



Predicting Emotional Intensity in Political Debates via Non-verbal Signals

Jeewoo Yoon^{1,2}, Jinyoung Han^{1,2,*}, Erik Bucy³, Jungseock Joo^{4,*}

¹Sungkyunkwan University, Seoul, Korea

²RAONDATA, Seoul, Korea

³Texas Tech University, Texas, USA

⁴University of California, Los Angeles, USA

{yoonjeewoo, jinyounghan}@skku.edu, erik.bucy@gmail.com, jjoo@comm.ucla.edu

Abstract

Non-verbal expressions of politicians are important in election. In particular, the emotional intensity of politician revealed in a debate can be strongly linked to voters' evaluation. This paper proposes a multimodal deep-learning model for predicting the perceived emotional intensity of a candidate, which utilizes voice, face, and gesture to capture the comprehensive information of one's emotional intensity revealed in a debate. We collect a dataset of political debate videos from the 2020 Democratic presidential primaries in the USA, and train the proposed model with randomly sampled clips from the debate videos. By applying the proposed model to 23 candidates in 11 debate videos, we show that the standard deviation of the perceived emotional intensity is positively correlated with the changes in candidates' favorability in public polls.

1. Introduction

Non-verbal expressions of politicians are important in election. Scholars have shown that non-verbal signals such as facial expressions, eye blink, and vocal pitch can affect voters' attitudes and evaluations on political leaders [1, 2, 3, 4, 5, 6, 7, 8]. For instance, Sullivan *et al* [3] showed that both American and French voters respond positively to the positive emotional displays of political leaders. Boussalis and Coan [4] found that an angry face can positively affect voters' attitudes toward politicians in televised debates.

The emotional intensity of politicians revealed in videos, e.g., debates or news, is an important factor that can affect the decision making process of voters [6, 7, 8, 9]. For example, Klofstad *et al* [8] showed that the voters prefer female politicians with lower pitched voices. Dietrich *et al* found that the Congresswomen who show high emotional intensity while talking about 'women issues' receive positive evaluations from women [9].

While prior work mostly used the vocal pitch of politicians in measuring the emotional intensity for understanding the voting behavior, however, relatively little attention has been paid to utilize multiple modalities such as facial expression, gesture, and vocal expression in modeling the emotional intensity. Instead of simply calculating the pitch of a person's voice to measure the intensity, we consider diverse types of emotional intensity by capturing multiple factors including facial expression [10], body movements [11], etc. In this way, our work goes one step further by comprehensively quantifying and measuring the *perceived emotional intensity* with a computational approach and investigating how it can affect voting behavior.

* Corresponding authors.

To this end, we propose a multimodal deep-learning model that takes non-verbal features as inputs and predicts the perceived emotional intensity of a candidate revealed in a video clip. Inspired by prior work [10, 11, 9], we focus on non-verbal features such as vocal, facial, and gestural features to train the proposed model. As a case study, we collected a dataset of political debate videos of the 2020 Democratic presidential primaries in the USA. Note that the televised political debates can be representative data for capturing and analyzing the non-verbal behaviors of politicians. We then recruit annotators to label the perceived emotional intensity of each candidate who is shown in a random subset of short clips, which are used to train the proposed model. By applying the proposed model to 23 candidates in 11 debate videos, we show that the standard deviation of the perceived emotional intensity is positively related with the changes in candidates' favorability in public polls.

2. Debate Data

In this section, we introduce a debate data used in our study. We first describe the collected televised debate videos, and then introduce (i) how to annotate emotional intensity of the collected videos, and (ii) how to extract non-verbal features.

2.1. Data Collection

We collected videos for each of the 11 debates from the source of the televised debate. Each debate is sponsored by a news channel or a TV station such as NBC, ABC, and CBS, who made high-quality videos available for download. We use the maximum quality available (generally 720p or higher) for better fine-grained feature extraction. The total length of the collected videos is 34 hours, 25 minutes, and 58 seconds with each debate lasting 2-2.5 hours on average. Note that we use the timestamped transcripts of each debate with speaker information.

2.2. Data Annotation

To build a training set for the emotional intensity prediction task, we first randomly sampled 40 short video clips (i.e., 5 secs long) for each candidate. We then recruited eight college students to annotate the emotional intensity score for the given video clips. The annotators are asked to rate the emotional intensity score for the given short video clips on a scale of 1 to 5 (1 being the lowest and 5 being the highest). As a result, we obtain 920 short video clips with emotional intensity scores (i.e., 40 video clips \times 23 candidates). The mean and standard deviation of the annotated (perceived) emotional intensity score were 0.55 (95% C.I. 0.54 - 0.56) and 0.17, respectively. The inter-rater reliability for emotional intensity annotation was ex-

cellent, with an ICC (2, 8) of 0.85 (95% C.I. 0.81 - 0.87).

Figure 1 shows the average annotated intensity score of 23 candidates. As illustrated in Figure 1, Kirsten Gillibrand shows the highest emotional intensity, whereas Michael Bloomberg shows the lowest. Note that we rescaled the emotional intensity score to 0–1.

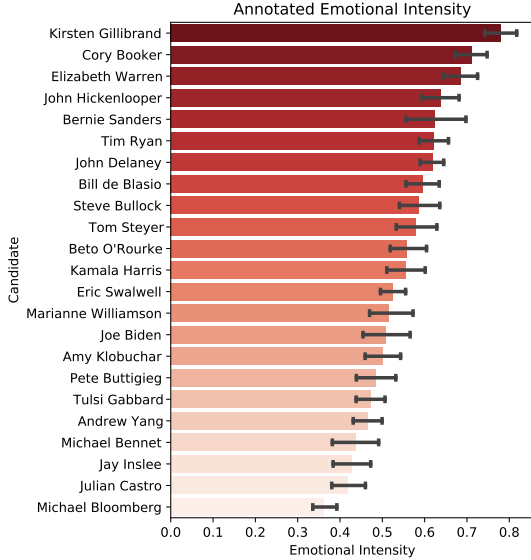


Figure 1: Annotated emotional intensity score (with 95% C.I.) per candidate.

2.3. Non-verbal Feature Extraction

2.3.1. Vocal Features

To extract vocal features from a given video clip, we employ Librosa, a well-known open-source toolkit for audio processing. We first split the audio waveform of each video clip into seconds (i.e., one segment = one second). In each second, we then extract 13 Mel-Frequency Cepstral Coefficients (MFCCs) features and obtain flattened 572-dimensional vectors¹. Hence, our final vocal features have a shape of 5 (seconds) \times 572 per sample.

2.3.2. Facial Features

For facial feature extraction, we utilize dlib [12], a popular open-source software for computer vision tasks such as face recognition and verification. We extract 68 facial landmarks (i.e., x and y coordinates) for each frame (1 FPS) in the video clip, and obtain 136-dimensional vectors by concatenating each x and y coordinates. If a face is not detected in a frame, we replace it with zero vectors. Our final facial features have a shape of 5 (seconds) \times 136 per sample.

2.3.3. Gestural Features

We use OpenPose [13], a state-of-the-art open-source pose estimation software, to extract gesture information. As most of the movements of candidates come from the upper part of the body, we consider 13 key points from nose, neck, shoulders, elbows, wrists, eyes and ears. Similar to the facial landmark extraction

¹13 (number of features) \times 1 (second) \times 22,050 (sample rate) / 512 (hop length).

process, we extract 13 body key points for each frame in the video clip. We then obtain 26-dimensional vectors by concatenating each x and y coordinates. Our final gestural features have a shape of 5 (seconds) \times 26 per sample.

3. Emotional Intensity Prediction

In this section, we first introduce the problem statement that describes the objective of the proposed model. We then present the overall architecture of the proposed model to predict perceived emotional intensity in the given video clip.

3.1. Problem Statement

The goal of the proposed model is to predict perceived emotional intensity by learning low-level non-verbal features including vocal, facial, and gestural features. More specifically, we define this problem as a regression problem that predicts emotional intensity score ranging between 0–1 for a given short video clip. Suppose we have a set of video clips $C = \{c_n\}_{n=1}^{|C|}$ and each clip can be represented as $c_n = (X_m^n \in \mathbb{R}^{t \times d_m}, X_f^n \in \mathbb{R}^{t \times d_f}, X_b^n \in \mathbb{R}^{t \times d_b})$ where $X_m^n, X_f^n, X_b^n, d_m, d_f, d_b,$ and t represent the vocal features, the facial features, the gestural features, the dimension of vocal features, the dimension of facial features, the dimension of gestural features, and the length of sequences, respectively. By learning a set of clips C , the proposed model can predict the perceived emotional intensity score of the given clip by learning latent features of candidates' non-verbal behaviors in debates.

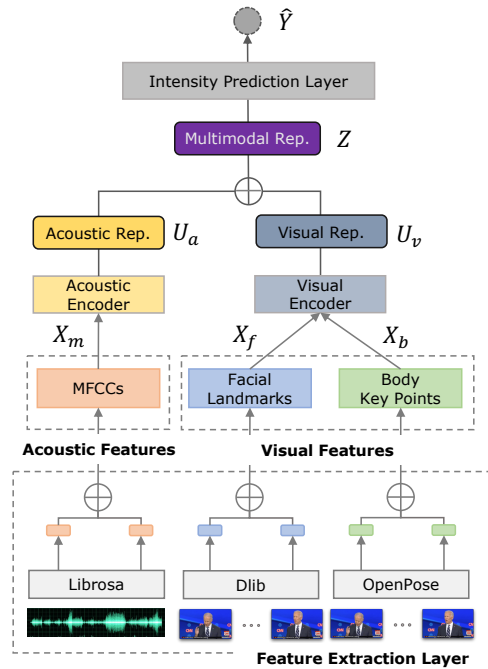


Figure 2: An overall architecture of the proposed model.

3.2. The Proposed Model

Figure 2 illustrates the overall architecture of the proposed model. To generate acoustic and visual representations from the low-level non-verbal feature vectors, the model employs two encoders, acoustic and visual encoders, respectively. The

acoustic encoder takes vocal features as input to generate acoustic representation as follows:

$$U_a = Dropout(AcousticEncoder(X_m^n)) \quad (1)$$

where $AcousticEncoder(\cdot)$, $Dropout(\cdot)$, X_m^n , and U_a denote the acoustic encoder, dropout, vocal features, and acoustic representation, respectively. The acoustic encoder consists of a stacked long short-term memory (LSTM) [14] where the second LSTM layer takes all sequences returned from the first LSTM layer.

On the other hand, the visual encoder concatenates the facial features and the gestural features, and employs the stacked LSTM to make visual representation, U_v , calculated as follows:

$$U_v = Dropout(VisualEncoder(X_f^n \oplus X_b^n)) \quad (2)$$

where $VisualEncoder(\cdot)$, $Dropout(\cdot)$, X_f^n , X_b^n , and U_v denote the visual encoder, dropout, facial features, gestural features, and visual representation, respectively. Note that both acoustic and visual encoders apply the same architecture (i.e., stacked LSTM).

Finally, the intensity prediction layer colligates acoustic and visual representations to generate a multimodal representation and predict a perceived emotional intensity for the given clip as follows:

$$Z = Dropout(\mathcal{F}_{fuse}(U_a \oplus U_v)) \quad (3)$$

$$\hat{Y} = Sigmoid(\mathcal{F}_{pred}(Z)). \quad (4)$$

where $F_{fuse}(\cdot)$, $F_{pred}(\cdot)$, $Sigmoid(\cdot)$, and Z denote the fully connected layer for fusion, fully connected layer for intensity prediction, sigmoid activation function, and multimodal representation, respectively. The emotional intensity prediction task is defined as a regression problem, and the mean squared error [15] and Adam [16] as the loss function and optimizer, respectively, are used.

4. Experiments

4.1. Experimental Settings

We split the dataset into the train and test sets with a 8:2 ratio. We also set the number of units, dropout rate, batch size, epochs, and learning rate to 8, 0.25, 128, 300, and 0.0002, respectively. Note that all weights are randomly initialized.

4.2. Baseline Methods

To evaluate the overall performance of the proposed model, we compare with the following five methods: (i) Support Vector Machine [17] (*SVM*), (ii) K-Nearest Neighbors [18] (*KNN*), (iii) Random Forest [19] (*RF*), (iv) Early Fusion LSTM (*EF-LSTM*), and (v) Late Fusion LSTM (*LF-LSTM*). For *SVM*, *KNN*, and *RF*, we aggregate all features by averaging and concatenating vectors. For *EF-LSTM*, we concatenate the vocal, facial, and gestural features, and pass them to the LSTM layer followed by a fully connected layer with sigmoid activation function. Lastly, for *LF-LSTM*, we concatenate the outputs of three different LSTMs (i.e., vocal/facial/gestural features), and add a fully connected layer with sigmoid activation function to predict emotional intensity.

4.3. Experimental Results

4.3.1. Overall Performance

Table 1 shows the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) of the baseline models and the proposed model. As shown in Table 1, the proposed model shows high emotional intensity performance (i.e., 0.12 of RMSE and 0.20 of MAPE), which outperforms other baseline models. This indicates that the proposed fusion method using the stacked LSTM can capture distinct indicators in predicting emotional intensity. Among the baselines, we find that *RF* shows the higher performance than the traditional machine-learning methods. This is because *RF* randomly selects subsets of high dimensional input features (i.e., vocal + facial + gestural features) via the bagging process to avoid overfitting. We also find that *LF-LSTM* achieves higher performance than *EF-LSTM*. This suggests that aggregating features at the decision-level (i.e., late fusion) helps the model capture important signals for emotional intensity prediction. We provide the prediction results of the proposed model available at: <https://dsail-skku.github.io/INTERSPEECH2022/>.

Table 1: Performance comparisons between the five baseline models and the proposed model.

Model	RMSE	MAPE
<i>SVM</i>	0.15	0.24
<i>RF</i>	0.13	0.23
<i>KNN</i>	0.15	0.23
<i>EF-LSTM</i>	0.15	0.26
<i>LF-LSTM</i>	0.14	0.22
The Proposed Model	0.12	0.20

4.3.2. Analysis on Different Modalities

To analyze the importance of each input modality for predicting emotional intensity, we conduct a performance analysis on the unimodal and bimodal models. For the unimodal model, we use a stacked LSTM layer to generate unimodal representation followed by the intensity prediction layer in Figure 2. As shown in Table 2, the model trained with the vocal features achieves higher performances (0.15 of RMSE and 0.25 of MAPE) than the models trained with the facial and the gestural features. That is, the candidates' vocal characteristics are more useful than their faces and gestures in predicting perceived emotional intensity. For the bimodal model, we first concatenate two different input features, and feed them into a stacked LSTM layer followed by the intensity prediction layer. Here, most of the bimodal models show better performance than unimodal models, implying that considering multiple modalities except V+F improves the performance. It is worth to note the performance improvement by the F+G model, which reveals that the candidates' facial and gestural features play a complementary role in predicting emotional intensity.

5. Case Study: 2020 Democratic Party Presidential Primaries in the USA

We now investigate whether there is a relation between the perceived emotional intensity and the changes in candidates' favorability. We first define the debate performance (i.e., net favorability) on politicians. We next apply the proposed model

Table 2: Performance comparisons on unimodal, bimodal, and trimodal emotional intensity prediction models. V, F, and G denote vocal, facial and gestural features, respectively.

Type	Modality	RMSE	MAPE
Unimodal	V	0.15	0.25
	F	0.17	0.30
	G	0.16	0.26
Bimodal	V + F	0.15	0.26
	V + G	0.14	0.24
	F + G	0.14	0.23
Trimodal	V + F + G	0.12	0.20

to the rest (i.e., not in our training set) of the videos in 2020 Democratic Party presidential primaries in the USA, to extract the emotional intensity of each candidate. Finally, we calculate the linear correlation between the perceived emotional intensity and the net favorability.

5.1. Net Favorability

To measure debate performance (i.e., voting behavior) on politicians, we leverage the before and after polls of FiveThirtyEight² and Morning Consult³ for obtaining favorability and unfavorability ratings [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. Note that FiveThirtyEight claims that the “poll is based on a nationally-representative probability sample of adults age 18 or older. Questions presented in this document were only asked of those who are likely to vote in the Democratic primary or caucus (n=3,360).” We calculate the performance metric as the net change in the (favorability-unfavorability percentages) for before and after the debate, which we shorten to the name as *net favorability*. Change in favorability provides an equal baseline for all candidates and thus is expected to be more representative of debate skill rather than an accumulation of historic events compared to other metrics.

Formally, we compute the net favorability as follows:

$$N_{ij} = \text{Before Favorability} - \text{Before Unfavorability} \quad (5)$$

$$N'_{ij} = \text{After Favorability} - \text{After Unfavorability} \quad (6)$$

$$P_{ij} = N'_{ij} - N_{ij} \quad (7)$$

where i denotes debate number, j denotes candidate appearing in one or more debate, and (7) represents the net favorability.

5.2. Emotional Intensity Extraction

To extract the emotional intensity of each candidate from the videos the test set, we first segment video clips where only one candidate appears and speaks (i.e., over 15h). We next apply the proposed model to predict the emotional intensity score every 5-second time window. For example, if a video clip is 8 seconds long, our model predicts emotional intensity during 1 to 5, 2 to 6, 3 to 7, and 4 to 8 seconds, respectively. We finally calculate the average score of the predicted emotional intensities for the video clip.

²<https://fivethirtyeight.com/>.

³<https://morningconsult.com/>.

5.3. Emotional Intensity and Net Favorability

We investigate whether there is a relation between a candidate’s emotional intensity and his/her net favorability. To this end, we first calculate the mean and the standard deviation (std.) of emotional intensity for each candidate in 11 different debates. Note that a candidate can appear in one or more debates. We then measure the Pearson correlation coefficient between mean and std. of emotional intensity and net favorability. As a result, we find no significant relation between the mean emotional intensity and the net favorability ($r = -0.08, p = 0.39$), implying that a candidate’s mean emotional intensity does not affect performance. However, as shown in Figure 3, we find that there is a positive correlation between std. of emotional intensity and the net favorability ($r = 0.21, p < 0.05$), implying the more candidate shows dynamics during his/her speech, the more positive changes in his/her favorability can happen.

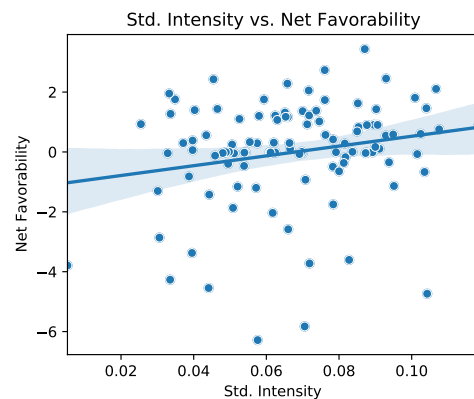


Figure 3: Scatter plot for standard deviation of emotional intensity and net favorability.

6. Conclusion

In this paper, we proposed a multimodal deep-learning model that uses non-verbal features as inputs and predicts the perceived emotional intensity. To this end, we collected a novel multimodal debate dataset from the 2020 Democratic Party presidential primaries in the USA. To train our proposed model, we extracted non-verbal features including vocal, facial, and gestural features. Our model achieved high emotional intensity prediction performance that outperforms other baseline methods. Our case study on the 2020 Democratic Party presidential primaries in the USA shows that the standard deviation of the perceived emotional intensity is positively correlated with the changes in candidates’ favorability in public polls. We believe the proposed model is useful in understanding non-verbal communications of political leaders. Furthermore, the model can be used as a research tool to extract the perceived emotional intensity from videos.

7. Acknowledgements

This research was supported by the framework of international cooperation program managed by the National Research Foundation of Korea (NRF-2020K2A9A2A11103842) and the National Research Foundation (NRF) of Korea Grant funded by the Korean Government (MSIT) (No. 2021R1A4A3022102).

8. References

- [1] R. D. Masters, D. G. Sullivan, J. T. Lanzetta, G. J. McHugo, and B. G. Englis, "The facial displays of leaders: Toward an ethology of human politics," *Journal of Social and Biological Structures*, vol. 9, no. 4, pp. 319–343, 1986.
- [2] R. D. Masters and D. G. Sullivan, "Nonverbal displays and political leadership in france and the united states," *Political Behavior*, vol. 11, no. 2, pp. 123–156, 1989.
- [3] D. G. Sullivan, "Emotional responses to the nonverbal behavior of french and american political leaders," *Political Behavior*, vol. 18, no. 3, pp. 311–325, 1996.
- [4] C. Boussalis and T. G. Coan, "Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation," *Political Communication*, vol. 38, no. 1-2, pp. 75–97, 2021.
- [5] C. Boussalis, T. G. Coan, M. R. Holman, and S. Müller, "Gender, candidate emotional expression, and voter reactions during televised debates," *American Political Science Review*, vol. 115, no. 4, pp. 1242–1257, 2021.
- [6] R. C. Anderson and C. A. Klofstad, "Preference for leaders with masculine voices holds in the case of feminine leadership roles," *PloS one*, vol. 7, no. 12, p. e51216, 2012.
- [7] R. C. Anderson, C. A. Klofstad, W. J. Mayew, and M. Venkatchalam, "Vocal fry may undermine the success of young women in the labor market," *PloS one*, vol. 9, no. 5, p. e97506, 2014.
- [8] C. A. Klofstad, R. C. Anderson, and S. Peters, "Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1738, pp. 2698–2704, 2012.
- [9] B. J. Dietrich, M. Hayes, and D. Z. O'brien, "Pitch perfect: vocal pitch and the emotional intensity of congressional speech," *American Political Science Review*, vol. 113, no. 4, pp. 941–962, 2019.
- [10] S. Nowicki Jr and J. Carton, "The measurement of emotional intensity from facial expressions," *The Journal of social psychology*, vol. 133, no. 5, pp. 749–750, 1993.
- [11] M. Sun, Y. Mou, H. Xie, M. Xia, M. Wong, and X. Ma, "Estimating emotional intensity from body poses for human-robot interaction," *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.
- [12] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, p. 1755–1758, Dec. 2009.
- [13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computing Research Repository (CoRR)*, vol. abs/1412.6980, 2015.
- [17] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.
- [18] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 4, pp. 580–585, 1985.
- [19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [20] J. Wolfe, "A final look at who won and lost the first democratic debates," Jul. 2019. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-poll/>
- [21] —, "Who won the third democratic debate?" Sep. 2019. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-september-poll/>
- [22] —, "Who won the fourth democratic debate?" Oct. 2019. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-october-poll/>
- [23] —, "Who won the fifth democratic debate?" Nov. 2019. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-november-poll/>
- [24] —, "Who won the december democratic debate?" Dec. 2019. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-december-poll/>
- [25] —, "Who won the january democratic debate?" Jan. 2020. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-january-poll/>
- [26] —, "Who won the new hampshire democratic primary debate?" Feb. 2020. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-first-february-poll/>
- [27] —, "Who won the south carolina democratic debate?" Feb. 2020. [Online]. Available: <https://projects.fivethirtyeight.com/democratic-debate-south-carolina-poll/>
- [28] —, "Who won the biden-sanders debate?" Mar. 2020. [Online]. Available: <https://projects.fivethirtyeight.com/biden-sanders-debate-poll/>
- [29] N. Silver, "Polls since the second debate show kamala harris slipping," Aug. 2019. [Online]. Available: <https://fivethirtyeight.com/features/polls-since-the-second-debate-show-kamala-harris-slipping/>
- [30] E. Yokley, "Bloomberg loses ground following debate debut in las vegas," Feb. 2020. [Online]. Available: <https://morningconsult.com/2020/02/21/michael-bloomberg-polling-post-debate-las-vegas/>