2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

Quantifying the Intensity of Toxicity for Discussions and Speakers

Samiha Samrose
University of Rochester
ssamrose@cs.rochester.edu

Ehsan Hoque
University of Rochester
mehoque@cs.rochester.edu

***Abstract*—In this work, from YouTube News-show multimodal dataset with dyadic speakers having heated discussions, we analyze the toxicity through audio-visual signals. Firstly, as different speakers may contribute differently towards the toxicity, we propose a speaker-wise toxicity score revealing individual proportionate contribution. As discussions with disagreements may reflect some signals of toxicity, in order to identify discussions needing more attention we categorize discussions into binary high-low toxicity levels. By analyzing visual features, we show that the levels correlate with facial expressions as Upper Lid Raiser (associated with 'surprise'), Dimpler (associated with 'contempt'), and Lip Corner Depressor (associated with 'disgust') remain statistically significant in separating high-low intensities of disrespect. Secondly, we investigate the impact of audio-based features such as pitch and intensity that can significantly elicit disrespect, and utilize the signals in classifying disrespect and non-disrespect samples by applying logistic regression model achieving 79.86% accuracy. Our findings shed light on the potential of utilizing audio-visual signals in adding important context towards understanding toxic discussions.**

***Index Terms*—toxicity, discussion, audio-video analysis, classification**

I. INTRODUCTION

Conversation involving a difference in opinions bears the potential to turn into a disrespectful interaction [1], [2]. In recent times when socio-political divide has increased among people with polarized viewpoints [3], understanding such interaction is crucial to prevent further toxicity. For example, in the first US presidential debate in October of 2020, significant number of interruptions or overlapping speeches barred the flow of the debate[1] (e.g., 90 interruptions in a 90-minute debate in which one speaker contributed to 72 of them). Such discussion dynamics are also prevalent among others who experience conflicting interactions in day-to-day life [4].

Understanding and identifying such problematic behaviors are important towards applying moderation for a conversation [5], [6]. Toxic, abusive, or disrespectful conversation has mostly been studied from textual aspect [7]–[10]. Besides real-life face-to-face interactions, increased usage of video and audio-based mediums (e.g., video-conferencing, podcasts) for communication has intensified the need for and the opportunity of machine understanding of toxicity through audio-visual signals alongside linguistic properties [11].

This work was supported by National Science Foundation Award IIS-1750380.

[1]https://slate.com/news-and-politics/2020/09/trump-interruptions-first-presidential-debate-biden.html

Fig. 1. An example frame from the dataset. All conversations in the dataset are dyadic news channel teleconferences presented in a split-screen format.

In this paper, using a multimodal dataset, we explore categorizing the intensity level of toxicity of a group discussion, defining speaker-wise toxicity score, and revealing the potential of audio-visual signals in understanding conversational toxicity. Previous work has explored classification of disrespect through visual cues [12]. However, it does not address intensity of toxicity or speaker-wise contribution to it, or audio features in disrespect identification. We establish the scoring mechanism through the analysis of visual features of the discussion data, and statistically analyze audio features as well as build binary classification models for it.

Our dataset is collected from News channel videos from YouTube[2]. The videos involve dyadic conversations between a news host and a guest connected through teleconferencing. Such telecast follows a standard template with the screens split in half showing host and guest on each side of the split. Figure 1 shows the setup of such a conversation. To better understand the speakers' performances in a toxic discussion session, we provide a speaker-wise toxicity contribution score making the proportionate contribution transparent to the audience and the analysis. To understand a toxic session better, we construct an intensity analysis of toxicity as well as present an association between this intensity and facial expressions. We also show the potential of vocal features in understanding toxicity, and explore how the audio information can be used to classify such instances. Overall, the audio-visual analysis on our multimodal dataset reveals the potential of better understanding conversational toxicity.

[2]https://www.youtube.com/yt/about/press

978-1-6654-0021-3/21/$31.00 ©2021 IEEE

II. Related Work

Prior work has found that maintaining respectful dynamics is important while exchanging conflicting ideas [13], [14]. Respectful interactions create a safe space to exchange disagreements, whereas disrespectful behaviors can derail the conversation [8] and even destroy the team structure [14]. Mansbridge et al. [14] suggest that actions such as actively listening as well as speaking in a polite manner to disseminate the reasoning can establish the mutual respect. On the other hand, shouting and interruption can project dominance leading to perceived disrespect [15], [16]. Based on the inter-personal relation among speakers and the prospect of them as a group, the intensity of toxicity can also vary [17], [18].

Disrespect or unjustified disruption in group discussion has been explored in various settings (e.g., face-to-face [19], online [8]). Various NLP approaches [7], [20]–[24] have analyzed the toxicity of conversation by using linguistic properties. For example, detecting toxicity from Wikipedia comments [25], identifying impoliteness in the language of YouTube comments [26], etc. Intensity and expressivity have been found resourceful in understanding affect [27]–[30], which can also be extended for toxicity analysis. Even though audio-video signals bear inherent information rrelated to toxicity projection, such signals have not been extensively used in understanding conversational toxicity. Recently Samrose et al. [12] have explored visual cues in capturing such toxicity. However, the information elated to speaker-wise performance and audio prospects remain under-explored.

III. Definitions

- **Disrespect/Toxicity:** We define interpersonal disrespect during discussions as conversational toxicity inflicted towards another speaker involved in a discussion. In this paper, we use 'disrespect' and 'toxicity' interchangeably.
- **Toxicity Score for Speaker:** To better understand how each speaker performs in each discussion, we assign a score (i.e., percentage) reflecting the proportionate toxicity contribution during that discussion. Notably, in different discussion sessions, a speaker can have different scores which can be used to compare performances across sessions.
- **Intensity Level for Discussion:** To better understand each discussion, we assign a binary class (i.e., high and low) reflecting its intensity level. General disagreements may still reflect some form of *disrespect* signals, therefore this *Discussion Intensity Level* help identify those highly toxic discussions needing more attention.

Capturing both discussion/session intensity (high-low) and speaker score (percentage) are important as these provide crucial context. For example, just mentioning *"Speaker-A was in a highly toxic discussion"* can disproportionately affect this speaker who might have stayed respectful in a highly toxic session. Likewise, *"A session had a low intensity of disrespect"* reveals that the session was not very toxic to begin with, so the speaker score might not be as intense. As a unit, the discussion intensity level and the speaker contribution score provide better context for evaluation.

TABLE I
METADATA PER VIDEO

Variable	Value	Description
Start	$0 \leq$ timestamp $\leq$ video Duration	The timestamp when a member initiates an act of disrespect
End	start $<$ timestamp $\leq$ video Duration	The timestamp when the ongoing act of disrespect terminates
Responsible	Host or Guest or Both	The member(s) who is responsible for inflicting the disrespectful incident
Modality	Face-gesture and/or Voice and/or Language	What modality(s) of the member contributing to the disrespectful act

IV. Dataset

A. Discussion Setup: We target naturalistic professional discussions happening in the wild, and thus collect YouTube News Channel videos primarily focusing on two news channels - Fox News and CNN. We also limit the conversation to be happening only in a dyadic setting to reduce any ambiguity towards whom a *disrespectful* act is inflicted.

B. Sample Collection: We include a three-stage review to make sure the selected videos in fact have *disrespect* markers on them: (1) While crawling videos from YouTube, the primary search was done with relevant keywords:{*"heated+disagreement+debate+discussion+news"*}; (2) Then a researcher went through individual videos to roughly assess whether the videos had heated discussions; (3) Finally, each video was labeled by three trained annotators, each of whom individually watched the videos and provided labels with metadata for each video.

C. Annotation Guideline: If annotation guidelines are not carefully prepared and annotators are not properly trained, developing such datasets can incorporate biases. For example, based on cultural norms, an older adult interrupting a younger person might be perceived differently. Therefore, to minimize bias, instead of crowd-sourcing-based labels, we prepared annotation guidelines based on related literature and trained the annotators [31]–[33]. The constraints included: (1) consideration of disrespect towards each other (speakers), not towards the discussion topic; (2) assumption that both speakers have the same and the highest level of self-esteem; (3) exclusion of demography or rank-based disrespect. Once all three sets of metadata are collected, those intersecting clip regions where two or more raters agreed on the *disrespect* label are trimmed and extracted. The detailed metadata is included in Table I. Notably, it holds which modality (i.e., visual, audial, linguistic) contributed to each *disrespectful* act.

V. Intensity of Toxicity

A. Speaker-wise Toxicity Score: Not all speakers in a toxic discussion may contribute equally towards expressing disrespectful behaviors. Identifying a discussion as toxic but assigning that label for all involved speakers is unfair and misleading information, since even in a highly toxic discussion one speaker may stay respectful while the other may escalate the situation. Therefore, it is important to assign a speaker-wise toxicity score for individual speakers in any toxic discussion. Having that score not only adds more insights into the particular discussion but also helps keep track of a speaker's performance across multiple sessions over time. We propose a $DisrespectContributionScore(DCscore)$ to evaluate the speaker-wise performance in each discussion session.

Figure 2 shows an example consistent with our dataset in which two speakers (s_1 and s_2) are having a heated disagreement. s_1 and s_2 individually inflict x and y counts of *disrespectful* acts, respectively. The duration of each of the acts is denoted by d_k. Notably, the total duration of disrespect is not necessarily the summation of the individual speaker's disrespect duration, as the instances can overlap (e.g., interruption). Fig 2 depicts the way overlapping zones can be distributed among speakers. A speaker's toxicity score is independent of another speaker. This means that both speakers can choose to have overlapping zones with *disrespectful* acts throughout the video, and thus each can gain a score of 100%. For i number of speakers in a discussion and m being number of disrespect instances for a speaker:

$$\textit{Toxicity duration per speaker:} \qquad d_{s_i} = \sum_{j=1}^{m_i} d_j$$

$$\textit{Toxicity duration in discussion:} \qquad d = \sum_{j=1}^{n} d_j$$

$$\textit{Properties:} \qquad d_{s_i} \leq d \qquad \textit{and} \qquad \sum d_{s_i} \geq d$$

$$\textit{Speaker-wise toxicity score:} \qquad DCscore_{s_i} = \frac{d_{s_i}}{d} * 100$$

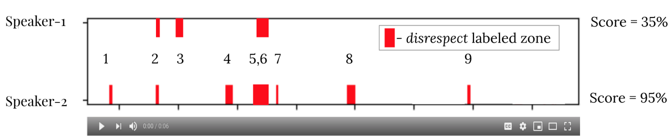

Fig. 2. Toxicity contribution score for two speakers in a session.

B. Discussion Intensity Level: Once we have the DC_{score} for each speaker in a session, we use the individual speaker clips to explore whether toxicity intensity varies with facial expressions. The intensity can vary from discussion session to session. We hypothesize that the facial expressions in the videos can also vary showing a correlation with the intensity metric. First, we assign *Low Intensity* level to an individual speaker's clip if the DC_{score} is less than 50%, otherwise *High Intensity*. Then we measure the intensity of facial expressions within the clips. We extract the facial action unit intensity scores and take the average of the different AU intensities over all frames corresponding to each speaker in each video. For each AU, we calculate the median of these average values over all the videos in the dataset. We label a video to have a *high intensity* for an AU, if the average AU intensity in that video is higher than the median value for that AU across all

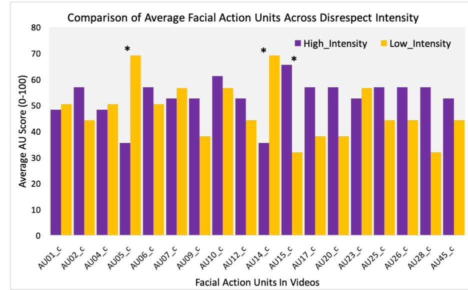

Fig. 3. Comparison of average AU score between *High* and *Low* intensity of disrespect clips for 18 AUs. Asterisks denote a statistically significant difference under the Mann-Whitney U test. These AUs are AU05 (Upper Lid Raiser), AU14 (Dimpler), and AU15 (Lip Corner Depressor).

videos. Otherwise, the video gets a *low intensity* for that AU. Now we compare all the AU intensity scores between these two groups we formed: *High intensity of disrespect* vs *Low intensity of disrespect*. Even though we specify the importance of having binary levels of toxicity, it is possible to compute a toxicity spectrum and place a discussion on it to capture the granularity, if necessary, by comparing the standard deviation of the clip's average feature score from all clips' average scores can be computed. Also, even though we show it for the action unit (AU), this process can be for any other feature from any modality. The binary level assignment steps are below:

1) Compute the average score (or average estimated intensity) of each feature f within a clip c having x frames (e.g., average facial action unit feature AU01 in clip-1 having 100 frames):

$$\overline{EI}_c^{AU^f} = \frac{1}{x} \sum_{i=1}^{x} AU_i^f$$

2) Compute the median of the average scores of each feature for all clips:

$$\widetilde{EI}_{all}^{AU^f}$$

3) Compare the average feature score of a clip with the median feature score and assign intensity level:

$$EI_{level}(c) = \begin{cases} high, & \text{if } \overline{EI}_c^{AU^f} > \widetilde{EI}_{all}^{AU^f} \\ low, & \text{otherwise} \end{cases}$$

C. Facial Feature Extraction: We analyze the facial expressions by using OpenFace, which provides the intensity of 18 facial Action Units (AU) based on the Facial Action Coding System (FACS) [34]. To do the extraction per speaker, we mask the video to have only the host or guest visible, and then extract the AU scores. Masking is done as OpenFace performs better on single- rather than multi-face videos. The AU scores are extracted from the video with 15 fps. The boolean values of the corresponding features are extracted for our dataset.

D. Results: Figure 3 shows the average AU score comparison between these two groups. We apply the Mann-Whitney U test showing a statistically significant difference ($\alpha < 0.05$) for AU05 (Upper Lid Raiser), AU14 (Dimpler), and AU15 (Lip Corner Depressor). Our intuition is that AU05 (Upper

Lid Raiser), which is associated with “surprise”, is more prominent in *non-disrespect* samples expressing genuine interest in the received information. AU14 (Dimpler) is higher in *non-disrespect* samples as speakers may be projecting comparatively more smiles during their conversation. AU15 (Lip Corner Depressor) is higher in *disrespectful* samples as this signal corresponds to negative emotion such as disgust. This suggests that facial expressions can differentiate between low and high *disrespect* intensity.

VI. Identifying *Disrespect* vs *Non-Disrespect*

A. Sample Extraction: To explore the power of audio signals in identifying *disrespectful* acts, we extract audio-based *disrespect vs non-disrespect* samples. First, we identify the intersecting zones in which two or more annotators marked having *disrespect*. Next, if any annotator included audio as a relevant modality for that zone, then we include that sample in the audio-based samples. We find 38 videos containing audio-based *disrespect* instances, from which we extracted 226 clips for our audio-based sub-dataset. To generate audio-based *non-disrespect* samples, we consider the zones which no rater labeled as *disrespectful*, and thus collect 176 samples by enforcing that the total duration of the *disrespect*ful audio samples for a particular video matches that of the *non-disrespect* audio samples collected from that video. This ensures the samples remain balanced in terms of speaker and discussion.

B. Audio Feature Extraction: We use Praat [35], an open-source audio processing software, to extract audio features related to amplitude, intensity, pitch, harmonicity, jitter (localJitter, localabsoluteJitter, rapJitter, ppq5Jitter, ddpJitter), shimmer (localShimmer, localdbShimmer, apq3Shimmer, apq5Shimmer, apq11Shimmer, ddaShimmer). We compute the average feature values per clip for further analysis.

C. Pattern Difference Analysis: To compare the audio characteristics and understand whether the patterns are different within audio-based *disrespect* and *non-disrespect* samples, we apply the Mann-Whitney U test [36] on the base features: *pitch, intensity, amplitude, harmonicity*. We find that pitch ($p < 0.001$), intensity ($p < 0.001$), harmonicity ($p < 0.01$) can differentiate between the two classes. Fig 4 shows the boxplot per feature.

D. Classification: As there is a statistically significant difference in audio characteristics of *disrespect* and *non-disrespect* classes, next we investigate a classification approach. We incorporate all the extracted audio features for this analysis. We build a logistic regression model with 5-fold cross-validation with random split, and ran it for 30 epochs. These days logistic regression is being widely used in signal processing [37], [38]. Fig 5 shows the ROC curve for one such epoch. The audio-based model achieves 79.86% accuracy with 84.08% recall (Table II). This reveals that even from a single modality perspective, audio features provide a rich context in a discussion. By investigative the observational characteristics, we find that the signals can be under 3 major groups: interruption, raised voice or shouting, other (e.g., disapproval or satire tone).

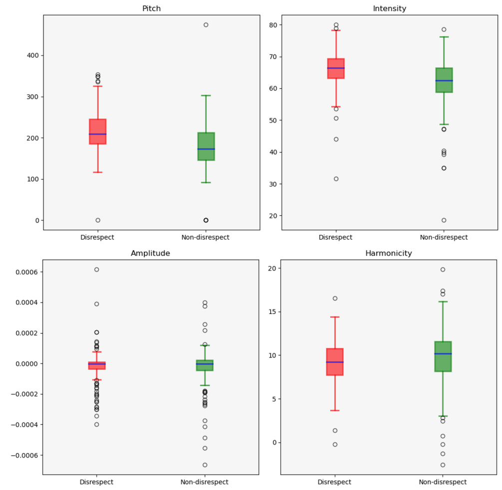

Fig. 4. Comparison of base audio features between *disrespect* and *non-disrespect* samples. Mann-Whitney U test shows that pitch ($p < 0.001$), intensity ($p < 0.001$), harmonicity ($p < 0.01$) are statistically significant.

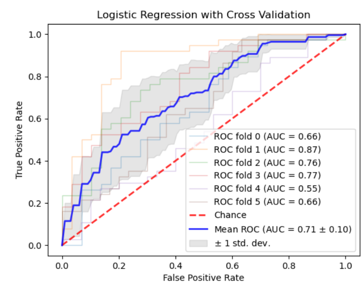

Fig. 5. ROC curve for Logistic Regression Model

TABLE II
Classification Performance: Our audio-based model compared to the video-based model of [12]

Model	Accuracy	Precision	Recall	F1-score
$Log.Reg_{audio}$	**79.86**	0.81	0.84	0.82
$Log.Reg_{video}$	62.61	0.65	0.63	0.64

VII. Discussion & Conclusion

On a multimodal dataset, we showcase that audio and video modalities can be crucial in revealing the signals of conversational toxicity. Such explorations are important, as conversational toxicity is mostly explored through linguistic signals. Collecting more data can enable better analyses with deep learning models and speaker/video out based validations. Our exploration opens up opportunities for audio-visual signals to be incorporated in understanding, and eventually mitigating, toxic discussions. With cautious and mindful incorporation, the applications can be adapted for conducting better classroom or professional meetings. For public discussions, such audio-visual analysis with the corresponding scores can provide better context to the audience. For private discussions, that information can be kept to the individual speakers and used as self-reflection-based feedback to improve ways in which people handle heated discussions.

ACKNOWLEDGMENT

The authors are thankful to Zhen Bai, Amanda Stent, Mary Czerwinski, and ACII reviewers for their invaluable feedback during various stages of the work.

REFERENCES

[1] H. Mercier and H. Landemore, "Reasoning is for arguing: Understanding the successes and failures of deliberation," *Political psychology*, vol. 33, no. 2, pp. 243–258, 2012.

[2] J. Mansbridge, J. Bohman, S. Chambers, D. Estlund, A. Føllesdal, A. Fung, C. Lafont, B. Manin, and J. L. Martí, "The place of self-interest and the role of power in deliberative democracy," *Journal of political philosophy*, vol. 18, no. 1, pp. 64–100, 2010.

[3] L. Boxell, M. Gentzkow, and J. M. Shapiro, "Greater internet use is not associated with faster growth in political polarization among us demographic groups," *Proceedings of the National Academy of Sciences*, vol. 114, no. 40, pp. 10 612–10 617, 2017.

[4] A. Rychwalska and M. Roszczyńska-Kurasińska, "Polarization on social media: when group dynamics leads to societal divides," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[5] J. Costa, M. F. Jung, M. Czerwinski, F. Guimbretière, T. Le, and T. Choudhury, "Regulating feelings during interpersonal conflicts by changing voice self-perception," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 631.

[6] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "Affectaura: an intelligent system for emotional memory," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 849–858.

[7] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper attention to abusive user content moderation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1125–1135.

[8] J. Zhang, J. P. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, N. Thain, and D. Taraborelli, "Conversations gone awry: Detecting early signs of conversational failure," *arXiv preprint arXiv:1805.05345*, 2018.

[9] B. van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for toxic comment classification: An in-depth error analysis," *arXiv preprint arXiv:1809.07572*, 2018.

[10] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, 2012, pp. 19–26.

[11] P. J. Moor, A. Heuvelman, and R. Verleur, "Flaming on youtube," *Computers in human behavior*, vol. 26, no. 6, pp. 1536–1546, 2010.

[12] S. Samrose, W. Chu, C. He, Y. Gao, S. S. Shahrin, Z. Bai, and M. E. Hoque, "Visual cues for disrespectful conversation analysis," in *8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, Cambridge, United Kingdom, September 3-6, 2019*. IEEE, 2019, pp. 580–586. [Online]. Available: https://doi.org/10.1109/ACII.2019.8925440

[13] R. C. Maia and T. A. S. Rezende, "Respect and disrespect in deliberation across the networked media environment: Examining multiple paths of political talk," *Journal of Computer-Mediated Communication*, vol. 21, no. 2, 2016.

[14] J. Mansbridge, J. Bohman, S. Chambers, D. Estlund, A. Follesdal, A. Fung, C. Lafont, B. Manin, and J. l. Marti, "The place of self-interest and the role of power in deliberative democracy," *Journal of Political Philosophy*, vol. 18, no. 1, 2009.

[15] N. E. Dunbar and J. K. Burgoon, "Perceptions of power and interactional dominance in interpersonal relationships," *Journal of Social and Personal Relationships*, 2005.

[16] T. A. Lamb, "Nonverbal and paraverbal control in dyads and triads: Sex or power differences?" *Social Psychology Quarterly*, pp. 49–53, 1981.

[17] T. Tyler and S. Blader, *Cooperation in groups: Procedural justice, social identity, and behavioral engagement*. Routledge, 2013.

[18] D. De Cremer, "Respect and cooperation in social dilemmas: The importance of feeling included," *Personality and Social Psychology Bulletin*, vol. 28, no. 10, pp. 1335–1341, 2002.

[19] M. Friedman, "The so-called high-conflict couple: A closer look," *The American Journal of Family Therapy*, vol. 32, no. 2, pp. 101–117, 2004.

[20] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. ACM, 2017, pp. 1217–1230.

[21] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?" *Information Processing & Management*, vol. 58, no. 3, p. 102524, 2021.

[22] R. Beniwal and A. Maurya, "Toxic comment classification using hybrid deep learning model," in *Sustainable Communication Networks and Application*. Springer, 2021, pp. 461–473.

[23] Y. Xia, H. Zhu, T. Lu, P. Zhang, and N. Gu, "Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–23, 2020.

[24] H. Almerekhi, H. Kwak, J. Salminen, and B. J. Jansen, "Are these comments triggering? predicting triggers of toxicity in online discussions," in *Proceedings of The Web Conference 2020*, 2020, pp. 3033–3040.

[25] 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*. International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.

[26] N. Lorenzo-Dus, P. G.-C. Blitvich, and P. Bou-Franch, "On-line polylogues and impoliteness: The case of postings sent in response to the obama reggaeton youtube video," *Journal of Pragmatics*, vol. 43, no. 10, pp. 2578–2593, 2011.

[27] M. T. Uddin and S. Canavan, "Quantified facial temporal-expressiveness dynamics for affect analysis," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3955–3962.

[28] S. Lei, K. Stefanov, and J. Gratch, "Emotion or expressivity? an automated analysis of nonverbal perception in a social dilemma," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 544–551.

[29] V. Lin, J. M. Girard, M. A. Sayette, and L.-P. Morency, "Toward multimodal modeling of emotional expressiveness," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 548–557.

[30] V. Lin, J. M. Girard, and L.-P. Morency, "Context-dependent models for predicting and characterizing facial expressiveness," *arXiv preprint arXiv:1912.04523*, 2019.

[31] D. T. Miller, "Disrespect and the experience of injustice," *Annual review of psychology*, vol. 52, no. 1, pp. 527–553, 2001.

[32] J. Blanchard and N. Lurie, "Respect: Patient reports of disrespect in the healthcare setting and its impact on care," *Journal of Family Practice*, vol. 53, no. 9, pp. 721–731, 2004.

[33] H. J. Smith, T. R. Tyler, and Y. J. Huo, "Interpersonal treatment, social identity, and organizational behavior," *Social identity at work: Developing theory for organizational practice*, pp. 155–171, 2003.

[34] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[35] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[36] J. S. Milton and J. C. Arnold, *Probability and Statistics in the Engineering and Computing Sciences*. McGraw-Hill Higher Education, 1986.

[37] Z. Deng, A. Kammoun, and C. Thrampoulidis, "A model of double descent for high-dimensional logistic regression," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 4267–4271. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053524

[38] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14 516–14 525, 2019.