

Multimodal Modelling of Persuasion Strategies in Social Deduction Games - An Undergraduate Research Thesis

Aryan Pariani

apariani3@gatech.edu

1 Introduction

Out of all possible mental states that a human can ascribe to oneself or to others (Premack and Woodruff, 1978), a human may often need to deduce the belief state of another human by using persuasive techniques. Humans tend to use a unique set of characteristic behaviors or strategies, verbal and non-verbal, in social interactions when they want to lie to, hide the truth from, or persuade someone to believe their statements or to engage in some behavior. These strategies have the potential desired effect of instilling new or changing previously held beliefs of the intended audience, hence affecting their belief state (Lonigro et al., 2017). However, the verbal and non-verbal cues for such persuasion strategies are arguably more implicit than for other social phenomena. Moreover, data for the use of persuasion strategies in real-world social interactions is not readily available because such strategies are usually utilized in private or intentionally covert social interactions. These issues make it difficult to collect reliable conversational data from realistic social interactions in open-ended, uncontrolled domains for the study of persuasion. Hence, the need arises for designing automated social deduction systems with the capability of recognizing the use of persuasion strategies in human social interactions and inferring the belief state of the target audience. Such systems can accelerate the study and detection of the social dynamics of persuasion and collaborative decision-making processes in realistic conversations.

Inspired by previous works like (Peskov et al., 2020) that focuses on the study of deception in *Diplomacy*, the aforementioned problems encourage us to turn toward conversations in multiplayer social deduction games like *One Night Ultimate Werewolf*, where the overarching self-serving motive of every player is to win the board game by any means which are within the scope of the game rules that call for them to employ deceptive and persuasive tactics. We can reliably collect realistic, rich dialog data from conversations in such game-based domains so that we can conduct an effective computational study of linguistic persuasion strategies without privacy concerns, that can be applicable to other domains characterized by persuasion in social interactions. Peskov et al. (2020) present a self-annotated dataset that does not allow for identification of specific linguistic strategies of deception beyond the detection of lies and their dataset is collected from a constrained environment where players play an online version of *Deception* with each other. In contrast, Chawla et al. (2021a) investigate identifying specific negotiation strategies in dialogues between pairs of participants acting as campsite neighbors negotiating with each other for additional camp resources, but their corpus consists of more artificial dialogues collected from a more heavily constrained online environment with an imagined scenario and set of items as the basis for negotiations. Most such previous works studying persuasion, deception or negotiation strategies capitalize solely on linguistic verbal cues in text. However, the multimodal nature of social interactions warrants also analyzing non-verbal cues like gestures in the visual modality for a more comprehensive study of such social phenomena. Furthermore, the indirectly observable belief states of participants are not

captured in datasets for most related previous works.

Therefore, to fuel reliable computational studies of the language of persuasion, we introduce a large, rich corpus of videos and conversation transcriptions of recorded in-house playthroughs of the role-based social deduction board game *One Night Ultimate Werewolf* among familiar players mostly consisting of close friends and family members who engage in naturalistic social interactions and behaviors without any inhibitions. This dataset is presented with annotations of identified strategies for every utterance by a speaker that exhibits the annotated strategies from a taxonomy of predefined 6 persuasion strategies commonly used in social deduction games. The dataset is also annotated with annotations of the final voting decision of every player in every game playthrough, that serve to model the belief state of every player, along with annotations of the players' actual game-based roles.

As a step towards the realization of aforementioned practical automated social deduction systems, we will present two primary prediction tasks for the study of persuasion in social deduction games, as well as baseline models and results for the same. Our first primary task is the task of predicting the possible persuasion strategies exhibited in an utterance by a player, based on the text of the utterance and immediately previous conversational context of varying lengths. A key subtask of this task incorporates multimodal (linguistic and visual) cues for strategy prediction, requiring additional input in the form of the corresponding extracted video frames spanning the duration of the utterance. Our second primary task is that of predicting the belief state of every player at the end of a game playthrough in the form of predicting their final voting decision, based on players' dialogues, their game roles, and the persuasion strategies employed by every player in every utterance.

2 Related Work

[Chawla et al. \(2021a\)](#) present CaSiNo, a rich corpus of negotiation dialogues between pairs of participants acting as campsite neighbors negotiating for additional camping resources. They also present a multi-task model architecture based on a common BERT-base encoder for the task of prediction of 9 possible negotiation strategies in such dialogues. The language used for negotiation has a lot of similar linguistic aspects and styles to the language used for persuasion, albeit the striking difference between negotiation and persuasion is the motivation and the possible outcomes, that is further delineated in the differing contexts of campsite negotiations and conversations in social deduction games. So we use a similar annotation schema as in this work, with 6 identified persuasion strategies specific to persuasive tactics commonly used in such social deduction games. Data for CaSiNo is collected from a constrained environment based on a digital chat interface for an imagined campsite negotiation scenario, and [Chawla et al. \(2021a\)](#) recognize that this does not promote a reliable study of free-form, open-ended negotiation dialogues that truly resemble realistic scenarios. On the other hand, our data is collected from natural conversations in social deduction games amongst friends and family, where the rules of the game naturally incentivize players to employ strategies of persuasion in their language as a means to win the game, without any reservations. Similarly to CaSiNo, a large majority of utterances in our dataset are annotated with a "No Strategy" label. So we propose an adaptation of the BERT-based multi-task learning framework as presented in this work, which allows for parameter sharing in order to address the problem of high asymmetry in our dataset.

[Peskov et al. \(2020\)](#) present a dataset of dialogues for the study of deception in language, collected from players playing an online version of *Diplomacy*, a social deduction board game of strategic conquest of territories by countries based on "diplomatic phases" often marked by persuasive or deceptive language. This work presents a model trained on this dataset for learning

to detect lies and what can be perceived as lies in such dialogues in *Diplomacy*. Although lying is a key characteristic in deceptive language, the dataset does not allow for the study of the more nuanced strategies and tactics that characterize the language of deception, like the negotiation strategies identified by [Chawla et al. \(2021a\)](#) or the persuasion strategies we identify for our strategy prediction task for similar social deduction games including *One Night Ultimate Werewolf*. According to [Peskov et al. \(2020\)](#), previous work identifies that it is a difficult and intractable task to collect data from the use of deception in language in real-world scenarios where deception is used much more discreetly for obvious reasons, rather than in short-term artificial contexts like games. This attests to the capability of games to be used as powerful test-beds to further AI research on social phenomena that are too covert in real-world scenarios to collect ample and reliable data from.

[Fornaciari et al. \(2021\)](#) present methods of incorporating conversational context information for detecting deception in Italian criminal proceedings with the help of language models, and an elaborate evaluation of their results with respect to the added value of BERT-based models and contextual information. Based on their results, [Fornaciari et al. \(2021\)](#) hypothesize that the best approach to fully leverage contextual information for similar prediction tasks with models based on pre-trained models like BERT is to combine such models with attention mechanisms to pinpoint the stylometric cues that characterize such strategies. This further validates our proposal to use the BERT-based models for our strategy prediction task. [Peskov et al. \(2020\)](#) hypothesize that the lack of observed performance improvement on account of the BERT encoder in their results can be attributed to their *Diplomacy* dialogue dataset being too small. However, [Fornaciari et al. \(2021\)](#) provide an alternate hypothesis based on their contrasting promising results with BERT-based models that the *Diplomacy* dialogues collected by [Peskov et al. \(2020\)](#) from an online version cover a narrow and restricted range of topics diminishing the possibility of leveraging BERT's broad knowledge for additional value. This exposes a potential limitation of using BERT-based models for our dataset for similar social deduction games where topics of conversation are likely to be less varied and more game-specific, albeit the major *Ego4D* subset of our dialogue data was collected by [Grauman et al. \(2022\)](#) from traditional game-night settings accompanied by spontaneous and naturalistic social behaviors.

[Chawla et al. \(2020\)](#) present a novel task of predicting negotiation outcomes with respect to the most likely final agreed price, based on a fraction of buyer-seller negotiation dialogues and task-specific features like product category, the buyer's target price. They propose a relatively straightforward approach to addressing the problem of intractability of training pre-trained language models like BERT on data with linguistic and non-linguistic task-specific features in the embedding space, by concatenating via a separator token all dialogue utterances with a "scenario" of multiple sentences of a predefined structure representing all the non-linguistic task-specific features to form the input for BERT. While our task entails a much more complex representation for the non-linguistic input with embeddings for multiple players, their true game-based identities and sequences of persuasion strategies, this serves as a good foundation to build the input for BERT for our task of predicting the voting decision for every player, with modifications to account for the increased complexity of our feature representation. This work also presents an interesting probing technique with a simple neural network to predict language features on both the buyer's and the seller's side that mainly contributed to the final agreed price prediction by the fine-tuned BERT model, using classifier tokens before and after training the model for the task. We can apply this technique for our social deduction task as well to predict not only the language features but also the persuasion strategies on both the voter's and the predicted voted player's sides that mainly contribute to the voting decision prediction.

[Chen and Yang \(2021a\)](#) present a taxonomy of very generalizable persuasion strategies and

a large multi-domain text corpus of dialogues along with strategy annotations, spanning three domains. While the presented generic taxonomy applicable to multiple contexts and domains sets this work apart from previous related studies on persuasive language and our work, our goal of understanding persuasive language is specific to the domain of social deduction game settings among close acquaintances, where broad persuasion strategies in the presented taxonomy like "Politeness" and "Scarcity" are much less likely to be employed. On the other hand, the more frequent outright accusations and the claims of players' own identities due to the emphasis on deducing hidden game-based identities of players in this domain warrant incorporating more specific persuasion strategies in our taxonomy like "Accusation" and "Identity Declaration" for our task objective. Nonetheless, the other strategies in the taxonomy presented by [Chen and Yang \(2021a\)](#) most notably serve as an inspiration for a handful of the identified persuasion strategies in our taxonomy, especially "Evidence" and "Call for Action".

[Girlea et al. \(2016\)](#) present a task of deception detection in *One Night Ultimate Werewolf* - one of the two social deduction games represented in our dataset - by predicting which players possess the game-based role of the werewolf. [Girlea et al. \(2016\)](#) present simple approaches of basic machine learning algorithms that do not involve any neural architectures or language models, like logistic regression, SVM and decision trees, based on 19 handcrafted psycho-linguistic features extracted from every player's utterances in the entire game, relating to stylometric features characterizing deceptive language based on previous literature. The presented task of werewolf role prediction entails similar inputs and feature representations as our social deduction task of predicting vote outcomes, although we incorporate utterances in their raw form instead of handcrafted linguistic features in addition to non-linguistic features of players' roles, positional embeddings, and the persuasion strategies they used throughout the game as our input to a complex BERT-based language model. Regardless, the results achieved by [Girlea et al. \(2016\)](#) are very promising given the simplicity of their approaches, which give good grounds for expecting that our BERT-based language models would perform even better, with a potential information gain in our more complex and arguably more cohesive input feature representations compared to handcrafted features.

[Grauman et al. \(2022\)](#) present *Ego4D*, a massive high-quality egocentric video dataset of hundreds of scenarios and activities spanning multiple domains. One of the two major sources of our original social deduction game video dataset to fuel our tasks is the subset of *Ego4D* collected by Georgia Tech of egocentric (first-person view) videos of friends and family members playing social deduction games of *One Night Ultimate Werewolf* or *The Resistance: Avalon* with each other and partaking in eating and drinking behaviors in their own homes in traditional game-night settings. As pointed out by [Grauman et al. \(2022\)](#), the realistic game-night setting promoted more free-flowing conversations and social behaviors richly captured in both the visual modality from the first-person perspective of every player and the auditory modalities, inclusive of realistic verbal and non-verbal, visual cues of social phenomena. Thus, the *Ego4D* subset allows us to obtain very rich textual transcriptions of realistic and free-flowing conversations in playthroughs of social deduction games for the study of social phenomena, setting our dataset of game dialogues sourced from *Ego4D* for the study of persuasion apart from other dialogue datasets for the study of similar social phenomena, like the CaSiNo corpus collected by [Chawla et al. \(2021a\)](#) or the *Diplomacy* game dialogue dataset collected by [Peskov et al. \(2020\)](#). Furthermore, with access to the rich visual modalities in high quality from the first-person perspective of every player, we take advantage of the egocentric video features for our sub-task of the strategy prediction task, aimed at using a multimodal approach for more accurate prediction of persuasion strategies by incorporating the potential information gain from non-verbal cues, on top of the utterance-level verbal cues.

Ta et al. (2022) present a collected dataset of message threads from the *r/ChangeMyView* subreddit, an online public forum for Reddit users to initiate debates with each other to try and change each others' views, and the tasks of identifying the handcrafted linguistic features with the most influence on the perceived persuasiveness of such messages. An interesting argument provided by Ta et al. (2022) is that their self-annotated dataset serves as a more effective and reliable dataset for the study of linguistic persuasion because persuasiveness in *r/ChangeMyView* is evaluated in terms of completely changing another's beliefs from their previous held beliefs, rather than simply coercing another to engage in a specific behavior or change their behavior without necessarily challenging their inherent beliefs. Changing one's usually previously long-standing beliefs is highlighted as a more effective and reliable metric of persuasiveness because it entails a more sophisticated and calculated use of persuasion strategies, supported by previous literature. This sets their dataset apart from other similar datasets like the multi-domain dataset in (Chen and Yang, 2021a) or the CaSiNo corpus in (Chawla et al., 2021a) involving domains where linguistic persuasion or negotiation strategies are used in the context of convincing others to engage in more surface-level behaviors like making or borrowing loans, or lowering personal stakes in a negotiation. Our dataset of dialogues are sourced from playthroughs of social deduction games where persuasion strategies are used for strengthening or changing players' beliefs about the true game-based roles held by players in order to win, naturally enforced by the rules and the maintenance of a high degree of ambiguity about true roles of players in a game. Thus, our dataset similarly promotes a reliable study of most effective persuasion strategies of changing one's beliefs.

Pérez-Rosas et al. (2015) present one of the first few works on multimodal detection of deception in realistic scenarios in the textual, verbal and visual, non-verbal modalities. Pérez-Rosas et al. (2015) implement three types of handcrafted verbal features. Engineering and extracting such handcrafted linguistic features from transcripts provides a focused yet trimmed linguistic feature representation that mollifies the need for excessive training data for simple classifiers like the ones proposed here to learn relatively quickly from. Neural architectures like the ones that we propose, require more training data to learn linguistic features of deception from transcripts in their raw form without any feature engineering, but can capture and learn from the more subtle and complex linguistic cues of deception from utterance-level data that cannot be captured by handcrafted features or learned by the proposed simple models. Pérez-Rosas et al. (2015) also propose a scheme to manually label video data with handcrafted visual, non-verbal features of deception including hand gestures and facial expressions in terms of the most frequently observed cues. This proposed scheme of handcrafted feature engineering for the visual cues is not only time-consuming and tedious, requiring further annotation work and overhead costs, but also unreliable because annotators are required to make subjective, estimated guesses of the most frequent visual cues. For our multimodal strategy prediction subtask, we do not propose any handcrafted feature engineering techniques for the visual cues, but instead propose first learning the most relevant visual feature representations with pre-trained video transformers from clips of video frames in their raw form corresponding to an utterance. During training with baseline BERT-based models, we propose a simple concatenation of the output linguistic feature representation from BERT with this learned visual feature representation, which forms the input for a subsequent fully connected layer.

Morio et al. (2019) present annotations on conversational persuasion data extracted from the same domain as the dataset in (Ta et al., 2022) - the *r/ChangeMyView* subreddit - with combinations of 5 types of identified "elementary units" (EUs) or persuasion strategies. They also present baseline neural models for a sequence tagging task to identify the employed persuasion strategies at the sentence-level and the post-level, along with positional information of the

identified strategies. Unlike Ta et al. (2022) for the same *r/ChangeMyView* domain, Morio et al. (2019) do not advocate for handcrafted feature engineering and instead propose neural models for directly processing sentences or posts in their raw form to capture more complex feature representations. Morio et al. (2019) elaborate on their supporting analysis of the positional roles of EUs as persuasion strategies in themselves at the post-level, indicating with strong evidence showing that specific persuasion strategies can be strategically used earlier or later in a single post, affecting the persuasiveness of the post and in turn the likelihood of the post changing the views of the original poster. This is a valid argument for extending the persuasion strategy classification task to a sequence tagging task, warranted by the *r/ChangeMyView* domain where a greater emphasis is laid on the post-level analysis rather than the sentence-level analysis because entire posts are marked as having changed the views of the original poster rather than single sentences. On the other hand, our domain of conversational data in social deduction games, is composed of a larger number of more granular "utterances" of fewer and shorter sentences in every game's dialogue transcript, where positional information of persuasion strategies within every utterance by a speaker is much less relevant to our task, supporting our decision to consider a persuasion strategy prediction task only at the utterance-level without any sequence tagging.

Chiu et al. (2021) present multiple experiments with GPT-3, the largest Transformer language model built by OpenAI, in different settings of zero-, one-, and few-shot learning for one and both of the sexist and racist categories, for the tasks of detection of hate speech and classification of identified hate speech as racist or sexist. Chiu et al. (2021) notably observe from their results that sub-par accuracy and F-1 scores were reported for the tasks using GPT-3 in the zero-shot, one-shot and single-category few-shot learning settings whereas impressive scores were reported in the multi-category few-shot learning setting. They attempt to find evidence to explain these observations by showing that GPT-3 tended to incorrectly classify a number of examples of online comments with intentionally or unintentionally misspelled words that retain the same semantic context in the eyes of a human. This leads Chiu et al. (2021) to make the conclusion that this exposes a potential downside of using GPT-3 in that it fails to recognize more current patterns that might emerge in the language data, if those patterns were not captured in GPT-3's pretrained knowledge base in the time before GPT-3 was pre-trained, with unique trends of linguistic usage arising very frequently over time, especially in racist and social contexts on online platforms. Comparatively, our conversational data is manually transcribed from videos, ensuring no misspellings or typos in the language. Moreover, conversational themes in this domain are not likely to be focused on more current affairs that can cause GPT-3 to mis-classify utterances. Therefore, we propose similar approaches for our strategy prediction task, with the GPT-3 model in zero-shot, one-shot and few-shot settings.

3 Dataset

3.1 Data collection

Our dataset is curated from two major sources of videos that capture complete play-throughs of social deduction games. In both sets of videos, the games are played by groups of friends of family members in traditional game-night settings while engaging in naturalistic group conversations and various social behaviors. To ensure that our dataset can be used for training models for the game outcome deduction task, all videos in both sets have fully visible game outcomes.

Ego4D: Our first source is a subset of the publicly released Ego4D Social benchmark dataset, capturing 7.3 hours of aggregate footage of 40 complete games of One Night Ultimate Werewolf and 8 games of The Resistance: Avalon. Although the Ego4D dataset features egocentric first-

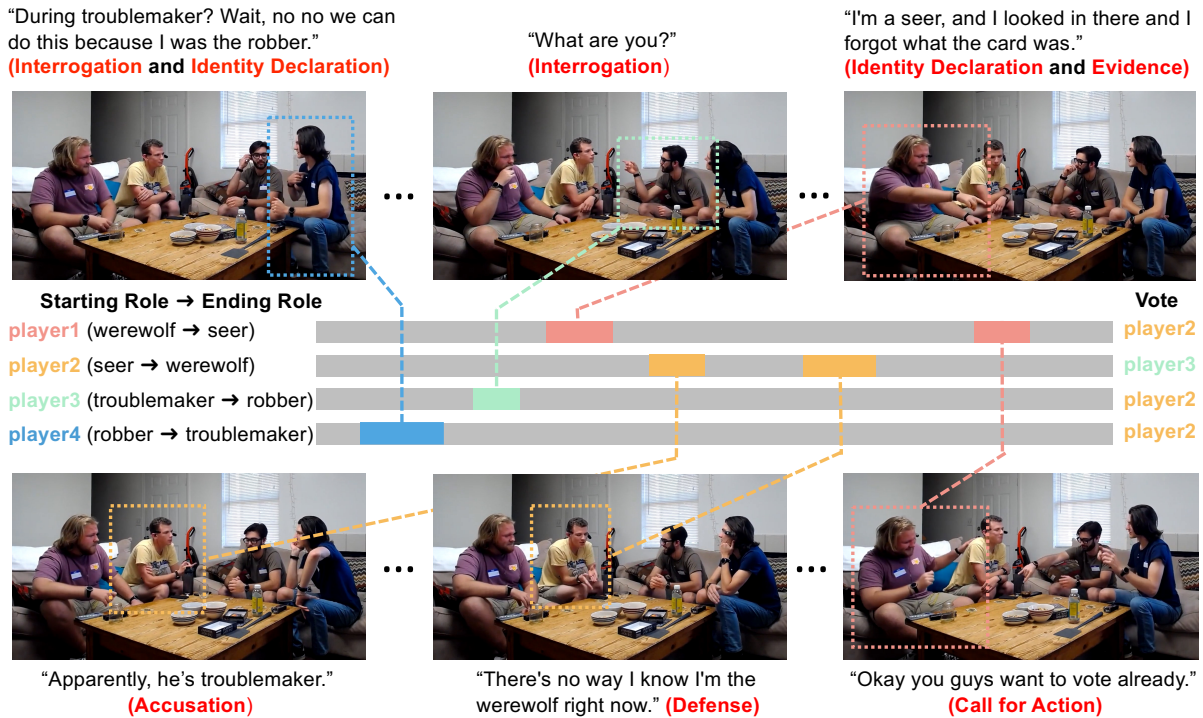


Figure 1: Demonstration of the six persuasion strategies included in our dataset and the corresponding video of a group playthrough of One Night Ultimate Werewolf. This figure also appears in a publication of this work (Lai et al., 2022).

person videos from the perspective of every player, we leverage only the third-person videos from this dataset. This ensures that every player is visible in a frame for generating transcripts and visual experiments. Since the Avalon subset is of a relatively small size, we use Avalon game data only for the purpose of testing our models’ generalization ability when applied on data from different kinds of games.

YouTube: We leverage the public YouTube Data API V3 to programmatically scrape the web for candidate YouTube videos. By making API GET requests from the YouTube Data API V3 (Developers), we collect, filter, and download task-relevant video files for around 400 YouTube videos of group game play-throughs of One Night Ultimate Werewolf, with search keyboards like “one night ultimate werewolf” and “ultimate werewolf”. We finally shortlist a dataset comprising 14.8 hours of video footage spanning 151 clips of completed games in similar settings as Ego4D.

3.2 One Night Ultimate Werewolf

Since majority of the video data captured in our dataset are based on playthroughs of One Night Ultimate Werewolf, it is imperative that we provide a description of the game. In this game, there are two teams competing with each other- Villagers and Werewolves. Their team roles are determined by the cards they are dealt. At the start of the game, nobody knows what role the other players have been dealt with. This status continues for most players and only some special role players are asked to identify themselves without the others’ knowledge. One player is the ‘Announcer’ and this player calls out the instructions for the others to follow. Upon the order by the Announcer, all the players close their eyes (this is the ‘night phase’) and depending on the role they have been assigned, some players are asked by the Announcer to take some actions in the night phase (eg. the “Robbers” are asked to open their eyes and swap their cards, the ‘Werewolves’ to look for their fellow werewolves etc.). During the ‘Night Phase,’ some players may have their roles changed without their knowledge, while some special role players are in the loop. Once the Announcer asks all players to open their eyes, the negotiations and

Label	Example	Ego4D			YouTube		
		Count	AUL	α	Count	AUL	α
Identity Declaration	“I’ll just come out and say I was a villager, so I have no idea what’s going on.”	293	9.87	0.90	1066	10.43	0.87
Accusation	“So James might be the werewolf.”	669	11.28	0.74	2830	11.06	0.67
Interrogation	“Who did you rob?”	695	7.56	0.80	3407	7.66	0.90
Call for Action	“We shouldn’t vote to not kill anyone. And then there could also be no werewolf.”	236	9.99	0.78	1163	9.53	0.71
Defense	“I think that you accused me of being a Werewolf very quickly.”	570	10.04	0.62	2696	9.75	0.80
Evidence	“If you swapped these two, he is not the werewolf.”	489	11.45	0.75	1740	9.80	0.60

Table 1: Utterance-level persuasion strategy annotations. AUL refers to the average utterance length in terms of the number of words in an utterance and α refers to Krippendorff’s alpha. This table also appears in a publication of this work (Lai et al., 2022).

persuasion strategies come into play. Using our six proposed persuasion tactics, Villagers try to discover who are the Werewolves, who, in turn, use their negotiation and persuasive skills to refute accusations by Villagers and try to mislead them so that their true identity remains hidden. Players are then asked to vote out one other player whom they suspect to be the Werewolf. The player with the maximum votes is declared ‘dead’ and quits the game. The game ends when at least one Werewolf player is voted out-this is declared a win for the Villagers. However, if a Villager is voted out and at least one player is a Werewolf and no Werewolf is voted out/ killed, then the Werewolf team wins.

3.3 Data Annotation

Dialogue Transcription: Using the automatic transcription service offered by rev.com, we upload videos and specify names and roles of speakers in every game. In this way, the tool generates transcripts with every utterance labelled with the speaking player’s name and the timestamp. Upon visual inspection, annotators carefully align the generated transcripts to the videos, making corrections where necessary.

Persuasion Strategy Annotation: We now introduce the persuasion strategy taxonomy and corresponding definitions adopted in our paper. Persuasion and negotiation strategy taxonomies from related works (Chawla et al., 2021b; Yang et al., 2019; Chen and Yang, 2021b) guided our proposed persuasion strategies enumerated as follows.

- **Identity Declaration:** *State one’s own role or identity in the game.* Such claims of one’s own role or identity are enforced frequently in persuasive conversations in our dataset due to the nature of social deduction games.
- **Accusation:** *Claim someone has a specific identity or strategic behavior.* Chawla et al. (2021b) categorises a similar proself tactic as "Undervalue-Partner".
- **Interrogation:** *Questions about someone’s identity or behavior.*
- **Call for Action:** *Encourage people to take an action during the game.* This strategy definition is derived from the prosocial tactic of "Coordination" for persuasion as defined by Chawla et al. (2021b).
- **Defense:** *Defend oneself or someone else against an accusation or defend a game-related argument.* For instance, a player’s utterance would be categorized as a "Defense" strategy

when the player uses persuasive tactics like citing credentials to establish trust or rationalize their own or another player’s choices. Similar strategies have been referenced in previous works under the categories of "Credibility" (Wang et al., 2019) and "Commitment" (Yang et al., 2019).

- **Evidence:** *Provide a body of game-related fact or information.* This strategy is also described under the term "Concreteness" (Yang et al., 2019)

Similar to (Chawla et al., 2021b), we label and analyze persuasion strategies at the more granular utterance level. This helps to easily identify and analyze linguistic cues of persuasion. By analyzing utterances in context and by correlating these with visual cues, we try to understand how different persuasion strategies are used in different situations and with different audiences. We adapted the annotation interface employed by (Hayati et al., 2020) for our annotators. The annotators were first tasked with annotating a common set of utterances. The results were then analyzed for the degree of agreement amongst the three annotators by the metric of the Krippendorff’s alpha (Krippendorff, 2018). The discrepancies among the annotators’ versions were then discussed together after which we strategized a standard set of rules to refer to whenever such discrepancies arise during the annotation process. This process was repeated until a Krippendorff’s alpha value greater than 0.6 was attained for each strategy (Refer to Table 1). Following this training, each of the annotators annotated the remaining transcripts in the dataset, working independently of each other.

Game Dynamics: Much of the video footage in our dataset revealed participants playing more than one game per clip. This necessitated annotating the start and end times of every game. The annotators also label each player’s starting role, ending role, and voting decision. These annotations are essential for the game outcome deduction task.

Annotation Statistics: Of the 5,815 utterances transcribed from the Ego4D data, more than 49.2% are annotated with "No Strategy" vis-a-vis only 37.9% of the 20,832 utterances in the YouTube subset. We hypothesize that this difference arises because Ego4D games were played in a more intimate social setting whereas players in YouTube games tend to be more focused on game-based social interactions.

4 Strategy Prediction Task

Humans rely on both verbal and visual clues to assess the speaker’s intent. Besides these cues, utterances need to be assessed within the context of past utterances as well. Keeping these points in mind, we propose that our models will be able to make predictions effectively when the visual and verbal cues are correlated in tandem. In light of this, we examine a given utterance and its corresponding visual in our experiments with the aim of predicting persuasion strategies exhibited.

We encode the input utterance to extract the corresponding word embeddings using pre-trained large language models such as BERT (Devlin et al., 2019; Liu et al., 2019) and RoBERTa (Liu et al., 2019). As for extracting the visual feature representation from the video segment, we use a vision transformer (Fan et al., 2021) as a video encoder. The concatenation of these visual and textual representations forms the basis for classifying the speaker’s persuasion strategies. We also conduct experiments for the strategy prediction task by incorporating context of previous utterances in the input.

4.1 Methodology

Modeling with single utterance: As pointed out by (Chawla et al., 2021b) and as exemplified in Figure 1, an utterance from naturalistic conversations can exhibit several persuasion or

negotiation strategies. So we approach this task primarily as a binary classification problem for each persuasion strategy. An utterance is labelled with "No Strategy" if negative labels are predicted for each strategy. We introduce a *Base* model, the pre-trained language model that we fine-tune for this task.

Modeling with Context Embedding: For some persuasion strategies defined in our taxonomy, especially "Defense" and "Evidence", often the verbal cues most likely lie in immediately preceding utterances in the game rather than in the current utterance itself. In the case of "Defense" and "Evidence", the speaker of the current utterance may defend their position against possible accusation in a prior utterance or provide evidence to support a claim (Fornaciari et al., 2021). To tackle such cases, we also present a model denoted as *Base + C*, incorporating context of prior utterances for each utterance in the input.

Modeling with Video Representation: In *Base + V*, we utilize visual, non-verbal cues from corresponding video segments, in addition to verbal cues of current utterance, to improve modelling of persuasion strategies. With the help of a fine-tuned pre-trained vision transformer, we extract visual feature representation of video segment, which is combined with the textual feature representation of the utterance. This aggregate of multimodal features is then fed into a classification layer, as shown in Figure 2. We call this model "Base + V". Dropout is adopted to mitigate overfitting. During training, the vision encoder is frozen and all other modules are trained end-to-end.

Modelling with Video and Context Encodings: We finally present the *Base + C + V* model, utilizing both video features and context cues for persuasion strategy prediction. This model features a simple late fusion layer that aggregates the context representation output of *Base + C* and the visual representation output of *Base + V* at the decision level.

4.2 Model Details

BERT and **RoBERTa** were found to be the best-performing models by (Dimitrov et al., 2021) for the detection of persuasion techniques in multimodal data like memes. Such promising results support our proposal to employ the pre-trained transformers of **BERT** (Devlin et al., 2018) implemented with `bert-base-uncased` and **RoBERTa** (Liu et al., 2019) implemented with `roberta-base` from Huggingface (Wolf et al., 2020) as baseline models for the text encoder. For the video encoder, we use the MViT-B-24 model (Fan et al., 2021) pretrained on Kinetics-400 (Kay et al., 2017).

Moreover, as an additional baseline, we also implement **MT-BERT**, a multi-task framework based on a common BERT-base encoder (Chawla et al., 2021b) for simultaneous prediction of multiple persuasion strategies rather than binary classification of a single strategy. This model features a task-specific self-attention mechanism, which allows for parameter sharing between the different prediction tasks in order to address the problem of imbalance of non-strategic and strategic utterances in our dataset. Six heads are set for the six strategies, while the backbone is shared. We add the fusion layer right after the last self-attention layer in each head in the same way as independent models. Context and video features are also incorporated into MT-BERT similarly.

Base Model: In these models, only the text of the current utterance forms the input \mathcal{T} . \mathcal{T} is input into the BERT-based or RoBERTa-based text encoder ϕ the output of which is then fed into a classifier to obtain the strategy prediction.

Base + C: The context C is formed by the concatenation of k previous utterances C_1, C_2, \dots, C_k , alternating with $[EOS]$ tokens as delimiters. The final input \mathcal{T} of the Base + C model is formed by the concatenation of this context C with the current utterance U , separated by a $[SEP]$ token.

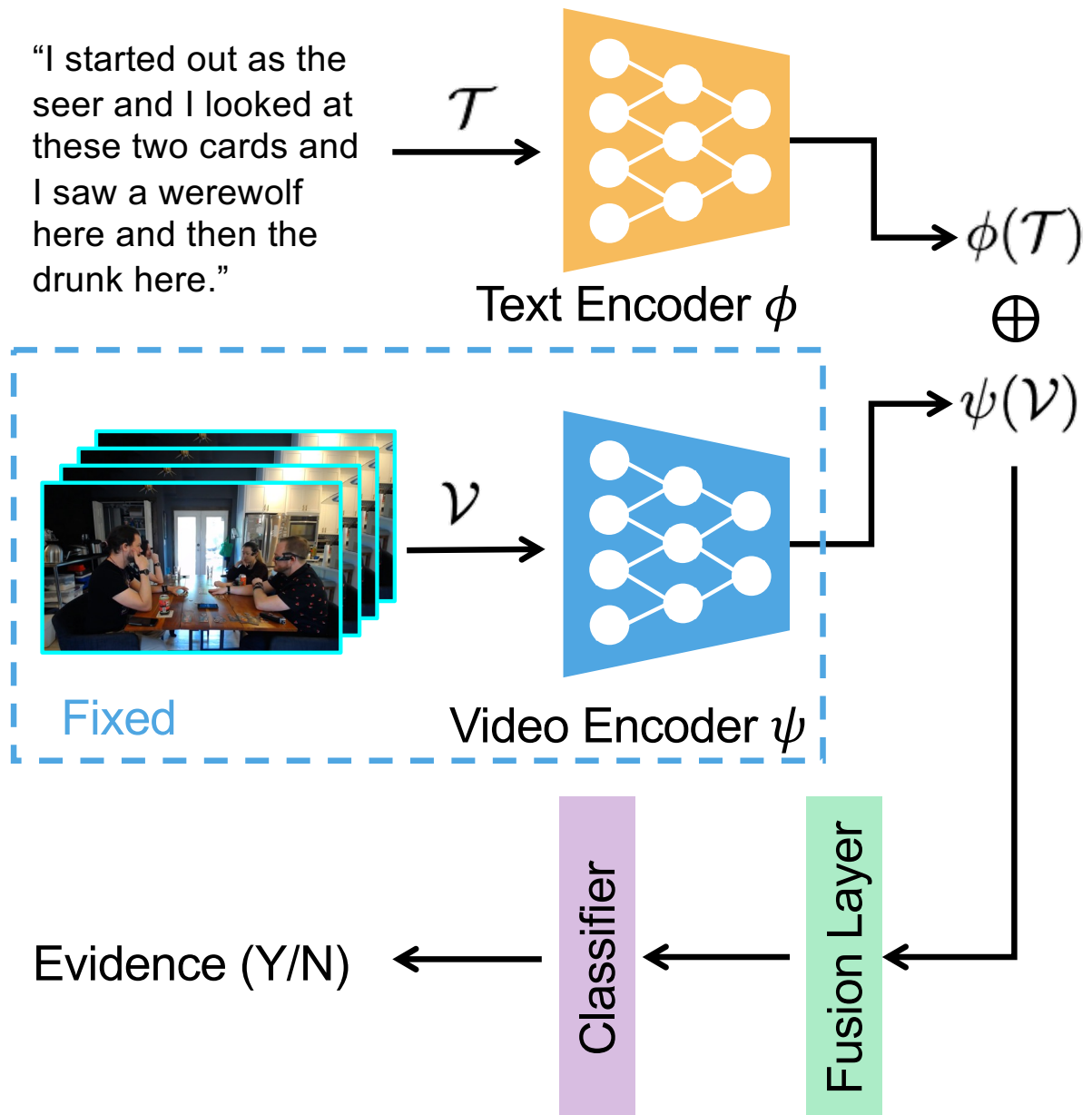


Figure 2: Architecture of the independent model for each strategy. We fix the parameters in the video encoder and train the other modules end-to-end. \oplus denotes the concatenation of two feature representations. This figure also appears in a publication of this work (Lai et al., 2022).

Formally, input \mathcal{T} is represented as below:

$$C = C_1 [\text{EOS}] C_2 [\text{EOS}] \cdots C_k, \quad (1)$$

$$\mathcal{T} = C [\text{SEP}] U. \quad (2)$$

Base + V: Visual representation $\psi(\mathcal{V})$ of the video segment \mathcal{V} corresponding to the current utterance is extracted by the video encoder ψ . During training, the parameters of the video encoder are fixed and the concatenation of $\psi(\mathcal{V})$ and text representation $\phi(\mathcal{T})$ of the current utterance is fed into a fusion layer with a linear mapping function W_F^T and an activation function $\text{Tanh}(\cdot)$. The prediction logits are finally obtained by applying a linear classifier W_P^T , as follows:

$$\text{logits} = W_P^T \cdot \text{Tanh} (W_F^T \cdot (\phi(\mathcal{T}) \oplus \psi(\mathcal{V}))), \quad (3)$$

where \oplus denotes vector concatenation.

Base + C + V: Here, we late fuse the softmax output P_C of Base + C with softmax output P_V of Base + V, formulated as below:

$$P_{C,V} = (1 - \lambda)P_C + \lambda P_V, \quad (4)$$

where λ is a scalar that balances P_C and P_V .

4.3 Training Details

1. **Loss function:** Cross-Entropy Loss
2. **Learning Rate:** $\{1e - 5, 3e - 5, 5e - 5\}$
3. **Batch Size:** $\{16, 8\}$
4. **Optimizer:** AdamW (Loshchilov and Hutter, 2017)

We perform grid search over the ranges of learning rates and batch sizes mentioned above for training each *Base* model with the cross-entropy loss function and finalize the best hyperparameters found for training the corresponding *Base + C*, *Base + V*, *Base + C + V* models. We run every experiment for three trials and report the mean and standard deviation over all trials of average F-1 scores.

4.4 Experiment Results, Analysis and Discussion

We report and evaluate performance of our models on the respective test sets in all experiments in terms of the Average F-1 score for every persuasion strategy individually as well as over all strategies, and Joint Accuracy. As defined by Chawla et al. (2021b), Joint Accuracy is the percentage of utterances for which the model predicts all the strategies correctly.

Ablations on Additional Context: We present in Table 2 and Figure 3 the average F-1 score results achieved for persuasion strategy prediction by every Base model with no previous context and the corresponding context model Base + C with different context lengths of 1, 3, 5, 7 and 9 previous utterances, each trained on the Ego4D Werewolf and YouTube subsets of our dataset. This ablation study shows how different context lengths may improve the performance of persuasion strategy prediction. We can see that for every one of the three baseline models, incorporating some previous context can boost the performance of persuasion strategy prediction but incorporating context of too many previous utterances can trip up the model possibly due to accumulation of too much irrelevant information. Instances of the latter are observed especially in utterances exhibiting persuasion strategies that can be confidently predicted based only on the

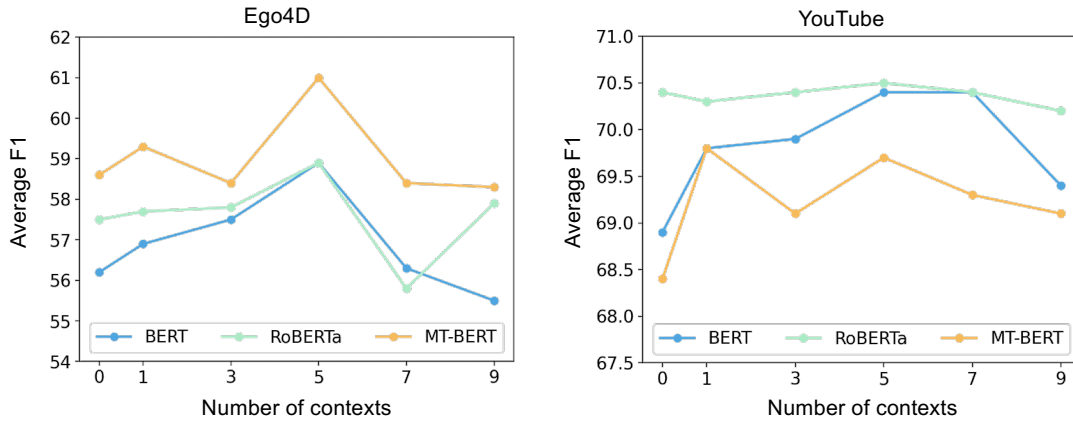


Figure 3: Ablation study of adopting different context lengths for persuasion strategy prediction. This figure also appears in a publication of this work (Lai et al., 2022).

current utterance. We consistently achieve the best improvement in performance with a context length of 5 for every baseline model, so for all other experiments incorporating context, whether with Base + C or Base + C + V, we use a context of 5 utterances as a default setting.

Ablations on Visual Features: We present in Table 2 the average F-1 score results achieved for persuasion strategy prediction by Base models with no video feature embeddings, Base + V models with visual features of corresponding video segment as well as Base + C + V models with visual features and context of previous 5 utterances. Incorporating video features in BERT + V notably showed an improvement of 0.8% over BERT for both Ego4D and YouTube domains but in the case of RoBERTa for the YouTube domain, the video model improves by only 0.2%. We hypothesize that this can be attributed to the significantly larger size of the YouTube dataset, which reduces the information gain from additional video features. We empirically observe that participants tend to engage in more frequent non-verbal communication by pointing to each other, raising hands, etc. during utterances exhibiting “Accusation”, “Interrogation”, and “Call for Action”. This could explain the larger improvement in model performance for these persuasion strategies by incorporating video features.

Off-the-shelf GPT-3 Inference: We also experiment with large language models off-the-shelf for direct inference for our strategy prediction task after seeing none or a very few examples from our dataset. Particularly, we experiment with GPT-3-175B (Brown et al., 2020) on our dataset under three settings: zero-shot, one-shot and five-shot. We obtain deterministic inferences by GPT-3 using the text-davinci-002 engine from OpenAI’s API¹ with temperature 0. The average F-1 score results of GPT-3 on the strategy prediction task are shown in Table 3. We can see that GPT-3 performs moderately well, and improves with more input examples, although outperformed by our fine-tuned baseline models.

Data Domain Generalization: Figure 4 shows average F-1 score results showcasing the generalization ability of our models across data source domains for persuasion strategy prediction based in the text modality. Particularly, we evaluate our Base and context-based Base + C models trained on the much larger YouTube subset but tested on Ego4D without and after fine-tuning with Ego4D. We observe that these models outperform models only trained on Ego4D due to the larger size of training data in the YouTube subset. This implies that there is not much difference between the Ego4D and YouTube domains in the text modality. After fine-tuning, the models show a small performance boost. These promising results give us hope that we can reliably implement transfer learning for persuasion modelling in social interactions, with the extensive

¹<https://beta.openai.com/>

	Method	Identity	Accusation	Interrogation	Call for Action	Defense	Evidence	Avg F1	Joint-A
Ego4D	BERT	82.6±1.1	48.8±4.8	82.8±0.2	39.4±9.6	29.3±5.5	54.2±2.5	56.2±2.5	65.1±1.6
	BERT + C	79.9±1.6	52.0±3.3	81.0±1.1	49.5±3.2	33.8±0.5	57.1±1.6	58.9±0.6	65.0±0.2
	BERT + V	81.5±3.5	52.1±1.9	83.3±1.6	42.4±3.8	28.4±5.1	52.8±1.0	56.7±1.2	64.5±1.2
	BERT + C + V	84.5±4.6	52.8±2.0	82.7±0.4	47.3±3.4	34.5±1.7	54.9±1.1	59.4±1.6	66.5±0.3
	RoBERTa	81.7±2.6	51.7±0.9	83.4±0.9	43.3±8.7	33.1±2.2	51.7±2.1	57.5±1.4	63.4±0.5
	RoBERTa + C	81.5±0.7	59.4±2.4	83.5±1.1	43.7±3.7	33.0±3.1	52.4±2.9	58.9±1.2	64.6±0.7
	RoBERTa + V	79.8±0.6	51.4±1.0	82.8±2.1	50.1±5.3	31.3±3.1	54.6±3.2	58.3±0.7	64.0±0.9
	RoBERTa + C + V	82.7±0.2	58.5±2.3	83.8±1.2	46.1±4.5	35.4±3.4	53.4±3.3	60.0±0.8	66.1±0.9
	MT-BERT	80.9±1.3	51.5±3.3	83.0±1.3	56.6±2.3	25.9±2.0	53.6±1.3	58.6±0.3	65.5±0.8
	MT-BERT + C	79.8±2.2	54.4±0.8	83.2±0.7	50.8±7.2	36.5±2.8	61.5±2.2	61.0±1.1	66.3±1.4
	MT-BERT + V	79.9±1.6	51.9±0.8	84.8±2.4	53.9±4.5	35.4±2.2	53.3±1.0	59.8±0.7	62.1±3.4
	MT-BERT + C + V	80.7±1.9	55.2±0.9	83.6±0.6	50.0±0.8	36.1±2.7	60.5±1.0	61.0±0.3	66.3±1.0
YouTube	BERT	80.2±1.6	64.7±1.1	89.6±0.4	77.2±2.5	43.5±1.0	58.3±0.7	68.9±0.0	64.6±0.8
	BERT + C	82.6±0.7	66.7±1.0	89.6±1.5	78.1±2.4	45.7±1.1	59.7±1.1	70.4±0.3	64.4±1.0
	BERT + V	82.4±0.5	65.4±1.4	89.7±0.1	78.0±0.8	45.3±2.8	58.4±1.3	69.9±0.4	66.2±0.5
	BERT + C + V	83.6±0.1	67.2±1.2	90.2±1.0	78.5±1.6	46.6±1.1	59.9±1.0	71.0±0.2	66.7±0.5
	RoBERTa	84.3±0.1	67.2±0.6	89.4±0.1	78.2±0.8	44.3±0.4	59.0±1.7	70.4±0.2	64.8±0.7
	RoBERTa + C	82.4±0.3	67.0±1.1	90.2±0.0	77.1±1.0	46.1±0.7	59.9±0.7	70.5±0.3	64.7±0.6
	RoBERTa + V	83.4±0.4	66.4±0.3	89.5±0.1	78.7±2.0	46.6±0.6	59.0±1.0	70.6±0.1	65.3±1.2
	RoBERTa + C + V	83.7±0.6	67.4±0.4	89.8±0.3	78.5±1.2	48.2±0.7	60.4±0.8	71.3±0.2	66.4±0.7
	MT-BERT	80.7±0.4	65.1±1.5	88.5±0.8	76.2±2.2	42.3±1.5	57.4±1.3	68.4±0.3	65.6±1.1
	MT-BERT + C	83.1±1.1	65.0±1.5	90.1±0.3	74.6±2.4	46.5±0.8	59.2±0.3	69.7±0.6	66.7±0.5
	MT-BERT + V	82.8±0.6	68.5±1.0	89.3±0.7	75.6±2.8	47.8±0.3	59.6±0.8	70.6±0.8	66.9±0.4
	MT-BERT + C + V	84.4±0.6	68.4±1.0	89.5±0.6	76.5±2.1	47.3±0.5	60.6±0.2	71.1±0.5	68.1±0.2

Table 2: Experimental Results on incorporating visual features and prior context for persuasion strategy prediction. This table also appears in a publication of this work (Lai et al., 2022).

Setting	Ego4D		YouTube	
	Avg F1	Joint-A	Avg F1	Joint-A
Majority	0	52.5	0	38.8
Zero-Shot	35.4	58.5	40.3	52.0
One-Shot	40.7	56.3	47.2	53.2
Five-Shot	47.0	59.7	49.6	53.7

Table 3: GPT-3 direct inference results on Ego4D and YouTube. This table also appears in a publication of this work (Lai et al., 2022).

and ever-increasing video data at our disposal online. We also observe that multi-task models (MT-BERT) may not generalize well.

Game Domain Generalization: Figure 5 shows average F-1 score results showcasing the generalization ability of our models across game source domains for persuasion strategy prediction based in text modality. Specifically, we report testing performance of our *Base* and *Base + C* models trained without any finetuning on games of One Night Ultimate Werewolf in Ego4D but tested on the small set of games of The Resistance: Avalon in the larger Ego4D subset. We observe that our models achieve relatively poor but overall passable performance on the new domain of Avalon games. The relative performance drop across the two domains can be explained by the large domain gap that arises because of the significantly different rules and winning criteria in both games. For example, One Night Ultimate Werewolf enforces only one voting round in every game whereas The Resistance: Avalon has multiple.

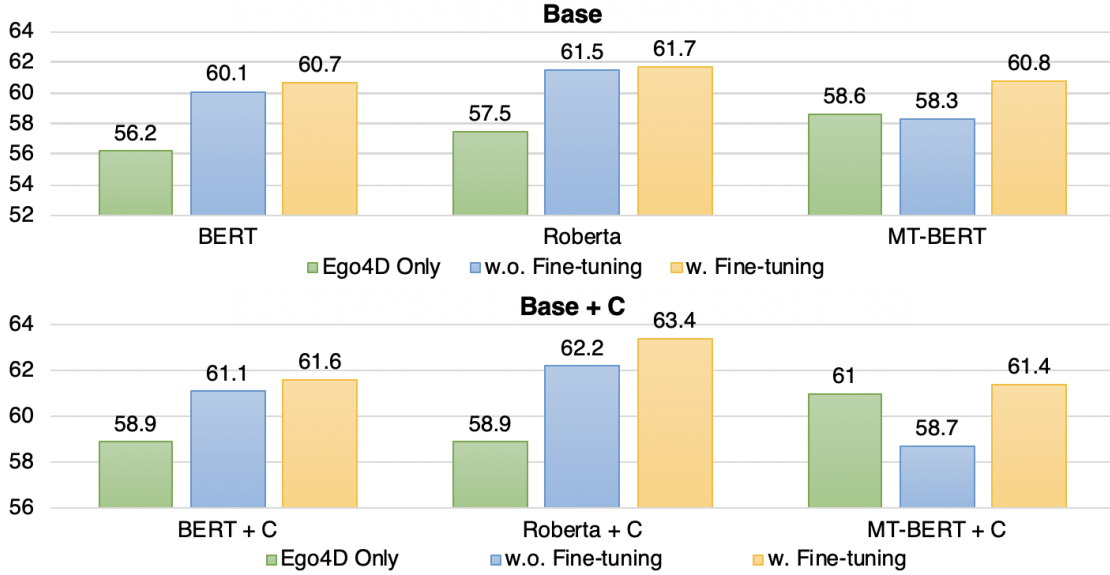


Figure 4: Data domain generalization experiments. We report testing performance (Average F-1) on Ego4D using models trained only on YouTube (w.o. Fine-tuning), and trained on YouTube and further fine-tuned with Ego4D (w. Fine-tuning). We also report performance (refer to Table 2) of models trained only on Ego4D (Ego4D Only) here for easy comparison.

5 Game Outcome Deduction

Finally, we show how persuasion strategies used in a game can be used to predict the game-level outcome. In this task, we try to predict the voting outcomes for each possible voter-candidate pair of players in the game. In other words, we try to predict whether the voter would vote for the particular candidate at the end of the One Night Ultimate Werewolf game, when players vote for one person to die in the game. To provide for a larger dataset size, we merge data points from the YouTube data and the Ego4D Werewolf data. We then partition the enlarged data set into train/val/test sets of 2741/427/827 data points, respectively.

5.1 Method

We formulate the input features for our voting outcome prediction task for every voter-candidate pair as follows: a 26-dimensional block feature vector formed by the concatenation of two 7-dimensional vectors representing the distribution of persuasion strategies (inclusive of "No Strategy") adopted by the voter and candidate in order, and a 12-dimensional one-hot vector representing the voter's starting role out of 12 possible roles designated in One Night Ultimate Werewolf. With such an input encoding, we run experiments for the binary classification task of whether or not a voter would vote for a candidate by training a simple logistic regression model with weighted binary classification loss.

5.2 Experiment Results, Analysis and Discussion

We report an average F-1 score of 32.7% and an AUC (area under ROC curve) of 54.7% achieved by our logistic regression model for the voting outcome deduction task. Comparatively, with random prediction, we obtain an average F-1 score of 28.6% and an AUC of 50.0%. Our observations hence speak to the importance of the use of persuasive language in determining outcomes of social deduction games. Using a model accepting only persuasion strategy embeddings in the input, we observe a negligible reduction in performance (by 0.5% in F-1 score and 0.1% in AUC). This strengthens the argument that compared to role embeddings, persuasion strategy embeddings are more conducive to inferring human belief states.

Visualizing the weights of our model as in Figure 6, we find that when a voter is predicted to

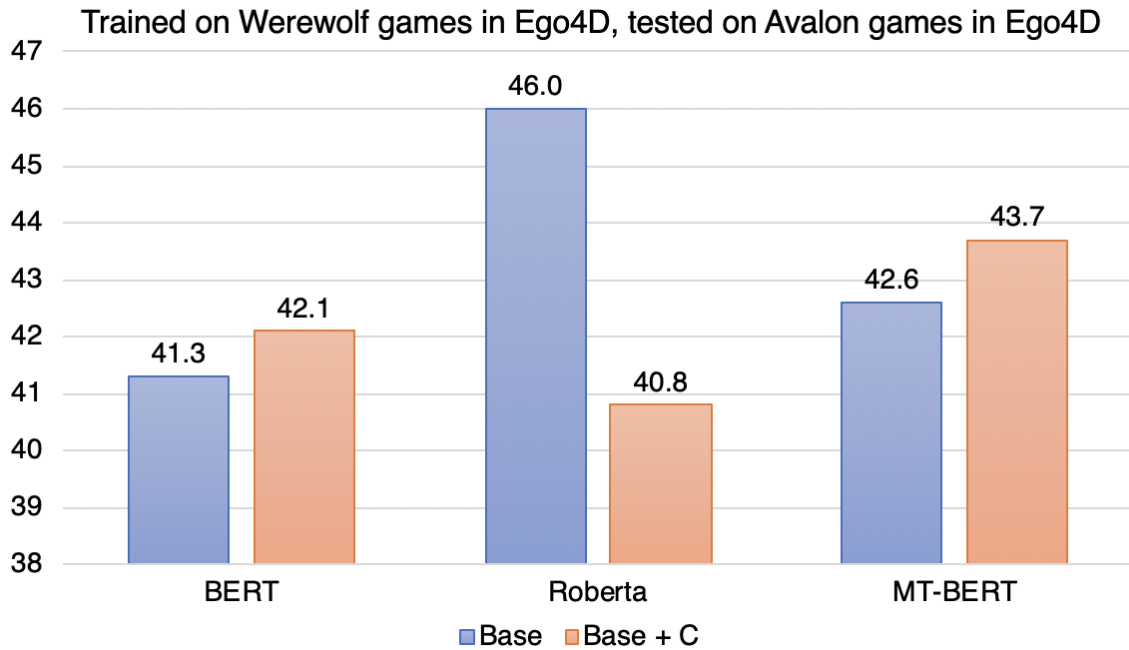


Figure 5: Game domain generalization experiments. We report testing performance (Average F-1) on Ego4D Avalon data using models trained only on Ego4D Werewolf data.

vote out a candidate, the candidate’s behavior has a greater weight in the prediction. This shows that the persuasive behavior of the candidate has more impact on the voting outcome than that of the voter, as expected. More interestingly, we observe that evidence is the strategy having the greatest weight in predicting that a voter will not vote for a particular candidate. This points to the importance of sharing of evidence and collaboration among players in implicating the werewolf.

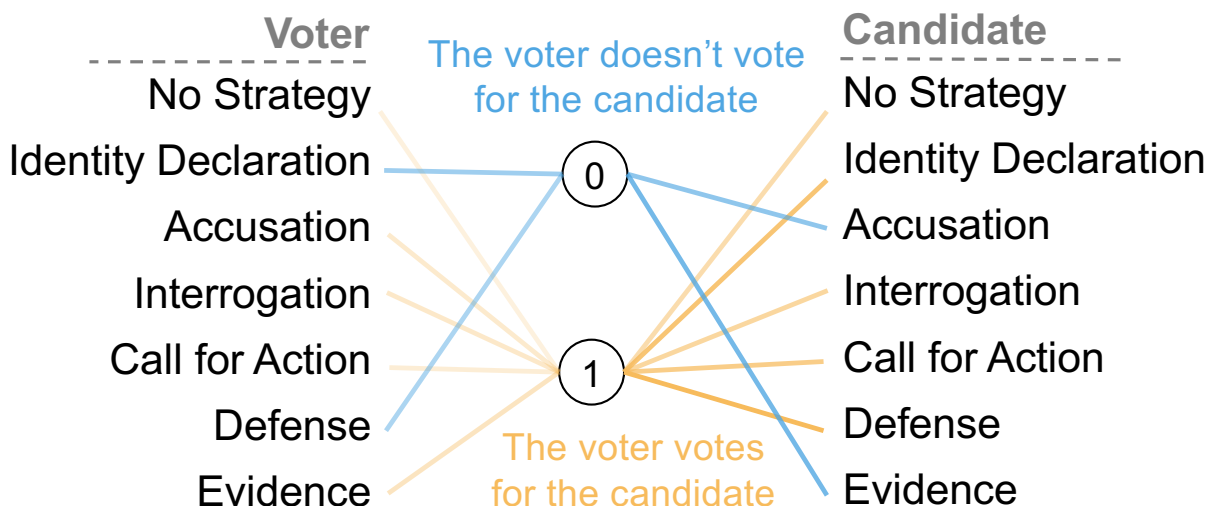


Figure 6: Weights visualization of candidate and voter persuasion strategy distributions in logistic regression. The transparency of connecting lines corresponds to the relative magnitude of the weights of the model and hence the relative degree of contribution to the prediction. This figure also appears in a publication of this work (Lai et al., 2022).

6 Conclusion

In this paper, we present the first benchmark dataset for persuasion modelling that is based in multiple modalities and that includes annotations of persuasion strategies at the utterance level. We build and train models that use both textual and visual representations to identify social persuasive behaviors. According to our research, visual signals can improve a model’s ability to predict persuasive tactics. We encourage more research on the potential use of the audio modality in persuasion modelling.

We only use commercially available, pre-trained video transformers to build the visual representation of the input. Other models, however, may be used in a more specialised manner to further enhance performance, similar to MT-BERT. The Ego4D collection also includes useful egocentric films and demographic data that we have not been able to incorporate in our models but that might provide richer, more predictive information for the task. We realize the need for a much larger dataset with many more completed games to achieve better and more practical results for the voting outcome deduction task, although this might prove to be more tedious and expensive. However, we can possibly overcome this issue with data augmentation techniques.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Kushal Chawla, Gale Lucas, Jonathan May, and Jonathan Gratch. 2020. [Exploring early prediction of buyer-seller negotiation outcomes](#).
- Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021a. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. *arXiv preprint arXiv:2103.15721*.
- Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021b. [CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.
- Jiaao Chen and Diyi Yang. 2021a. [Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests](#).
- Jiaao Chen and Diyi Yang. 2021b. [Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12648–12656.
- Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. [Detecting hate speech with gpt-3](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. [SemEval-2021 task 6: Detection of persuasion techniques in texts and images](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

- Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835.
- Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, and Dirk Hovy. 2021. [BERTective: Language models and contextual information for deception detection](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2699–2708, Online. Association for Computational Linguistics.
- Codruta Girlea, Roxana Girju, and Eyal Amir. 2016. [Psycholinguistic features for deceptive role detection in werewolf](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–422, San Diego, California. Association for Computational Linguistics.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.
- Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. [INSPIRED: Toward sociable recommendation dialog systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. 2022. [Werewolf among us: A multimodal dataset for modeling persuasion behaviors in social deduction games](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Antonia Lonigro, Roberto Baiocco, Emma Baumgartner, and Fiorenzo Laghi. 2017. Theory of mind, affective empathy, and persuasive strategies in school-aged children. *Infant and Child Development*, 26(6):e2022.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. [Revealing and predicting online persuasion strategy with elementary units](#). 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 6274–6279, Hong Kong, China. Association for Computational Linguistics.
- Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. [It takes two to lie: One to lie, and one to listen](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015. [Verbal and nonverbal clues for real-life deception detection](#). pages 2336–2346.
- Vivian P Ta, Ryan L Boyd, Sarah Seraj, Anne Keller, Caroline Griffith, Alexia Loggarakis, and Lael Medema. 2022. An inclusive, real-world investigation of persuasion in language and verbal behavior. *Journal of Computational Social Science*, 5(1):883–903.
- Xuwei 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.