, and consequently, we aimed at first collecting annotations for our COVID-19-related memes dataset. We recruited online participants from AMT to annotate our dataset. We developed an annotation process to establish the ground truth of memes based on the meme's content. We instructed participants to annotate any meme as hateful, that is: (1) directed towards an individual or a group of people, organization, or country, and (2) attacks victims using violent or dehumanizing speech, scandalization of personal appearance propagates harmful stereotypes or misrepresentation, makes statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation [18]. We directed the participants to annotate the memes into the two classes based on our instructions, hateful or non-hateful. To ensure reliable annotation, we only

[1] https://developer.twitter.com/en/docs/twitter-api

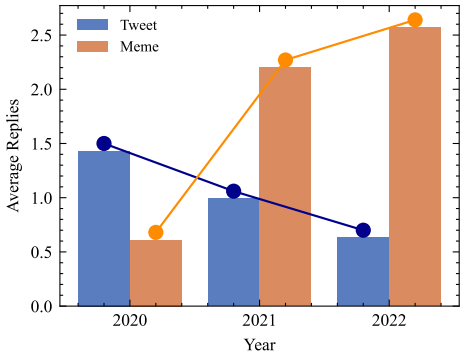

Fig. 2: Average replies to COVID-19-related hateful memes and text-based tweets, respectively.

recruited participants with an approval rating of 90% or higher and 1000 approved HITs to participate in our annotation task. We allowed each meme to be annotated by three distinct participants and chose the majority-voted category as the final annotation. Overall, 4,001 memes were annotated as hateful and 4,384 memes were annotated as non-hateful. The reliability of the agreement with Fleiss' kappa score (0.38 on average) represented fair reliability for human rating [19].

Collection of COVID-19-related Hateful Tweets

In our work, we are interested in studying the propagation of online hate via memes in comparison to the propagation of online hate via traditional text-based tweets. Thus, we proceeded to collect a set of text-based tweets that represented COVID-19-related online hate. We collected a new dataset of COVID-19-related text-based hateful tweets using precisely the same parameters (*i.e.*, hashtags, time frame, and tools) as the collection of memes, along with the needed metadata such as tweet IDs. We collected a total of 49,289 tweets in this data-collection process. Next, we used Google Jigsaw's Perspective API [20] which is trained on Wikipedia moderation verdicts for abuse as well as samples from other online communities [21] to classify the hateful tweets among the full set of collected tweets. Perspective API has been used in multiple existing works [22], [23], [24], [25] to classify social media posts, such as tweets as hateful or not. We considered only those tweets as hateful that received a confidence score of over 0.5 and above from the Perspective API. This process led to the classification of 5,854 as hateful tweets.

Collection of Replies

To capture the reaction of social media users towards both COVID-19-related hateful memes and tweets, we focused on collecting text-based replies made by other users directly to those hateful memes and tweets. Although there are other ways in which users can react to posts on Twitter, such as likes and retweets, we focused on replies since they are text-based and therefore would enable text-based methods for psycho-linguistic analyses, whereas the other options cannot be directly analyzed. We collected direct, text-based replies to those memes and tweets that were annotated or classified as hateful, respectively until November 21st, 2022. We collected a total of 4,118 replies to the hateful memes made by 2,038 users, and 6,725 replies to the hateful tweets made by 4,361 users respectively, until the mentioned date. Figure 2 depicts the average replies to COVID-19-related hateful text-based tweets (*i.e.*, blue bars) and hateful memes (*i.e.*, orange bars), respectively over the 3-year period. Interestingly, although the average replies to hateful tweets progressively decreased over the 3-year period (a decrease of 55.94%), the average replies to hateful memes increased over the same period (an increase of 321.3%). We surmise that this could be due to efforts by social media platforms to flag online hate speech, which however may not be as effective against memes.

Understanding COVID-19-related Hateful Memes

Since our primary objectives were to understand the differences in the nature of online hate that was propagated via memes in comparison to text-based tweets, we posed the following research question:

RQ: Did online hate propagate differently via memes in comparison to text-based tweets during the COVID-19 pandemic?

We conducted two studies to answer this research question. In the first study, we addressed the research question by studying the psycho-linguistic properties [26] such as emotional tone and negative emotions, and the presence of online hate and offensive language [11] of the responses (*i.e.*, the replies) to the memes in comparison to the tweets, respectively. We then explored linear modeling of these properties to investigate any linear trends in those properties.

Analyzing COVID-19-related Hateful Meme Reactions

Studying social media users' reactions to memes containing COVID-19-related hateful content gives a direct indication of the impact of such hateful content on users. Replies to posts are a direct way in which these reactions can be captured. Existing studies have shown that such replies to posts are a reliable way to study users' reactions [27]. In this study, we compared the users' reaction (*i.e.*, replies) to memes that were propagating COVID-19-related hateful content to the reactions to text-based tweets that were also found propagating COVID-19-related hateful content.

We used the psycho-linguistic properties from LIWC [26] for drawing comparisons. We used the latest available LIWC dictionary (*i.e.*, LIWC-22). In addition, we also used HateSonar [11] to measure the hate speech and offensive language content in the replies. HateSonar provides a probability score between 0 and 1 for both hate speech and offensive language respectively, or neither in text excerpts and has been previously used to study these measures in social media text [28]. We used Welch's independent sample t-tests [29] to compare the LIWC dictionary category scores to see if there exist any statistical differences in replies to COVID-19-related hateful memes and hateful text-based tweets, and linear modeling to observe any linear trends over the 3-year time period.

Years	Tone			Neg. Emotion			Anger			Sadness			Swear			HS			OL		
	Meme	Text	T	Meme	Text	T	Meme	Text	T	Meme	Text	T	Meme	Text	T	Meme	Text	T	Meme	Text	T
Overall	28.46	30.67	**-2.27***	1.46	1.42	-.08	.17	.25	**-3.15****	.07	.19	**-3.18****	.71	1.41	**-6.6*****	.06	.06	-1.21	.37	.38	**-2.15***
2020	27.66	30.87	**-1.98***	1.45	1.36	-.27	.21	.26	-.98	.11	.21	-1.67	.83	.67	**-4.29*****	.06	.06	-1.32	.37	.39	**-1.94***
2021	29.11	29.03	.96	1.44	1.5	**-2.01***	.14	.23	**-2.56***	.04	.16	-1.31	.65	.98	**-1.31***	.06	.06	-.41	.37	.37	.79
2022	29	34.47	**-3.09****	1.6	1.65	.41	.11	.27	-.35	.06	.14	-.81	.5	1.07	-1.86	.06	.05	**-4.53*****	.37	.37	1.7

Note. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$

TABLE I: Welch's independent t-test results comparing LIWC and HateSonar properties for replies to COVID-19-related hateful memes and text-based tweets.

Since we were interested in comparing the capability of COVID-19-related memes in inducing emotions of negativity and hostility in users with text-based tweets, we focused on collecting the negative psycho-linguistic properties from the memes and tweets replies. We collected properties closely related to negativity provided by LIWC, *i.e.*, *tone*, *negative emotion*, *anger*, *sadness* and *swear words*. Additionally, we also collected closely related properties via HateSonar, *i.e.*, *hate speech*, and *offensive language*. We compared all these properties for responses to memes and tweets for the overall period between January 1st, 2020 to November 21st, 2022, and also for each of the three years in this period, *i.e.*, 2020, 2021 and 2022. The results of this study are presented in Table I.

Tone. The *tone* property indicates the negativity or positivity of the tone of a text [26], by *taking the full text into context*. A lower *tone* value indicates that the emotion of the text is more negative, and a higher *tone* value indicates that the emotion of the text is more positive. A value below 50 indicates a negative tone. Overall, the replies to COVID-19-related memes had a significantly greater negative *tone* in comparison to the replies to text-based tweets ($t = -2.27$, $p < 0.05$). Especially, in 2020, the hateful memes were significantly greater negative replies than the text-based tweets ($t = -1.98$, $p < 0.05$) and the replies to memes became even more significantly negative in 2022 ($t = -3.09$, $p < 0.01$). More negative overall *tone* of replies to COVID-19-related memes in comparison to text-based tweets shows the difference in the impact of these two media in the propagation of hate during a major crisis. Furthermore, the capability of the memes to receive negative *tone* replies even in 2022 emphasizes the long-lasting effect on the emotions of social media users.

Negative Emotion, Anger and Sadness. The *negative emotion* property indicates the percentage of words in a text that express negative emotions, such as *anger* words (*e.g.*, ugly, nasty, etc) or *sadness* words (*e.g.*, hurt) [26]. In the replies to COVID-19-related hateful text-based tweets, users more directly expressed their emotions by using words that were indicative of negative emotions, compared to memes. Memes involve sarcasm and euphemisms, which do not directly use *negative emotion* words but are overall expressive of negative emotions through the subtle usage of text and image combinations. The replies to memes and text-based tweets also differed based on this difference in the usage of words. In 2021, the text-based tweet replies contained more proportion of *negative emotion* words ($t = -2.01$, $p < 0.05$) than memes. Overall, the text-based tweet replies also contained more percentage of *anger* words ($t = -3.15$, $p < 0.01$) and *sadness* words ($t = -3.18$, $p < 0.05$). The non-usage of direct words in meme replies poses a greater challenge to their moderation since existing dictionary-based moderation techniques are limited against subtle word usage.

Swear Words. The *swear words* property indicates the percentage of swear words (such as d*mn, f**k, and p*ss) in a text from a swear words dictionary [26]. Overall, the replies to text-based COVID-19 hateful tweets contained more proportion of swear words ($t = -6.6$, $p < 0.001$) compared to meme replies, wherein the replies to text-based tweets contained the direct usage of such words. In text-based tweet replies, on average every reply consisted of at least one swear word (overall: 1.41 ($t = -6.6$, $p < 0.001$), 2021: .98 ($t = -1.31$, $p < 0.05$), 2022: 1.07 ($t = -1.86$, $n.s.$)).

Hate Speech and Offensive Language. Finally, we also studied two properties of online hate from HateSonar [11]. This work proposed a machine learning-based approach to computing a score for hate speech, if a text just contains offensive language but no hate speech, or if there is neither hate speech nor offensive language. In the case of hate speech, COVID-19-related hateful memes got a more hateful reaction in 2022 compared to the reaction to text-based tweets ($t = 4.53$, $p < 0.001$), which could imply that memes induce a long-lasting effect on social media users. In the case of offensive language, observations were similar to the *swear words* property, wherein overall, replies to text-based tweets contained more offensive language ($t = -2.15$, $p < 0.05$).

Furthermore, we performed linear modeling of all these properties for memes, to find out if there are any linear trends in the 3-year time period. Table II summarizes the results of linear modeling. An increasing linear trend in *anxiety*

Property	R^2	β	p
Positive Emotion	.001	-.038	<.05
Anxiety	.001	.037	<.05
Anger	.002	-.04	<.01
Sadness	.002	-.05	<.001
Swear	.001	-.031	<.05

TABLE II: Linear model results for psycho-linguistic properties regarding meme replies over the three-year period.

($R^2 = 0.001$, $\beta = 0.037$, $p < .05$) for replies to memes was observed, indicating an extended stage of perception of COVID-19, wherein users depicted anxiety about uncertain conditions during the pandemic. Moreover, other properties such as *anger*, *sadness*, and *swear* showed extremely gradual linear decreasing trends (*e.g.*, $\beta = -.04$ for *anger*), indicating longer lasting harmful effects of COVID-19-related hateful memes.

Findings Regarding RQ: Reaction to Memes Vs. Tweets. To summarize the answer to RQ, COVID-19-related hateful memes elicited more strongly negative reactions than the replies to text-based tweets. Social media users reacted to memes in a more overall negative *tone*, although text-based tweet replies used more *negative emotion* words and *swear* words. Such memes also seemed to be a sustained, long-lasting problem. For example, in 2022 memes replies were both higher *negative emotion* and more *hateful*. Additionally, gradually decreasing linear trends of properties such as *anger*, *sadness*, and *swear* in meme replies were also found. Thus, the role of memes in the propagation of online hate during COVID-19 is significantly different from text-based tweets, and more studies about the damaging role played by memes should be investigated.

Conclusion

In this work, we studied the propagation of COVID-19-related online hate via memes. We focused on studying the reactions of social media users to these memes in comparison to tweets. We collected and annotated a large dataset of COVID-19-related memes from Twitter along with replies to the memes. Our analysis of psycho-linguistic responses towards COVID-19-related hateful memes and tweets shows that memes impact users more negatively than tweets. Linear modeling of the negative properties indicates a longer-lasting harmful effect of memes.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant No. 2245983, 2129164, 2114982, 2228617, 2120369, and 2237238.

References

[1] Macguire, E., "Macguire, E., Anti-Asian Hate Continues to Spread Online Amid COVID-19 Pandemic, in Al-Jazeera," https://www.aljazeera.com/news/2020/04/anti-asian-hate-continues-spread-online-covid-19-pandemic\-200405063015286.html, 2020.

[2] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis," *arXiv preprint arXiv:2005.12423*, 2020.

[3] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *arXiv preprint arXiv:2005.04790*, 2020.

[4] Y. Du, M. A. Masood, and K. Joseph, "Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 153–164.

[5] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil, "On the origins of memes by means of fringe web communities," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 188–202.

[6] Davey Alba, "How anti-Asian activity online set the stage for real-world violence," https://www.independent.co.uk/news/world/americas/anti-asian-online-hate-crime-real-world-b1820031.html, 2021.

[7] James Palmer, "Don't Blame Bat Soup for the Coronavirus," https://foreignpolicy.com/2020/01/27/coronavirus-covid19-dont-blame-bat-soup-for-the-virus/, 2020.

[8] Hannah Sparks, "Morbid 'boomer remover' coronavirus meme only makes millennials seem more awful," https://nypost.com/2020/03/19/morbid-boomer-remover-coronavirus-meme-only-makes-millennials-\seem-more-awful/, 2020.

[9] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, pp. 1–22, 2020.

[10] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "Ethos: an online hate speech detection dataset," *arXiv preprint arXiv:2006.08328*, 2020.

[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Eleventh international aaai conference on web and social media*, 2017.

[12] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[13] X. Lian, "Speech detection task against asian hate: Bert the central, while data-centric studies the crucial," *arXiv preprint arXiv:2206.02114*, 2022.

[14] L. Fan, H. Yu, and Z. Yin, "Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, p. e313, 2020.

[15] Financial Express, "Coronavirus outbreak: What is "covidiots" trending on twitter?" https://www.financialexpress.com/lifestyle/coronavirus-outbreak-what-is-covidiots-trending-on-twitter/1907432/, 2020.

[16] A. Kay, "Tesseract: an open-source optical character recognition engine," *Linux Journal*, vol. 2007, no. 159, p. 2, 2007.

[17] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[18] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[19] R. Falotico and P. Quatto, "Fleiss' kappa statistic without paradoxes," *Quality & Quantity*, vol. 49, no. 2, pp. 463–470, 2015.

[20] Google Perspective, "Perspective API," https://www.perspectiveapi.com/#/home, 2020.

[21] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1391–1399.

[22] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," 2018.

[23] M. Saveski, B. Roy, and D. Roy, "The structure of toxic conversations on twitter," in *Proceedings of the Web Conference 2021*, 2021, pp. 1086–1097.

[24] Y. Hua, M. Naaman, and T. Ristenpart, "Characterizing twitter users who engage in adversarial interactions against political candidates," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[25] A. Rajadesingan, P. Resnick, and C. Budak, "Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 557–568.

[26] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[27] S. Yardi and D. Boyd, "Dynamic debates: An analysis of group polarization over time on twitter," *Bulletin of science, technology & society*, vol. 30, no. 5, pp. 316–327, 2010.

[28] T. Grover and G. Mark, "Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 2019, pp. 193–204.

[29] M. Delacre, D. Lakens, and C. Leys, "Why psychologists should by default use welch's t-test instead of student's t-test," *International Review of Social Psychology*, vol. 30, no. 1, 2017.