Keeping Up Appearances: Computational Modeling of Face Acts in Persuasion Oriented Discussions

Ritam Dutt **Rishabh Joshi** **Carolyn Penstein Rosé**
Language Technologies Institute
Carnegie Mellon University
{rdutt, rjoshi2, cprose}@cs.cmu.edu

Abstract

The notion of *face* refers to the public self-image of an individual that emerges both from the individual's own actions as well as from the interaction with others. Modeling face and understanding its state changes throughout a conversation is critical to the study of maintenance of basic human needs in and through interaction. Grounded in the politeness theory of Brown and Levinson (1978), we propose a generalized framework for modeling *face acts* in persuasion conversations, resulting in a reliable coding manual, an annotated corpus, and computational models. The framework reveals insights about differences in face act utilization between asymmetric roles in persuasion conversations. Using computational models, we are able to successfully identify face acts as well as predict a key conversational outcome (e.g. donation success). Finally, we model a latent representation of the conversational state to analyze the impact of predicted face acts on the probability of a positive conversational outcome and observe several correlations that corroborate previous findings.

1 Introduction

Politeness principles, displayed in practice in day-to-day language usage, play a central role in shaping human interaction. Formulations of politeness principles are related to basic human needs that are jointly met in and through interaction (Grice et al., 1975; Brown et al., 1987; Leech, 2016). Natural language offers various ways to enact politeness. One of the most influential politeness theories from linguistics is proposed in (Brown and Levinson, 1978), in which a detailed exposition is offered of the individual actions whose cumulative effect results in *saving face* and *losing face*, along with a consideration of cost. Using this framework, it is possible to analyze how interlocutors make decisions about where and how these devices should be used based on an intricate cost-benefit analysis (Brown et al., 1987). We refer to these component actions here as **face acts**.

The idea of *face acts* appears quite attractive from a computational standpoint for their potential role in understanding what is "meant" from what is "said" (Grice et al., 1975; Brown et al., 1987; Leech, 2016). Consequently, politeness has been widely researched in various domains of language technologies (Walker et al., 1997; Gupta et al., 2007; Wang et al., 2012; Abdul-Mageed and Diab, 2012; Danescu-Niculescu-Mizil et al., 2013) in addition to foundational work in pragmatics and sociolinguistics (Brown et al., 1987; Grice et al., 1975; Leech, 2016). However, much prior work modeling politeness reduces the problem to a rating task or binary prediction task, separating polite and impolite behavior. Consequently, what the models end up learning is mainly overt markers of politeness or rudeness, rather than the underlying indirect strategies for achieving politeness or rudeness through raising or attacking face, even in the indirect case where no overt markers of rudeness or politeness might be explicitly displayed.

In contrast, the main contribution of this work is the investigation of eight major face acts, similar to dialogue acts, including an investigation of their usage in a publicly available corpus of Wang et al. (2019). In the selected corpus, a persuader (**ER**) is tasked with convincing a persuadee (**EE**) to donate money to a charity. The nature of the task prompts frequent utilization of face acts in interaction, and thus these face acts are abundantly present in the chosen dataset. We also provide a generalized framework for operationalizing face acts in conversations as well as design an annotation scheme to instantiate these face acts in the context of persuasion conversations (§2.1, §2.3). We offer the annotations we have added to this public dataset as another contribution of this work

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7473–7485,
November 16–20, 2020.

(§2.2). Additionally, we develop computational models to identify face acts (§3.1) as well as construct a latent representation of conversational state to analyze the impact of face acts on conversation success (§3.2). We achieve 0.6 F1 on classifying face acts (§5.1), and 0.67 F1 in predicting donation outcome (§5.2). We observe that the predicted face acts significantly impact the local probability of donation (§5.3)[1].

2 Framework

2.1 Face Representation

Face, based on the politeness theory of Brown et al. (1987), reflects the 'public self-image' that every rational adult member of society claims for themselves. It can be subdivided into *positive face*, referring to one's want to be accepted or valued by society, and *negative face*, referring to one's right to freedom of action and freedom from imposition.

We refer to 'face acts' as utterances/speech acts that alter the positive and/or the negative face of the participants in a conversation. We hereby refer to the acts that attack one's face as Face Threatening Acts (FTA) and those acts that raise one's face as Face Saving Acts (FSA). For example, criticizing an individual is an attack on the other's positive face, whereas refusing to comply with someone's wishes, raises one's own negative face. We also note that a single utterance or act can simultaneously affect the face of one or both participants in a conversation. For example, a refusal to donate to a charity because they do not trust the charity involves asserting one's negative face as well as decreasing the charity's positive face.

The implication of a face act between the participants is governed by several factors such as 'power' and relative 'social distance', as well as the relative threat ('ranking') of the face act (Brown and Levinson, 1978). For example, refusing to comply with the orders of one's superior is more threatening than requesting a friend for some change.

Moreover, face acts need to be contextualized for a particular situation based on the rights and obligations of the individual participants, such as in compliance-gaining episodes (Wilson et al., 1998). For example, a teacher has the responsibility and right to assign homework to the students. Such an action cannot be perceived as an attack on negative face, even though the student is reluctant to do so.

Based on the definition of face and face acts, we design a generalized annotation framework to capture the face dynamics in conversation. We instantiate our framework on a publicly-available corpus on persuasion dialogues.

2.2 Dataset Description

We use the pre-existing persuasion corpus of Wang et al. (2019). Each conversation comprises a series of exchanges where the persuader (ER) has to convince the persuadee (EE) to donate a part of their task earnings to the charity, *Save the Children*. This selected corpus is well-situated for our task since each conversation is guaranteed to have a potential face threat (i.e., a request for money) and hence, we can expect face act exchanges between the two participants. It also sets itself apart from other goal-oriented conversations such as restaurant reservations and cab booking (Budzianowski et al., 2018) since in those cases the hearer is obligated to address what might otherwise come across as an FTA (request/ booking), and thus in those cases non-compliance can be assumed to be due to logistic issues rather than an unwillingness to co-operate.

In the selected corpus, the participants are Amazon Mechanical Turk workers who are anonymous to each other, which controls for the 'social distance' variable. Moreover, the participants have similar 'power', with one role having some appearance of authority in that it represents an organization, but the other role representing possession of some desired object (i.e., money). Thus, we argue that although ER imposes an FTA by asking for donation, EE is equally at liberty to refuse. Moreover, ER does not incur a penalty for failing to persuade. In fact the corpus includes some conversations that do not talk about donation at all. We also emphasize that the task was set up in a manner such that EE come into the interaction blind to the fact that ER have been tasked with asking them to donate.

We assess the success of a conversation based on whether EE agrees to donate to the charity. We label successful conversations as *donor* conversations and *non-donor* conversation otherwise. We refer the reader to Wang et al. (2019) for more details about the dataset.

2.3 Annotation Framework

In a two-party conversation, a face act can either raise (+) or attack (-) the positive face (Pos) or

[1]We include our annotation framework in Appendix and the annotated dataset and code is publicly available at https://github.com/ShoRit/face-acts

Face Act	Description
SPos+	(i) S posit that they are **virtuous** in some aspects or they are **good**. (ii) S **compliment the brand or item they represent or endorse** and thus project their credibility. (iii) S state their **preference or want**, something that they like or value.
SPos-	(i) S **confess** or apologize for being unable to do something that is expected of them. (ii) S **criticise or humiliate themselves**. They damage their reputation or values by either saying they are not so virtuous or criticizes some aspect of the brand/item they endorse or support.
HPos+	(i) **S compliment H** either for H's virtues, efforts, likes or desires. It also extends to S **acknowledging the efforts of H** and **showing support for H**. (ii) S can also provide an **implicit compliment to incentivize H** to do something good. (iii) S **empathize /** sympathize or in general **agree with H**. (iv) S is **willing to do the FTA** as imposed by H (implying that the FTA is agreeable to S.)
HPos-	(i) S voice **doubts or criticize H** or the product/brand that H endorses. (ii) S **disagree with H** over some stance, basically contradicting their viewpoint. (iii) S is either **unaware** or **indifferent** to H's wants or preferences.
SNeg+	(i) S **reject** or are unwilling to do the **FTA.** Stating the reason does not change the circumstances of non-compliance but sometimes helps to mitigate the face act.
SNeg-	(i) S **offer to assist H**.
HNeg+	(i) S seek to **decrease the imposition of the FTA** on H by either decreasing the inconvenience such as providing alternate, simpler ways to carry out the FTA or decrease the threat associated with the FTA. (ii) S **apologize for the FTA** to show that S understood the inconvenience of imposing the request but they have to request nevertheless.
HNeg-	(i) S **impose an FTA on the H**. The FTA is some act which H would not have done on their own. (ii) S **increase the threat** or ranking of the FTA (iii) S **ask/request H** for assistance?

Table 1: Generalized framework for situating and operationalizing face acts in conversations. The predicates for each of the face act are highlighted in bold.

negative face (Neg) of either the speaker (S) or the hearer (H), leading to 8 possible different outcomes. For example, *HPos+* means raising the positive face of the hearer. We provide a generalized framework in Table 1 for labelling a speech act / utterance with one or more face acts, building upon the politeness theory of Brown and Levinson (1978). The framework is designed to be explicit enough to ensure the creation of a reliable coding manual for classifying face-acts, as opposed to the simple classification of requests and other directives as intrinsic FTAs (Brown and Levinson, 1978). Moreover, since we also seek to operationalize FSA, we make some departure from the original classification of directives. For example, we feel that compliments directed at the hearer, should be *HPos+* rather than *HNeg-* (as observed in Brown et al. (1987)) since an appreciation for someone's efforts is more desirable.

We highlight the predicates that result in a particular face act in bold in Table 1. For example, S claiming to be virtuous or doing some good deed amounts to raising their own positive face (*SPos+*). Although the framework is designed to be generalizable across domains, the predicates themselves need to be instantiated based on the domain of choice. For example, in this particular corpus, the act of requesting someone for donation counts as an FTA. We refer the readers to Table S3 in Appendix A which outlines how the face acts are instantiated for the specific persuasion dataset.

Each conversation in the dataset consists of 10 or more turns per participant with one or more utterances per turn. Each utterance is labeled with one or more face acts according to our annotation framework. Otherwise we label the utterance as 'Other' if no face act can be identified, or if the utterance contains no task-specific information (Eg: small talk). We consider ER to be a representative of the charity since ER advocates for donations on their behalf. We show the flowchart detailing the annotation framework in Figure S3 of Appendix A.

Validating the annotation scheme: Two authors of the paper annotated 296 conversations in total. The annotation scheme underwent five revisions, each time with three different conversations, eventually yielding a high Cohen's Kappa score of 0.85 across all face acts (Cohen, 1960). The revised scheme was then used to annotate the remaining conversations. We show an annotated conversation snippet in Table 4.

2.4 Summary Statistics

Our annotated dataset comprises 231 donor conversations and 65 non-donor conversations. Table 2 shows the distributions of different face acts em-

ployed by ER and EE respectively for both donor and non-donor conversations. We also note that certain face-acts do not occur in our corpus, such as *SPos-* for ER, presumably because ER does not have a reason to debase themselves or the charity they endorse. We provide a detailed explanation of the occurrence of such acts in the supplementary section. We observe multiple statistically significant differences in face act prevalence based on whether EE is a donor or non-donor. Some findings are intuitive, such as an increase in *HPos+* for Donor conversations (for both ER and EE). We argue that EE had acknowledged the efforts of the charity and was willing to donate, and was thus rewarded with compliments from ER. Likewise, *SNeg+* occurs significantly more in Non-donor situations, due to a refusal to comply. We note that a majority of the turns labeled 'Other' involve greetings or conversation exchanges unrelated to the main business of the conversations.

Face Acts	ER		EE	
	D	N	D	N
SPos+	19.95	23.03	8.29	6.51
SPos-	0.00	0.00	0.18	0.96^{*}
HPos+	23.08^{***}	16.24	36.17^{***}	21.07
HPos-	0.70	2.65^{*}	4.37	10.73^{**}
SNeg+	0.00	0.00	3.85	11.97^{***}
HNeg+	5.50	4.81	0.00	0.00
HNeg-	10.47	10.85	9.20	13.03
Other	40.31	42.42	37.94	35.73

Table 2: Distribution of different face acts for the donor (D) and non-donor (N) for ER and EE. *, **, and *** signify that the specific act is statistically significant for D and N according to the independent t-test with p-values ≤ 0.05, 0.01, and 0.001 respectively.

3 Methodology

3.1 Face act prediction

We model the task of computationally operationalizing face acts as a dialogue act classification task. Given an dialogue with n utterances, $D = [u_1, u_2, ..., u_n]$, we assign labels $y_1, y_2 ... y_n$ where $y_i \in Y$ represents one of 8 possible face acts or 'Other'. Although, we acknowledge that an utterance can have multiple face acts, we observe that multi-labeled utterances comprise only 2% of our dataset, and thus adopt the simplification of predicting a single face-act for each utterance[2]. Several tasks in the dialogue domain, such as emotion recognition (Majumder et al., 2019; Jiao et al., 2019), dialogue act prediction (Chen et al., 2018; Raheja and Tetreault, 2019) and open domain chit-chat (Zhang et al., 2018b; Kumar et al., 2020), have achieved state-of-the-art results using a hierarchical sequence labelling framework. Consequently, we also adopt a modified hierarchical neural network architecture of Jiao et al. (2019) that leverages both the contextualized utterance embedding and the previous conversational context for classification. We hereby adopt this as the foundation architecture for our work and refer to our instantiation of the architecture as BERT-HIGRU.

Architecture of BERT-HIGRU: An utterance u_j is composed of tokens $[w_0, w_1, ..., w_K]$, which are represented by their corresponding embeddings $[e(w_0), e(w_1), ..., e(w_K)]$. In BERT-HIGRU, we obtain these using a pre-trained BERT model (Devlin et al., 2019). We pass these contextualized word representations through a BiGRU to obtain the forward $\overrightarrow{h_k}$ and backward $\overleftarrow{h_k}$ hidden states of each word, before passing them into a Self-Attention layer. This gives us corresponding attention outputs, $\overrightarrow{ah_k}$ and $\overleftarrow{ah_k}$. Finally, we concatenate the contextualized word embedding with the GRU hidden states and Attention outputs in our *fusion layer* to obtain the final representation of the word. We perform max-pooling over the fused word embeddings to obtain the j^{th} utterance embedding, $e(u_j)$. Formally,

$$\overrightarrow{h_k} = \text{GRU}\left(e\left(w_k\right), \overrightarrow{h_{k-1}}\right)$$
$$\overleftarrow{h_k} = \text{GRU}\left(e\left(w_k\right), \overleftarrow{h_{k+1}}\right)$$
$$\overrightarrow{ah_k} = \text{SelfAttention}(\overrightarrow{h_k})$$
$$\overleftarrow{ah_k} = \text{SelfAttention}(\overleftarrow{h_k})$$
$$e_c(w_k) = \tanh(W_w[\overrightarrow{ah_k}; \overrightarrow{h_k}; e(w_k); \overleftarrow{h_k}; \overleftarrow{ah_k}] + b_w)$$
$$e(u_j) = \max(e_c(w_1), e_c(w_2), ...e_c(w_K)) \quad (1)$$

Similarly, we calculate the contextualized representation of an utterance $e_c(u_j)$ using the conversation context. In departure from Jiao et al. (2019), we pass $e(u_j)$ through a uni-directional GRU that yields the forward hidden state $\overrightarrow{H_j}$. Masked Self-Attention over the previous hidden states, yields $\overrightarrow{AH_j}$. We fuse $e(u_j)$, $\overrightarrow{H_j}$ and $\overrightarrow{AH_j}$ before passing it through a linear layer with tanh activation to obtain $e_c(u_j)$. This ensures that current $e_c(u_j)$ is not influenced by future utterances, enabling us to observe change in donation probability over time

[2]For each utterance with multiple labels, one is randomly select from that set to be treated as the Gold label.

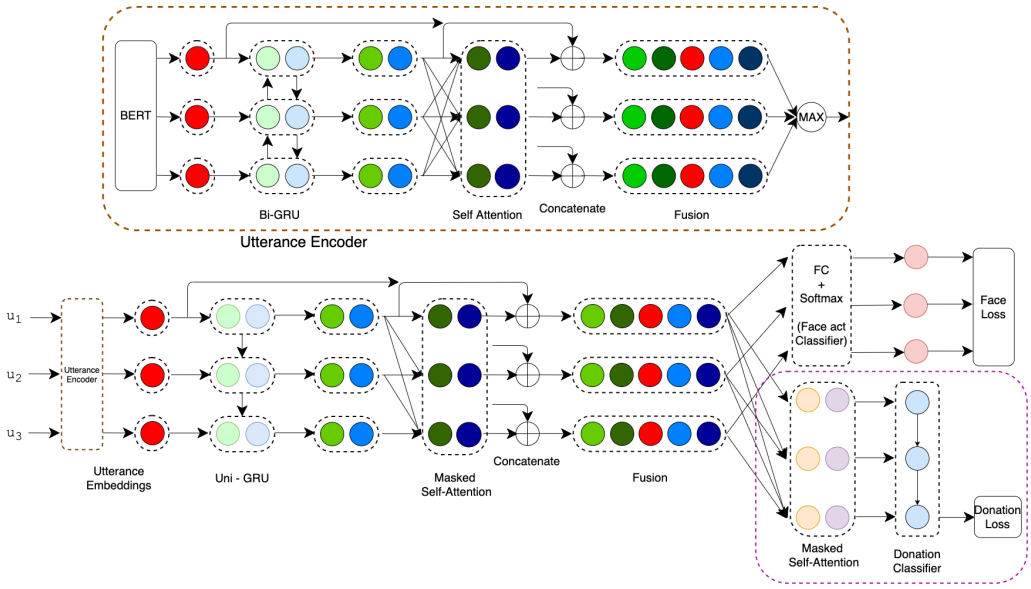

Figure 1: Overview of BERT-HIGRU. We first encode the utterances by passing the BERT representations of the token through a BiGRU layer followed by Self Attention. The BERT, BiGRU and Self-Attention outputs are then fused to get the final token representation before max pooling. This utterance representation is passed through a uni-directional GRU followed by Masked-Self-Attention and fusion. One part of the model uses the face classifier to predict the face-act of each utterance while the other model uses another layer of Masked-Self-Attention to predict the donation probability. The details are in Section 3

in Section 3.2

$$\overrightarrow{H_j} = \text{GRU}\left(e\left(u_j\right), \overrightarrow{H_{j-1}}\right)$$
$$\overrightarrow{AH_j} = \text{MaskSelfAttention}(\overrightarrow{H_j})$$

$$e_c(u_j) = \tanh(W_u[\overrightarrow{AH_j}; \overrightarrow{H_j}; e(u_j)] + b_u) \quad (2)$$

We explore different hierarchical architecture variants which differ in the creation of contexualized embeddings $e_c(w_k)$ and $e_c(u_j)$ in Equation 1 and 2. (1) BERT-HIGRU includes only the final hidden state $\overrightarrow{H_j}$; (2) BERT-HIGRU-f additionally employs the utterance embedding $e_(u_j)$; and (3) BERT-HIGRU-sf, which also includes the attention vector $\overrightarrow{AH_j}$.

We feed the final contextualized utterance embedding $e_c(u_j)$ through a FC layer with dropout and project it onto the state space of face-acts. We then apply softmax to obtain a probability distribution over the face-acts, with negative logarithmic loss as the loss function. Given the true labels y and the predicted labels y', the loss is computed for all n utterances in a conversation as:

$$L_f = -\sum_{i=1}^{n}\sum_{y_j \epsilon Y} y_j log(y'_j) \quad (3)$$

3.2 Impact of face acts on donation

Donation Outcome Prediction: Brown et al. (1987) notes that the exchange of face acts contributes towards an evolving conversational state. We seek to view the evolving state representation within our sequence model and analyze its impact on conversation success. The best reflection of what the evolving conversational state accomplishes in the context of persuasion is whether a donation occurs or not. We thus add the prediction task as supervision and interpret the resulting conversation state based on how the probability of donation changes. We accomplish the supervision by incorporating another loss, called donation loss, in addition to the loss obtained for face acts.

For each utterance u_j, we apply masked self attention over the set of contextual utterance embeddings $e_c(u_j)$ till the j^{th} utterance and project it through a linear layer with tanh activation to obtain the donation score don_j. The tanh non-linearity ensures that the donation score remains between -1 and 1 and intuitively denotes the delta change in scores from the previous step. We finally compute the probability of donation o'_j at the j^{th} step, by applying sigmoid activation over the sum of probability at the previous step and the delta change don_j. This ensures that the o'^{th}_j probability is restricted between 0 and 1. We obtain the donation loss L_d similar to Yang et al. (2019) by taking the mean squared error of the donation probability at the last step o'_n and the actual donation outcome o_n. o_n is 1 if successful, otherwise 0. We also exper-

iment with Binary Cross Entropy loss and obtain similar results.

$$\begin{aligned} e_d(u_j) &= \text{MaskSelfAttention}(e_c(u_j)) \\ don_j &= \tanh(W_d[e_d(u_j)] + b_d) \\ o'_j &= \sigma(o'_{j-1} + don_j) \\ L_d &= (o'_n - o_n)^2 \end{aligned}$$

The donation loss is combined with the original face-act loss in a weighted fashion using some hyperparameter α, such that $\alpha \in [0, 1]$.

$$L_{tot} = \alpha L_f + (1 - \alpha) L_d \tag{4}$$

Correlating face acts with donation outcome: The aforementioned formulation enables us to obtain the donation probability at any given step and assess the impact of difference in conversational state (due to a specific face act) on the local assessment of the probability of donation. To quantify the impact, we perform linear regression with the donation probability at each time step (y_i) as the dependent variable. The independent variables includes the predicted face acts for that step (f_i^k) and the donation probability at the previous step y_{i-1}.

$$y_i = \beta_0 * y_{i-1} + \sum_{f^k} \beta_k * f_i^k \tag{5}$$

Here, β_k represents the coefficient of the corresponding face-act and β_0 the coefficient for y_{i-1}.

4 Experimental Setup

We describe the baselines and evaluation metrics here. We present the additional experimental details of our model in Appendix Table S1.

4.1 Baselines

Face act prediction: We employ different variants of BERT-HIGRU described in Section 3, namely the vanilla BERT-HIGRU, BERT-HIGRU-f (with residual connections (fusion)) and BERT-HIGRU-sf with self-attention and fusion.

To observe the effect of incorporating conversation context, we pass the utterance embedding $e(u_j)$ obtained from the utterance encoder directly into the face act classifier. We denote the different variants of utterance encoder employed as BERT-BiGRU, BERT-BiGRU-f, and BERT-BiGRU-sf with the same notation as the hierarchical variants.

To explore the impact of embedding choice on model performance, we experiment with pretrained Glove embeddings (Pennington et al., 2014) in addition to BERT tokens. We denote the hierarchical models with GloVe embeddings as HiGRU and those without conversation context as BiGRU.

Donation Outcome Prediction: We use the baselines mentioned above for predicting face acts and augment them with the donation loss component. We explore different values of α for weighing the two losses as described in Equation 4.

4.2 Evaluation Metrics

Face act prediction: We observe the model performance in predicting the face acts for (i) only the persuader (ER), (ii) only the persuadee (EE), and (iii) both ER and EE (All). We perform five-fold cross-validation due to the paucity of annotated data. We report performance in terms of mean accuracy as well as macro F1-scores across the five folds due to the high class imbalance.

Donation Outcome Prediction: For a given conversation, we observe the probability of donation at the final step as o'_n. We choose an appropriate threshold on o'_n across the five validation folds, such that a conversation with (o'_n) greater than the threshold is considered successful and vice versa. The F1 score is then computed on the binarized outcome. We choose macro F1-score as the evaluation metric due to the highly skewed distribution of the number of donor and non-donor conversations.

Correlating face acts with donation outcome: We quantify the impact of face acts on donation probability for both ER and EE through the corresponding coefficients obtained from the regression framework. A positive coefficient implies that the face act is positively correlated with the local donation probability and vice versa. We also note the fraction of times a particular face act contributed to a local increase in donation probability and denote it by *Frac*. A value of *Frac* > 0.5 for an act implies that the act increased the local donation probability more number of times than it decreased it.

5 Results

In this section we put forward the following research questions and attempt to answer the same.

Q1. How well does BERT-HIGRU predict face acts? (Section 5.1)

Q2. How well are we able to predict the outcome of the conversation? (Section 5.2)

Q3. How do individual face acts correlate with donation probability? (Section 5.3)

5.1 Face Act Prediction

Model Performance: We present the results of our models for face act prediction in Table 3 and glean several insights. Firstly, we observe that all models consistently perform better for ER than EE due to the more skewed distribution of EE and the presence of an extra face act (6 for ER vs 7 for EE). The difference in F1 is less noticeable between EE and All, despite an additional face act (8 for All), possibly due to the increase in labelled data.

Model	ER		EE		All	
	Acc	F1	Acc	F1	Acc	F1
BiGRU	0.68	0.58	0.59	0.49	0.62	0.49
BiGRU-f	0.69	0.57	0.59	0.50	0.62	0.49
BiGRU-sf	0.69	0.57	0.60	0.51	0.62	0.51
HiGRU	0.68	0.56	0.62	0.52	0.63	0.49
HiGRU-f	0.70	0.57	0.62	0.52	0.64	0.49
HiGRU-sf	0.69	0.59	0.62	0.56	0.64	0.52
BERT-BiGRU	0.73	0.63	0.66	0.57	0.67	0.56
BERT-BiGRU-f	0.72	0.62	0.66	0.58	0.66	0.57
BERT-BiGRU-sf	0.72	0.62	0.65	0.56	0.66	0.56
BERT-HiGRU	**0.74**	**0.63**	0.66	0.54	0.68	0.57
BERT-HiGRU-f	0.73	**0.63**	**0.67**	**0.61**	**0.69**	**0.60**
BERT-HiGRU-sf	0.73	0.62	**0.67**	0.60	0.68	0.59

Table 3: Performance of the various models on face act prediction. The best results are shown in bold.

Adding conversational context aids model performance as observed by the average increase of 1.3 in F1 scores across all model configurations (BiGRU to HiGRU). The highest gains are observed for BERT-HiGRU-sf and HiGRU-sf which attends over the set of previous utterances and thus can better reason about the current utterance.

The greatest boost, however, comes from incorporating BERT embeddings as opposed to pre-trained GloVe, which bears testimony to the efficacy of contextualized embeddings for several dialogue tasks (Yu and Yu, 2019; Lai et al., 2019).

In fact, with the inclusion of BERT, BERT-HiGRU-f performs as competitively as BERT-HiGRU-sf. We also note that the performance of BERT-HiGRU-f and BERT-HiGRU-sf are statistically significant as compared to the other baselines according to the McNemar's test (McNemar, 1947) with p-value ≤ 0.001.

Error Analysis: We present the confusion matrix of face act prediction for the BERT-HiGRU-f model for both ER and EE in Figures 2 and 3 respectively. We observe that a majority of the misclassification errors occurs when a face-act is predicted as 'Other' since it is the most frequent class and also shares a commonality with the remaining classes.

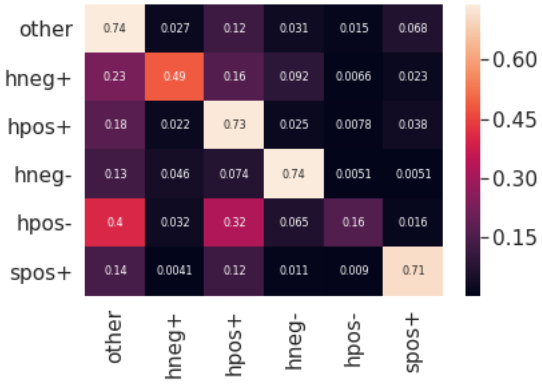

Figure 2: Confusion matrix for Persuader (ER)

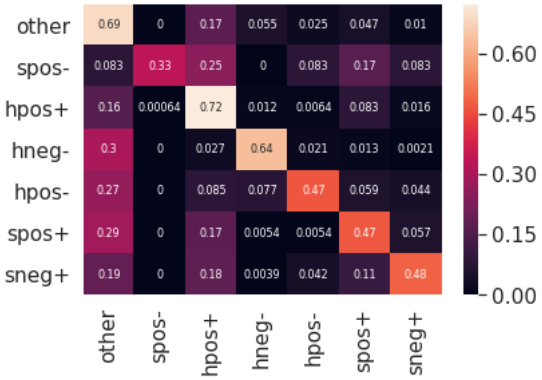

Figure 3: Confusion matrix for Persuadee (EE)

Some instances of misclassification errors for ER involves the labels *HPos+* and *HNeg+*. For example, utterances like "Every little bit will help the children we are trying to save." or "Anyways, even 50 cents goes a long way you know?" seek to decrease the imposition of the face threat by letting EE choose the donation amount, but also simultaneously provides EE with an opportunity to raise their face. For EE, the face act *SNeg+* is frequently misclassified as *SPos+* or *HPos+*, since a refusal to donate is often accompanied by EE stating their preference ("there are other charities I'd like to donate to over this one") or acknowledging the efforts of ER and the charity ("I appreciate you informing about this but I'll be donating at a later time"). The greatest misclassification happens for *HPos-* for ER and *SPos-* for EE, since they account for $\approx 1\%$ and $\approx 0.3\%$ of the dataset respectively. Consequently, they are often misclassified as *HPos+* and *SPos+* respectively since they affect the same face.

5.2 Donation Outcome Prediction

We use BERT-HiGRU-f as the model since it achieves the highest performance on face act prediction. Experiments with different values of alpha

Role	Utterance	True Face	Pred Face	Don Prob
ER	Hi	other	other	0.268
EE	hello	other	other	0.324
ER	How's it going	other	other	0.337
EE	it's going good so far this morning for me.	other	other	0.340
EE	what about you?	other	other	0.340
ER	Another day in paradise	other	other	0.340
EE	haha.	other	other	0.340
EE	do you live in the states ?	other	other	0.791
ER	Great state of Texas.	other	other	0.448
ER	How about you?	other	other	0.365
EE	chicago.	other	other	0.346
EE	so what is this about anyway?	other	other	0.792
EE	it says a child's charity.	other	other	0.448
EE	I am lost.	other	other	0.365
ER	I guess I'm supposed to persuade you to donate your bonus to the save the children charity.....	hneg-	other	0.346
ER	I want you to keep your bonus obviously	hneg-	other	0.342
EE	it would help to know what childrens charity as well with some information about it.	hpos+	other	0.341
EE	usually thats how you persuade ppl to donate.	other	other	0.341
EE	show a pic of a kid etc.	other	other	0.340
ER	How much do you like to donate to the charity now?	hneg-	hneg-	0.340
ER	Your donation will be directly deducted from your task payment.	hneg+	other	0.791
ER	You can choose any amount from $0 to all your payment	hneg+	other	0.856
EE	I do not wish to donate.	sneg+	sneg+	0.464
EE	I've been given no info about the charity.	hpos-	sneg+	0.369
ER	They help children in warzones and other poor nations to get food and clothes	spos+	spos+	0.797
EE	oh ok, who is they?	hneg-	hneg-	0.449
EE	what is the organization ?	hneg-	hneg-	0.365
ER	Save The Children is the name	spos+	other	0.796
EE	no, I do not wish to donate at this time.	sneg+	sneg+	0.449
EE	there are other charities I'd like to donate to over this one.	spos+	sneg+	0.802
EE	I'm sorry.	hpos+	spos-	0.858
EE	I don't have a lot to work with either.	sneg+	sneg+	0.464
ER	Think of the poor kids in Syria who could get so much for the price of a coffee in Chicago.	other	hpos+	0.812
ER	Do you really need all you have when they have nothing at all?	hpos-	other	0.453
EE	I don't have much money for myself either which is why I consider this to be my part time job.	sneg+	sneg+	0.366
EE	I already work full-time to make ends meet.	sneg+	sneg+	0.346
EE	I'm sorry	hpos+	spos-	0.793
ER	But you have a full time job, food, shelter and I'm sure you have family and friends.	hpos+	other	0.857
ER	These kids families were murdered and they live in rubble that used to be their home, nobody to care for them and they only eat what they can find off the street from the dead.	other	hpos+	0.464
ER	If they eat at all	hpos-	hpos+	0.369
EE	that is very sad.	hpos+	hpos+	0.797
EE	I'd like to look into this charity more before I donate as well.	sneg+	other	0.857
EE	I'd like to see how the money is dispersed in the company	hneg-	other	0.464

Table 4: An example conversation consisting of true and predicted face acts, along with donation probabilities. The persuader was unsuccessful in convincing the EE to donate. The utterances of the EE are in cyan.

reveal that $\alpha = 0.75$ achieves the highest F1 score on both face acts (0.591) and donation outcome (0.672). The F1-score for face acts is comparable to the best performing model in 5.1 but the performance for donation outcome increases significantly from 0.545 to 0.672. When α = 1.0, the donation outcome prediction is similar to random choice. On the other extreme, when α = 0, i.e. in the absence of L_f, the donation outcome behaves like random choice since, the model is unable to learn an appropriate latent state representation.

5.3 Correlation of face acts with donation.

We present the coefficients of face acts for ER and EE obtained from the regression framework 5 and the corresponding fraction (Frac) of times the face act increased the donation probability in Table 5.

We observe several face acts which are statistically significant based on the corresponding p-value of the coefficients (highlighted in bold).

	ER		EE	
Face Act	Frac	Coeff	Frac	Coeff
HPos-	0.464	-0.036	0.661	0.003
HPos+	0.646	0.004	0.605	-0.007
HNeg+	0.753	**0.029***	-	-
HNeg-	0.709	**0.023***	0.744	**0.043*****
SPos+	0.504	**-0.052*****	0.553	**-0.026***
SPos-	-	-	1.000	0.002
Other	0.742	**0.031*****	0.717	**0.023***
SNeg+	-	-	0.435	**-0.038***

Table 5: Coefficients of the face acts, for ER and EE obtained from linear regression. A positive coefficient implies positive correlation. *, *** indicate statistical significance with p values $\leq$ 0.05 and 0.001.

We observe unsurprisingly a positive correlation (0.029) for *HNeg+* for ER since decreasing the imposition of the FTA is likely to influence donation. Likewise, a criticism of EE (*HPos-*), decreases the likelihood of donation and thus has a negative cor-

relation (-0.036). We also observe a negative coefficient for *SPos+* (-0.052) which corroborates the finding of Wang et al. (2019), that appealing to the credibility of the organization correlates negatively with donation outcome.

Similarly, for EE, asserting one's face / independence (*SNeg+*) corresponds to a decrease in the donation probability (-0.038). Likewise, *HNeg+* corresponds to an increase in donation probability (0.043) since these face acts increase user engagement. We attribute the negative correlation for *SPos+* (-0.026) and *HPos+* (-0.007) to the fact that they often occur along with *SNeg+* and hence decreases the outcome probability. Nevertheless, *SPos+* and *HPos+* increase the donation probability 60.5% and 55.3% of the times. However, the learned model errs in assuming that *HPos-* (for EE) and *HNeg-* (for ER) increases the donation probability and requires further investigation

We illustrate the effect of face act on the local donation probability through a conversation snippet in Table 4. We observe a noticeable reduction in donation probability associated with *SNeg+* (for EE) and *HPos-* (for ER). Likewise, face acts corresponding to *HNeg+* (for ER) and *HPos+* (for EE) result in an increase in donation probability.

6 Related Work

Although politeness derailment and politeness evolution in dialogue have been previously investigated in the NLP literature (Chang and Danescu-Niculescu-Mizil, 2019; Danescu-Niculescu-Mizil et al., 2013), the prior work is distinguished from our own in that they do not explicitly model face changes of both parties over time. Rather, Danescu-Niculescu-Mizil et al. (2013) utilizes requests annotated for politeness to create a framework specifically to relate politeness and social power. Other previous work attempt to computationally model politeness, using politeness as a feature to identify conversations that appear to go awry in online discussions (Zhang et al., 2018a). Previous work has also explored indirect speech acts as potential sources of face-threatening acts through blame (Briggs and Scheutz, 2014) and as face-saving acts in parliamentary debates (Naderi and Hirst, 2018).

The closest semblance of our work is with Klüwer (2011, 2015), which builds upon the notion of face provided by Goffman (1967) and invents its own set of face acts specifically in the context of "small-talk" conversations. In contrast, our work specifically operationalizes the notion of the positive and negative face of Brown et al. (1987); Brown and Levinson (1978), which is well established in the Pragmatics literature and heavily acknowledged in the NLP community (Danescu-Niculescu-Mizil et al., 2013; Zhang et al., 2018a; Wang et al., 2012; Musi et al., 2018). Moreover, we focus on analysing the effects of face acts in a "goal-oriented" task like persuasion, where there is an explicit threat or attack on face as opposed to small-talk scenarios, where the goal is building rapport or passing the time. Thus our work can be considered to be complementary to the prior work of Klüwer (2011) and Klüwer (2015). It also enables us to draw insights from recent work in persuasion strategy to analyze face act exchanges in persuasion (Wang et al., 2019; Yang et al., 2019).

7 Conclusion and Future Work

In this paper, we present a generalized computational framework based on the notion of *face* to operationalize face dynamics in conversations. We instantiate these face act exchanges in the context of persuasion and propose a dataset of 296 conversations annotated with face acts. We develop computational models for predicting face acts as well as observe the impact of these predicted face acts on the donation outcome.

One important limitation of the current work is the assumption that all face acts have the same intensity/ranking. We seek to rectify this by separating the content and style of these face acts. We also wish to expand the current face framework to a more comprehensive politeness framework that incorporates notions of power and social distance between the interlocutors. We believe that our work may be extended to language generation in chatbots for producing more polite language to mediate face threats. Moreover, we intend to instantiate our proposed framework to other domains such as teacher/student conversations and other types of discourse such as social media narratives.

Acknowledgments

We thank the anonymous EMNLP reviewers for their insightful comments. We are also grateful to Xinru Yan, Meredith Riggs, and other members of the TELEDIA group at LTI, CMU. This research was funded in part by NSF Grants (IIS 1917668 and IIS 1822831) and from Dow Chemical.

References

Muhammad Abdul-Mageed and Mona T Diab. 2012. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, volume 515, pages 3907–3914. Citeseer.

Gordon Briggs and Matthias Scheutz. 2014. Modeling blame to avoid positive face threats in natural language generation. In *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, pages 157–161, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4742–4753, Hong Kong, China. Association for Computational Linguistics.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Erving Goffman. 1967. Interaction ritual: essays on face-to-face interaction.

H Paul Grice, Peter Cole, Jerry Morgan, et al. 1975. Logic and conversation. *1975*, pages 41–58.

Swati Gupta, Marilyn A Walker, and Daniela M Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *International Conference on Affective Computing and Intelligent Interaction*, pages 203–217. Springer.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406, Minneapolis, Minnesota. Association for Computational Linguistics.

Tina Klüwer. 2011. "i like your shirt"-dialogue acts for enabling social talk in conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer.

Tina Klüwer. 2015. Social talk capabilities for dialogue systems.

Gaurav Kumar, Rishabh Joshi, Jaspreet Singh, and Promod Yenigalla. 2020. AMUSED: A multi-stream vector representation method for use in natural dialogue. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 750–758, Marseille, France. European Language Resources Association.

Tuan Manh Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2019. A simple but effective bert model for dialog state tracking on resource-limited systems.

Geoffrey N Leech. 2016. *Principles of pragmatics*. Routledge.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2018. Changemyview through concessions: Do concessions increase persuasion? *Dialogue & Discourse*, 9(1):107–127.

Nona Naderi and Graeme Hirst. 2018. Using context to identify the language of face-saving. In *Proceedings of the 5th Workshop on Argument Mining*, pages 111–120, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Marilyn A Walker, Janet E Cahn, and Stephen J Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. *arXiv preprint cmp-lg/9702015*.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. 2012. Love ya, jerkface: using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue*, pages 20–29. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Steven R Wilson, Carlos G Aleman, and Geoff B Leatham. 1998. Identity implications of influence goals a revised analysis of face-threatening acts and application to seeking compliance with same-sex friends. *Human Communication Research*, 25(1):64–96.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018a. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Appendices

We present the hyper-parameters for all the experiments, their corresponding search space and their final values in Table S1. We also present additional details of our experiments below.
(i) Each run took at most 1 hour on a single Nvidia GeForce GTX 1080 Ti GPU.
(ii) We present the number of parameters for our model in Table S2.
(iii) All hyper-parameters were chosen based on the mean cross-validation performance.
(iv) Each validation split comprised 20% of all Donor conversations and 20% of all Non-Donor conversations, with the training split comprising the remaining 80%. All validation splits are mutually exclusive and exhaustive. The data splits are made available in the supplementary material.

Hyper-parameter	Search space	Final Value
learning-rate (lr)	1e-3, 1e-4	1e-4
Batch-size	-	1 conversation
#Epochs	50, 100	50
lr-decay	-	0.966
d_{h1}	300, 768	300
d_{h2}	300	300
d_{fc}	100	100
α	[0, 0.25,0.5,0.75,0.9,1.0]	0.75
L_D	BCE, MSE	MSE
Donation threshold	{0.001 - 0.999}	0.813

Table S1: Here we describe the search-space of all the hyper-parameters used in our experiments and describe the search space we used to find the hyper-parameters. All the experiments were run on a single 1080Ti GPU. d_{h1}, d_{h2} and d_{fc} represents the hidden dimensions of Utterance GRU, Conversation GRU, and the Face act classifier. α is the hyper-parameter used to combine the face-act loss and donation loss denoted by L_f and L_D respectively.

Model	Parameter Size
BiGRU	2.0M
BiGRU-f	2.1M
BiGRU-sf	2.2M
HiGRU	2.0M
HiGRU-f	2.1M
HiGRU-sf	2.4M
BERT-BiGRU	8.6M
BERT-BiGRU-f	9.2M
BERT-BiGRU-sf	10.4M
BERT-HiGRU	9.4M
BERT-HiGRU-f	10.2M
BERT-HiGRU-sf	11.5M

Table S2: Number of parameters for each model in our experiments

We show how donation probability changes for both Donors and Non-Donors in Figure S1.

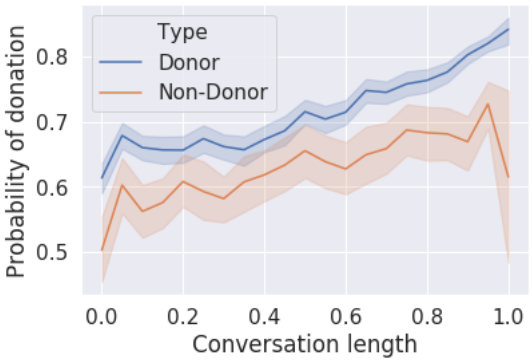

Figure S1: Plot of donation probability for Donor and Non-Donor conversations. We observe that the donation probability for Non-Donors at each step is lower than the corresponding probability for Donors.

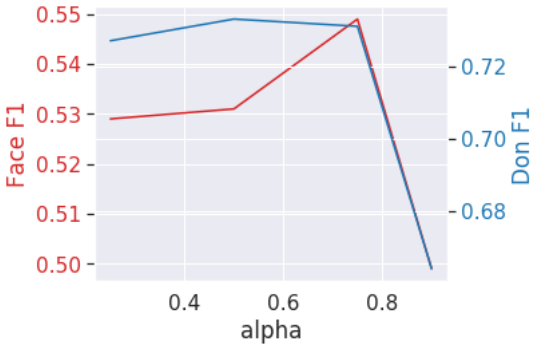

Figure S2: Plot showcasing the trade-off between donation F1 score and face-act F1 score for different values of α. As observed from the plot, a value of 0.75 achieves the highest overall score.

We also show the dependence of alpha on the overall F1 scores of Face acts and donation probability in Figure S2, which justifies our choice of $\alpha = 0.75$

Face act	Persuader (ER)	Persuadee (EE)
SPos+	(i) ER praises/promotes the good deeds of STC (ii) ER shows her/ his involvement for STC	(i) EE states her preference for other charities (ii) EE states that she does good deeds
HPos+	(i) ER appreciates/praises EE's generosity or time (ii) Incentives EE to do a good deed. (iii) Empathize/ agree with EE	(i) EE shows willingness to donate to discuss the charity (ii) EE acknowledges the efforts of STC. (iii) Empathizes/ agrees with ER
SPos-		(i) EE apologizes for not donating
HPos-	(i) ER criticizes EE	(i) EE doubts/ questions STC or EE (ii) EE is not aware of STC
SNeg+		(i) Rejects donation out-right (ii) Cites reason for not donating at all or not donating more.
HNeg+	(i) ER provides EE convenient ways to donate. (ii) ER apologizes for inconvenience/ intrusion. (iii) ER decreases the amount of donation.	
HNeg-	(i) ER ask's EE's time/ permission for discussion. (ii) ER asks EE for donation. (iii) ER asks EE to donate more.	(i) EE asks ER questions about STC.

Table S3: Instantiating predicates corresponding to the different face acts in the context of persuasion.

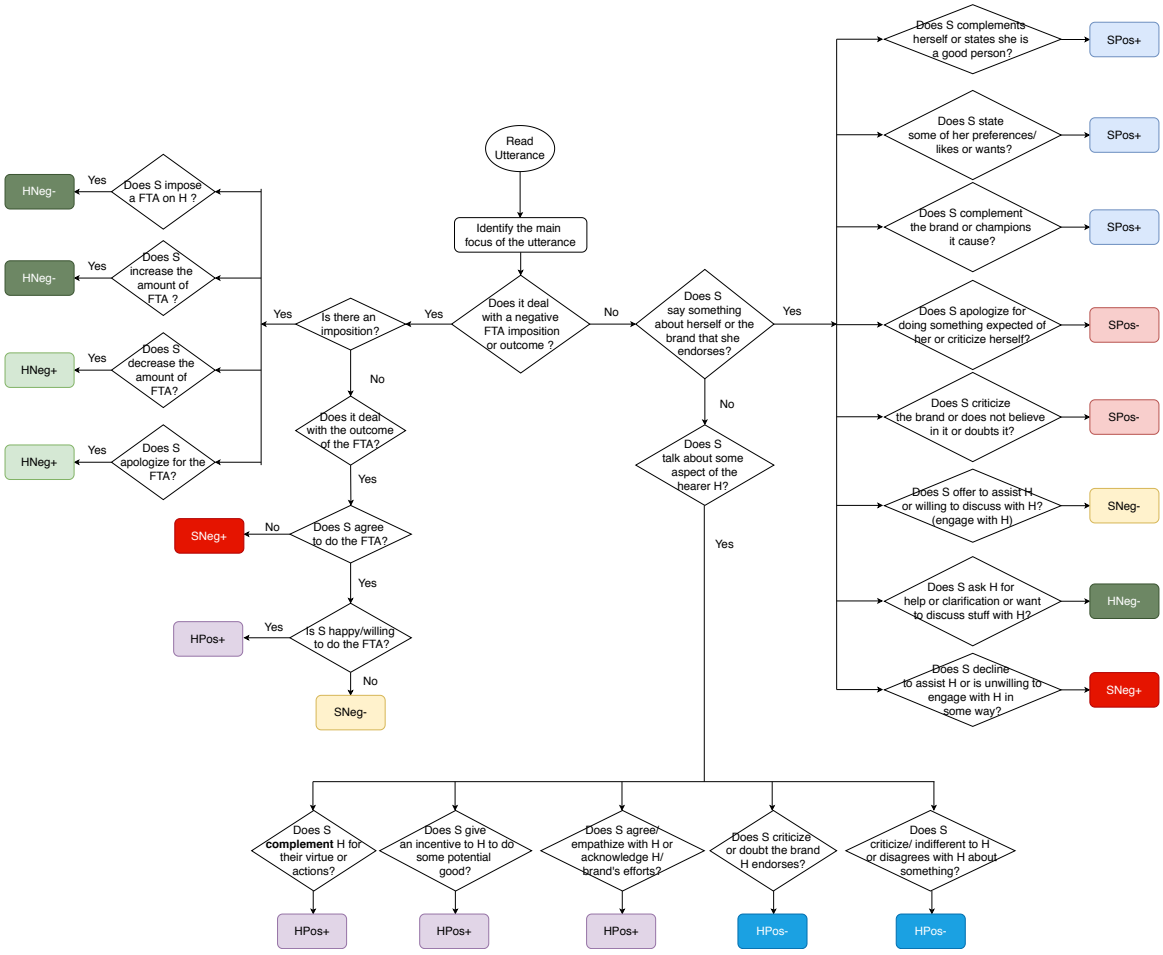

Figure S3: Flowchart outlining the annotation framework employed for labeling face acts in persuasion conversations.