The Language of Persuasion, Negotiation and Trust

José Lopes
Heriot-Watt University
Edinburgh, United Kingdom
jd.lopes@hw.ac.uk

Helen Hastie
Heriot-Watt University
Edinburgh, United Kingdom
h.hastie@hw.ac.uk

Abstract

There is a need to clearly understand the effect that interactive systems can have on users in the real world. This study explores whether aspects of social interaction (persuasion and negotiation) can be predicted purely from linguistic, politeness and collaborative features. Amongst other findings, we show that politeness cues (such as expressing gratitude) are important for successful negotiation dialogues and that collaborative features (such as repeated content between consecutively turns) are important for effective persuasion. We report here accuracy for automatic prediction methods based purely on interaction features using logistic regression, but also explore more opaque methods including neural models trained with dialogue embeddings. The two scenarios explored both involve economic decision-making, thus the subject has some stake in the outcome of the interaction, which is important for investigating trust.

1 Introduction

As interactive systems become more sophisticated, we can now look to various social aspects of interaction such as persuasion, negotiation and building of trusting relationships. However, there is a lack of understanding of how successful dialogues in this regard, manifest and what linguistic phenomena are observed. Designing and conducting studies to measure trust (subjectively or objectively) is particularly difficult because, in order to instill varying levels of trust in subjects, they have to be involved in the task and feel vulnerable to the outcome (Rousseau et al., 1998). One way to try to emulate this is to involve subjects in some kind of financial commitment to decisions made in the experimental set-up. The underlying assumption is that choices in such scenarios provide a reliable approximation of success in terms of persuasion, negotiation and consequently trust and trustworthiness (Camerer, 2011).

This paper reports an investigation into linguistic cues in two datasets that involve such economic decision-making: 1) where participants negotiate the price of items from real Craigslist advertisements (He et al., 2018); and 2) where one of the participants has to convince the other to donate part of their experimental reward to charity (Wang et al., 2019). The first of these datasets looks at a negotiation setting where one participant plays the role of a buyer and the other a seller and for the second dataset, one subject has the persuader role while the other is the persuadee. We posit that for both types of situations, in order for the interaction and transaction to succeed, there needs to be a trusting relationship between participants because these scenarios involve some emotional and financial investment. If we can establish trends and phenomena in language and dialogue that enable persuasion, effective negotiation and trust, then these can be used to inform dialogue management and natural language generation.

The importance of trust in human-robot interaction and conversational systems is a topic of much recent research (Kok and Soh, 2020). Levels of conflict of interest have been shown to be important for negotiation success (Cadilhac et al., 2013) and the role trust plays when coming to an agreement (Balliet and Van Lange, 2013). This form of cooperation depends on whether trust is conditional or unconditional. Conditional trust represents the minimum level of trust to facilitate social and economic exchanges toward a common goal (Jones and George, 1998). Rempel et al. (1985) state that trust evolves over time in interpersonal relationships, nurtured through interaction. However, trust can fall away rapidly, for example following an error (Nesset et al., 2021). In order to create interactive systems that are able to react and mitigate against

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

over-trust or undertrust/distrust (when perhaps the system is incorrect or misguided), we need to be able to monitor and infer the user's level of trust. Currently, measures of trust and trustworthiness are mostly collected from subjective questionnaires after the interaction (Schaefer, 2013; Jian et al., 2000; Ullman and Malle, 2018) or during the interaction (Khalid et al., 2019). Even if such methods have been validated, they can be considered intrusive as they break the flow of the interaction and are thus impractical for actually deployed systems. By observing linguistic phenomena, we hope to be able to develop an automatic method for predicting persuasion or whether a deal has been achieved without intrusive measures. In the future, we aim to extend the approach to predict trust in dialogues. This would allow for monitoring interaction, thus providing seamless mitigation through dialogue and language.

In this paper, we address the following research questions:

1. RQ1: Can we identify linguistic indicators present trustworthy interaction, in particular in successful persuasion/negotiation dialogues?

2. RQ2: Do role-specific linguistic indicators influence the outcome of the dialogue in a particular way?

3. RQ3: Can we use data-driven methods to predict the outcome of a persuasion/negotiation dialogue?

The contributions of the paper are thus two-fold: an in-depth analysis of linguistic indicators for successful dialogues, breaking this down by role, and providing discussion on how they may also influence trust in interaction. Secondly, we present data-driven methods, of varying transparency, for automatically predicting success in dialogue in terms of persuasion and whether a deal has been reached.

The paper starts by reviewing previous work on detecting trust and using linguistic indicators in predicting human behaviour (Section 2). The data used are described in detail in Section 3. The methodology is described in Section 4 and the results achieved in Section 5. We then discuss less transparent methods trained with dialogue embedding features and neural modelling in Section 6. In Section 7, we discuss the results and implications, and finally conclude the paper with Section 8.

2 Related Work

To achieve trustful interactions, systems have to become trustworthy. In order to do that, systems need to be equipped with resources to monitor the impact of their actions and how they affect the user's perception of the trustworthiness of the system. Therefore, there have been a number of studies where researchers have investigated specific cues that could be associated with trustworthiness. In Lucas et al. (2016), non-verbal cues have been studied in the context of negotiation dialogues between humans. Their goal was to predict both the perceived trustworthiness (i.e. partner perceptions of trustworthiness) and the reported perceived trustworthiness (i.e. if participants are honest). This study showed humans were actually poor predictors of trustworthiness, when compared with the proposed machine learning approach that used multimodal data. Still in the negotiation domain, Mell and Gratch (2017) found that negotiations were more likely to be successful when agents behaved aggressively. However, even if this strategy could lead to successful outcomes in the short-term, it is not necessarily advisable for human-robot interaction to display aggression in the long-term. Similar to our approach, Mell et al. (2019) used machine learning to predict the outcome of a negotiation using interaction features (e.g. number of turns), which were fed into both a linear model and a deep neural network. They do not, however, explore lexical features.

The above-mentioned approaches follow the intuition described in (DeSteno et al., 2012), that trust-related signals will likely emerge dynamically within the context of an interpersonal situation between individuals who are unfamiliar with each other. In addition, DeSteno et al. (2012) found that the accuracy of individuals in predicting trustworthy behaviour was higher when they had access to non-verbal cues. Examples of such cues were leaning forward or head nods. Lisetti et al. (2013), used a similar intuition to test whether different behaviours implemented in virtual agents were perceived more trustworthy. They found that the empathic version of the agent was generally preferred to its non-empathic counterpart on a number of dimensions related to trust, such as willingness to follow the agent's advice or politeness and willingness to continue the interview. Torre et al. (2018) also manipulated the agent's behaviour and measured the perceived trust. The virtual character

with a smiling face was perceived more trustworthy, knowledgeable and appealing. Kraus et al. (2020) modified the robot pro-activity and measured trust. The pro-active (contrasting with the reactive) version of the robot had a higher acceptance rate when it was possible to have natural dialogue, exemplifying the importance of dialogue and for acceptance and trust. Rapport building is also known to be a persuasion strategy that will likely increase trust. Therefore, Zhao et al. (2018) combined social dialogue with a model for task-oriented dialogue, including a first phase intended to build rapport. Examples of strategies used were self-disclosure, shared experience and praise.

2.1 Language and Trust

So far, we have focused on negotiation and persuasion as a means to maintain and manage trust, however, the above-mentioned studies mostly focus on non-verbal behaviour. Of specific interest here is whether we can observe linguistic indicators of these phenomena and use these to automatically predict varying levels of trust. Example studies looking into this area include Scissors et al. (2008), who investigated lexical mimicry (i.e. repetition of words or word phrases by both partners). They found that higher levels of mimicry were present in high-trusting pairs than low-trusting pairs. With regards to lexicon items, Rashkin et al. (2017) show that first-person and second person pronouns are used more in less reliable or deceptive news texts. On the other hand, Newman et al. (2003) found fewer self-references in people telling lies (so less trustworthy) about their personal opinions. These differences can perhaps be explained by the fact that the former is in relation to written facts, whilst the latter is about storytelling. With regards the use of superlatives and comparatives, Rashkin et al. (2017) found that trusted sources are more likely to use assertive words and less likely to use hedging words.

Continuing the theme of trustworthy news sources, Glenski et al. (2018) performed a study where they labelled bot and human users' reactions to (mis)information posted by various news sources and measured how bot and human users reacted to deceptive news sources compared to trusted news sources. However, the language aspect was not analysed. Volkova et al. (2017), on the other hand, found that incorporating linguistic and network features via a "late fusion" technique boosted performance of their suspicious tweet classifier. They found that verified news tweets contain significantly fewer bias markers, hedges and subjective terms.

Recent work has tried to use linguistic indicators to predict behaviours in interactive settings. Constructiveness has been one of the behaviours investigated in the context of an exploration game (Niculae and Danescu-Niculescu-Mizil, 2016) and disputes about Wikipedia articles (De Kock and Vlachos, 2021). In Zhang et al. (2018), politeness markers were used to predict if conversations were likely to fail at early stages. A conversation failure could be seen as a loss of conditional trust between interlocutors. In Niculae et al. (2015), sentiment, argumentation and discourse, politeness, subjectivity and talkativeness were used as linguistic cues to identify betrayal in a competitive game. As stated in Peskov et al. (2020), trust can be betrayed through deception, therefore some of these features might be relevant to our study. The most similar to our work is (Chawla et al., 2020), where BERT and linguistic features were used to predict the final price of successful negotiation in the Craigslist Bargain dataset. In our work, we use different lexical features and dialogue embeddings and have different tasks, namely the binary prediction of persuasion and whether a deal has been achieved. We believe this is an easier task for the model and thus would lead to further insights through the use of simpler more transparent modelling methods.

In this paper, we make use of some of the above-mentioned interaction cues, however, we apply them to negotiation and to the new domain of persuasion dialogues in scenarios of economic decision-making, where subjects in these types of scenarios have been shown to exhibit conditional trust.

3 Data

Two datasets were used in our analysis: the Craigslist Bargain dataset (He et al., 2018) and the Persuasion for Good dataset (Wang et al., 2019). In this section, we will provide a high-level description of these datasets. Further details can be found in the respective papers.

3.1 Craigslist Bargain

This dataset contains 6555 negotiation dialogues collected through crowd-sourcing. During data collection, crowd-workers were provided a real

Craigslist advertisement and were assigned roles of the buyer or the seller. They had to converse with another participant in order to negotiate an agreed price and thus close the transaction. Each participant was trying to push for a target price specified in the job (HIT) description. The datasets include information about these prices and the final closing price, and if participants eventually reach an agreement. This dataset has established partitioning for train/test/dev, which we have used in the research we present in this paper, in line with other work on the same dataset (He et al., 2018).

3.2 Persuasion for Good

The Persuasion for Good dataset is composed of 1017 dialogues between crowd-workers. Each participant had a specific role in the conversation. One crowd-worker, the persuader, had to convince the crowd-worker they were paired with, the persuadee, to donate a fraction of the amount they would receive for performing the task to a given charity (the same charity was used throughout the whole data collection). The persuader could also opt to donate part of their financial reward to the same charity at the end of the dialogue. The amount donated by each participant was recorded. The dataset also includes personality information gathered through pre-screening tests, in addition to demographics. A subset of dialogues was manually annotated for specific persuasion strategies and also for the intended donation verbalised by the persuadee during the dialogue (note that some persuadees actually did not donate the amount they verbally committed).

4 Method

In this section, we describe the method followed to perform two tasks: 1) predict the outcome of the dialogue and 2) identify the most relevant features in this prediction. Because we use a transparent method for prediction, we can do both of these tasks simultaneously. In both datasets, we have used the same features and extracted them from the conversations. We have drawn inspiration from the approach proposed in (De Kock and Vlachos, 2021) for feature-based models. We include the feature groups described below (a full reference of the features used can be found in Appendix A):

- Politeness strategies from (Zhang et al., 2018) for capturing tokens associated with greetings, apologies, directness and other politeness markers;
- Markers for collaboration from (Niculae and Danescu-Niculescu-Mizil, 2016) such as mutual pronoun usage or linguistic style accommodation (COLL).
- LIWC (Pennebaker, 2001) that provides counts of words associated with a given sentiment using pre-built lexicons.

All of the above-mentioned features were extracted at the turn-level, using Convokit (Chang et al., 2020). Similarly to De Kock and Vlachos (2021), at the end of the dialogue, for each feature we take the average (avg) and the gradient of a straight line fit of the feature value throughout the conversation (fit). The latter was done to assess how the feature value evolved throughout the dialogue. Then we have used the features, which will henceforth be called lexicon-based features, to train logistic regressions (LR). The LR method was chosen as it is reasonably transparent and allows the interpretation of the model by examining the weights of each feature.

5 Results

In this section, we present results for each of the datasets used. We used accuracy and F1-score as metrics, as all our tasks are binary classifications and the labels can be unbalanced (see the majority baselines in the results tables). We also report the McFadden R^2 score, the coefficient of determination, to provide a measure of how well the learned model fits the data. For the case of models trained with lexicon-based features, we report the 5 features with the highest absolute coefficients in the trained regressor.

5.1 Negotiation Dialogues

The task for this dataset is to automatically predict whether an agreement had been reached (binary deal/no-deal) and understand what features could help lead to this. We have used the dataset splits (5147 for train, 582 for validation, 826 for test) available in the data release. All sets of features used were able to improve over the majority baseline both in terms of accuracy and F1-score (it is a strong baseline given the dataset is highly unbalanced), as seen in Table 1. Regarding feature types, out of the 5 features in the best performing lexicon-feature based model, 4 were politeness features. In addition, dialogues with a trend of increasing turn length (fit_n_words) were more likely to lead

Features	Accuracy	F1-score	R^2	Top-5 features
Baseline Majority	0.769	0.869	-	-
COLL	0.810	0.886	-0.121	-avg_agree +fit_gap -fit_n_repeated_pos_bigram -avg_n_repeated_content -fit_n_repeated_stop
LIWC	0.815	0.886	0.016	-fit_n_words +avg_certain -avg_geo +avg_n_introduced_w_hedge +avg_n_introduced
Politeness	0.833	0.896	-1.123	-avg_has_negative -avg_apologising -avg_indicative -avg_direct_start -avg_indirect_greeting
COLL + LIWC + Politeness	**0.847**	**0.904**	**0.489**	-fit_n_words -avg_has_negative +avg_has_positive +avg_gratitude -avg_apologising
Buyer Features	0.832	0.896	**0.380**	-fit_pron_me +fit_pron_we +fit_1st_person +fit_indicative +avg_subjunctive
Seller Features	0.834	0.898	-0.222	+fit_n_introduced -avg_direct_start -fit_pron_you -fit_hedges +fit_indicative
Buyer+Seller Features	**0.857**	**0.910**	-0.519	-avg_seller_1st_person -avg_buyer_2nd_person_start +fit_seller_apologising +fit_buyer_please_start +fit_seller_n_adopted_w_hedge

Table 1: Accuracy, F1-score and McFadden's R^2 for predicting negotiation success in the Craigslist Bargain dataset. The speaker-independent features are in the top part of the table. Speaker-dependent features are in the bottom part of the table where the buyer and seller features include LIWC+Politeness separated out and calculated per role. The top-5 features are sorted according to the absolute coefficient value.

to a no-deal. This could indicate the use of longer, more elaborate utterances in an attempt to convince the other party. Dialogues where negative words were identified combined with a high number of apologetic words were also more likely to lead to no deal. On the other hand, dialogues where positive words were identified, combined with high rates of gratitude words (e.g. 'thank you') were more likely to result in a dialogue with a deal.

To further understand the impact of the behaviour of each participant in their various roles in the negotiation, speaker-dependent features were computed, specifically the LIWC and Politeness features for each speaker, be they a buyer or a seller. Since COLL features are meant to capture markers for collaboration, they are viewed as speaker-neutral. Thus in the lower part of Table 1, the results are split into the two buyer/seller roles. An interesting aspect when comparing results in the top half and bottom half of Table 1 is that the model trained with Buyer+Seller features from both speakers (i.e. LIWC+Politeness speaker-dependent features) has a better performance, both in terms of accuracy and F1-score, than the best model trained with speaker-independent features (COLL+LIWC+Politeness). Nevertheless, from the models trained with speaker-dependent features, only the buyer features achieved a R^2 above 0.2, the threshold to be considered a good fit between the trained model and the data. Therefore, when looking at the top speaker-dependent features for best performing model, some caution is warranted.

5.2 Persuasion Dialogues

For this dataset, we trained a LR to predict the persuasiveness, i.e., whether a donation was made by the persuadee. The Persuasion for Good dataset was not released with fixed splits, therefore we adopted a 5-fold cross-validation procedure following previous work with this dataset (Wang et al., 2019). In Table 2, we present the average accuracy and F1-score for all folds and their standard deviation. For each fold, we have saved the respective feature coefficients. Given that the metrics computed for the models have a small standard deviation, we assume that models in each fold are relatively similar and thus averaged the coefficient values for every feature across the 5 folds. The features presented in the tables are those with the highest absolute average coefficient values across folds. Similarly to the negotiation dataset, we also report the average R^2 across the different folds and respective standard deviation.

Using lexicon-based features, we observed a marginal improvement in terms of accuracy, when compared with the baseline majority, except when using only LIWC features. In the set of COLL features, the number of repeated content (avg_n_repeated_content) and stop words (avg_n_repeated_stop) in consecutive turns, and the number of agreement words (avg_agree) contributed to predicting persuasiveness. A high number of direct questions was one of the most valuable features to predict unpersuasive dialogues in the model trained with Politeness features (this feature

Features	Accuracy	F1-score	R^2	Top-5 features
Baseline Majority	0.536 (0.001)	**0.698 (0.001)**	-	-
COLL	**0.571 (0.029)**	0.653 (0.022)	-0.088 (0.063)	+avg_agree +avg_n_repeated_content +avg_n_repeated_stop -fit_disagree +fit_repeated_stop
LIWC	0.500 (0.044)	0.553 (0.033)	-0.107 (0.125)	-avg_geo +coordination_score +avg_n_adopted +avg_n_introduced +avg_n_introduced_w_hedge
Politeness	0.568 (0.031)	0.626 (0.049)	-0.088 (0.160)	+avg_has_positive -avg_direct_question -avg_has_negative +avg_gratitude +avg_2nd_person_start
COLL + LIWC + Politeness	0.556 (0.039)	0.591 (0.038)	**0.025 (0.058)**	-avg_geo -avg_has_negative +avg_has_positive +avg_agree -avg_direct_question

Table 2: Mean Accuracy, F1-score and McFadden's R^2 for predicting persuasion in the Persuasion for Good Dataset in the 5-folds. The figure between brackets represent the standard deviation across the different folds. The top-5 features are sorted according to the mean of absolute coefficient values.

was automatically detected by the occurrence of the initial wordings of "what, why, who or how").

6 Opaque Models for Prediction of Persuasion and Negotiation

As well as the traditional lexicon-based features described above, we have also used embedding-based features, specifically: RoBERTa-SE sentence embeddings (Reimers and Gurevych, 2019) trained for the STS task[1]; and a dialogue vector representation extracted from a ConvERT model (Henderson et al., 2019). For sentence-based models (RoBERTa-SE), for each turn an embedding was generated. The dialogue representation is then the average of the sentence embeddings for all dialogue turns. In the ConvERT model, given the context and the current utterance, the model would provide a dialogue embedding. We compare these two models in order to assess the impact of using a model that attempts to keep the sequential structure of the data (ConvERT) versus a model trained with a larger amount of data (RoBERTa-SE).

These embeddings were given as inputs either to a LR or a neural model composed by a linear layer and a softmax layer, which provides the probability distribution of the different classes (Linear-NN). The reasoning behind this was to see if the neural model was better at predicting whether the dialogue resulted in successful negotiation, even though this method is less transparent than LR.

Results from embedding-based dialogue representations predicting negotiation success in the Craigslist Bargain dataset are shown in Table 3. The fact that ConvERT keeps the sequential structure of the data seems to provide an advantage over RoBERTa-SE in terms of F1 and accuracy. It is interesting to observe a drop in performance from the LR-models to the NN-models. In any case, models based on pre-trained dialogue representations seem to improve the performance over models trained with lexicon-based features using LR, as well as observing a higher R^2 (as reported in Table 1).

For persuasiveness prediction, the neural models trained with ConvERT (see Table 4) outperform those using LR with embedding features and also LR with linguistic features (see Table 2). However, again, the disadvantage of this approach is that these models are less transparent.

7 Discussion

As we look at the features, we find some interesting results. Tables 5 and 6 show an example dialogue and corresponding features, from a Craigslist Bargain and a Persuasion for Good dialogue respectively. One of the features emerging as potentially contributing to no deal was an increasing number of words per utterance (fit_n_words) as the dialogue progresses (see Table 1). In the example of a dialogue where a deal was reached, shown in Table 5, there is a tendency for short utterances as the dialogue unfolds. One of the factors associated with increasing trustworthiness is transparency (Nesset et al., 2021). However, a direct consequence of increasing transparency in dialogue is an increase in the number of words per sentence. This seems an interesting avenue for future research, to instill the appropriate amount of trust while keeping the utterance short, along with the appropriate level of transparency.

Another interesting outcome is that the average number of apologetic words were higher in no deal dialogue compared to dialogues where a deal was reached. This may be due to the fact that people

[1] https://github.com/UKPLab/sentence-transformers

Features	Model	Accuracy	F1-score	R^2
RoBERTa-SE	LR	0.854	0.906	**0.560**
ConvERT	LR	**0.895**	**0.932**	0.533
RoBERTa-SE	Linear-NN	0.843	0.904	-
ConvERT	Linear-NN	0.859	0.913	-

Table 3: Accuracy, F1-score and McFadden's R^2 (in the LR models) for prediction negotiation success in the Craigslist Bargain dataset using dialogue embeddings.

Features	Model	Accuracy	F1-score	R^2
RoBERTa-SE	LR	0.611 (0.038)	0.638 (0.052)	0.050 (0.331)
ConvERT	LR	0.602 (0.022)	0.665 (0.027)	0.120 (0.003)
RoBERTa-SE	Linear-NN	0.607 (0.010)	**0.724 (0.013)**	-
ConvERT	Linear-NN	**0.622 (0.018)**	0.715 (0.004)	-

Table 4: Average accuracy, F1-score and McFadden's R^2 (for the LR models) in the 5 folds for predicting persuasion in the Persuasion for Good dataset using dialogue embeddings. Number between brackets is the standard deviation in the 5 folds.

Buyer: I am interested in purchasing this item, but since it is used I can only afford to pay about 25 **Seller**: I mean, we can work out a deal, but that is way too low. how about 60? **Buyer**: Shoot, I only have about 40 in my account right now.	avg_agree = 0.0 fit_gap = 0.016 fit_n_repeated_pos_bigram = −0.333 avg_n_repeated_content = 0.0 fit_n_repeated_stop = −0.333 **fit_n_words** = −0.214 avg_n_certain = 0.0 avg_n_geo = 0.0 avg_n_introduced_w_hedge = 0.0 avg_n_introduced = 0.0 **avg_has_negative** = 0.0 **avg_apologising** = 0.0 avg_indicative = 0.0 avg_direct_start = 0.0 avg_indirect_greeting = 0.0 **avg_has_positive** = 1.0 **avg_gratitude** = 0.0

Table 5: Example of a dialogue where a deal was reached from the Craigslist Bargain dataset with corresponding feature values. Underlined words have direct impact in the feature values reported. Top-5 features of COLL+LIWC+Politeness model in bold from Table 1.

apologise for the negotiation not being successful or being unable to adjust the price to the other person's requested price.

Collaborative features seem to be more important for success in persuasion than negotiation dialogues (see Tables 1 and 2) when compared to other lexicon-based features. This could be explained by the fact that the persuasion needs a high amount of collaboration, where both participants could benefit from a positive outcome, whereas in Craigslist Bargain the task of negotiation is competitive and both users have to compromise to achieve a trade-off.

The number of geographical-related words, given by the avg_geo feature seems to be influential, which is perhaps non-obvious. Since none of the datasets are likely to have a large number of geographical references (even if there is a section of Craigslist Bargain about housing). It could be that the geographical lexicon has several polysemic words (e.g. the word physical in part of this lexicon) and therefore introduces some extra noise in the data. This could be an example of lexicon that could be filtered out in future work.

In both corpora, the number of positive words and agreement expressions contributed to predicting the success of the dialogue. Interestingly, some of the most predictive features are not explicitly related to the success of task (such as the use of agreement words), but rather relate more to the language common to the various roles, e.g., whether the seller uses the first person or whether the buyer starts a sentence with a second person pronoun. On the other hand, gratitude has been one of the features which contributed to a correct prediction of a deal. Social norms might explain this, since it is considered polite to thank the other person at the end of the negotiation. Attempting to predict the outcome of the dialogue in earlier stages could help us to confirm this hypothesis.

The performance of the models is generally higher in the Craigslist than the Persuasion for Good dataset, which could likely be due to the larger size of the Craigslist dataset. Nevertheless, the set of features used by the best performing models with lexicon-based features (COLL+LIWC+Politeness and Buyer+Seller Features) was rather unexpected. One perhaps would have thought that features such as agreements/disagreements would be natural candidates to be relevant features for the task at hand, but these were not part of the top-5 most relevant feature when combining all the lexicon-based features. Also the best performance for lexicon-based feature

Persuader: Hello **Persuadee**: Hi there how are you? **Persuader**: Great and you **Persuadee**: I am doing fine, thanks for asking. **Persuader**: Can I ask you a question? **Persuadee**: You sure can **Persuader**: Do you have any kids? **Persuadee**: I don't have any children but there are many children in my family. **Persuader**: Im sure you love them as your own, I have a big family do you? **Persuadee**: I do have a big family, I enjoy it at times, other times I don't **Persuader**: I truly understand. But I know I would do anything for them especially if it was for their wellbeing, do you feel the same way **Persuadee**: Yes I do. **Persuader**: It's great to know that there are others who feel the same way. Have you heard of Save the Children? **Persuadee**: No I haven't. Could you QUICKLY tell me about it **Persuader**: Save the Children is an international non-governmental organization that promotes children's rights, provides relief and helps support underlinechildren in developing countries. **Persuadee**: Okay that sounds nice and an important service **Persuader**: And the money raised helps feed and clothe them. Its a lot of underlinechildren that are starving and need our help. Would like to help? **Persuadee**: I would like to help in the future when I am more financially stable. **Persuader**: I understand but even the smallest amount would be a BIG help. **Persuadee**: I am sure but I just am not able at this time	**avg_agree** $= 0.053$ avg_n_repeated_content $= 0.053$ avg_n_repeated_stop $= 2.053$ fit_disagree $= NaN$ fit_n_repeated_stop $= 1.285$ **avg_geo** $= 0.0$ coordination_score $= NaN$ avg_n_adopted $= 0.150$ avg_n_introduced $= 0.150$ avg_n_introduced_w_hedge $= 0.0$ **avg_has_positive** $= 2.2$ **avg_direct_question** $= 0.0$ **avg_has_negative** $= 0.0$ avg_gratitude $= 0.050$ avg_2nd_person_start $= 0.05$

Table 6: Example of unsuccessful dialogue from the Persuasion for Good dataset with corresponding feature values. Underlined words have direct impact in the feature values reported. Top-5 features of COLL+LIWC+Politeness model in bold from Table 2.

models in the Craigslist was achieved by separating the seller from the buyer features. This is an interesting outcome and reinforces the different roles of each speaker in the dialogue.

Finally, initial results suggest that dialogue embeddings are powerful representations that can be used to predict the outcome of the dialogue. In fact, for LR trained with dialogue embeddings, the R^2 was above 0.2 for negotiations, unlike most of the cases using lexicon-based features, which shows a better fit to the data. However, interpretable features and models can provide more explainable and transparent cues.

8 Conclusion and Future Work

We have investigated linguistic indicators that reflect two tasks, namely a successful negotiation and persuasion of a donation. These two interaction outcomes can be seen as examples of conditional trust (Jones and George, 1998), since they involve social and/or economic exchanges. In the case of negotiation, the task is competitive, whereas persuasion dialogues can be considered more of a collaboration. Various lexicon-based features were identified as being indicators of success through our method of training regressors. However, a role-based analysis showed differences in the relevant features. Therefore, considering the role will be important when designing trustworthy conversational agents. Future work will look into individual differences more deeply and explore variations of personality and propensity to trust of individual users.

Methods based on dialogue embeddings achieved the best performance in both problems, however these methods are opaque. Future work would involve combining recent work on transparent NLP methods for explaining embedding models (Hoover et al., 2020) and explainable AI (Ribeiro et al., 2016), so as to provide further insight into linguistic and dialogue features for these opaque but high performing features and models.

In the introduction, we mentioned that in both datasets used in this research we were using proxies for trust assuming that financial transaction between subject would only occur when a certain level of trust is achieved. This is a limitation of our work, which are trying to address at the moment by collecting trustworthiness ratings at turn level. This will allow us to confirm whether our assumption is correct and develop a fine-grained strategy to increase trustworthiness in conversational agents.

Finally, a discussion on the ethical implications is needed of using interactive systems for these types of interactions, where trust is conditional on the perceived behaviour of system.

Acknowledgements

This work was funded and supported by the EPSRC ORCA Hub (EP/R026173/1) and UKRI Node on Trust (EP/V026682/1).

References

Daniel Balliet and Paul AM Van Lange. 2013. Trust, conflict, and cooperation: a meta-analysis. *Psychological bulletin*, 139(5):1090.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Colin F Camerer. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Kushal Chawla, Gale M. Lucas, Jonathan Gratch, and Jonathan May. 2020. BERT in negotiations: Early prediction of buyer-seller negotiation outcomes. *CoRR*, abs/2004.02363.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

David DeSteno, Cynthia Breazeal, Robert H. Frank, David Pizarro, Jolie Baumann, Leah Dickens, and Jin Joo Lee. 2012. Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, 23(12):1549–1556. PMID: 23129062.

Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. How humans versus bots react to deceptive and trusted news sources: A case study of active users. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '18, page 654–661. IEEE Press.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Matthew Henderson, Inigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A visual analysis tool to explore learned representations in Transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71.

Gareth R. Jones and Jennifer M. George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, 23(3):531–546.

Halimahtun Khalid, Wei Shiung Liew, Bin Sheng Voong, and Martin Helander. 2019. Creativity in measuring trust in human-robot interaction using interactive dialogs. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, pages 1175–1190, Cham. Springer International Publishing.

Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports*, pages 1–13.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 107–116, New York, NY, USA. Association for Computing Machinery.

Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Trans. Manage. Inf. Syst.*, 4(4).

Gale Lucas, Giota Stratou, Shari Lieblich, and Jonathan Gratch. 2016. Trust me: Multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 5–12, New York, NY, USA. Association for Computing Machinery.

Johnathan Mell, Markus Beissinger, and Jonathan Gratch. 2019. An expert-model & machine learning hybrid approach to predicting human-agent negotiation outcomes. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 212–214, New York, NY, USA. Association for Computing Machinery.

Johnathan Mell and Jonathan Gratch. 2017. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, pages 401–409.

Birthe Nesset, David A. Robb, José Lopes, and Helen Hastie. 2021. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 313–317, New York, NY, USA. Association for Computing Machinery.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675. PMID: 15272998.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California. Association for Computational Linguistics.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1650–1659, Beijing, China. Association for Computational Linguistics.

James W Pennebaker. 2001. Linguistic inquiry and word count: LIWC 2001.

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.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404.

Kristin Schaefer. 2013. *The perception and measurement of human-robot trust*. Ph.D. thesis.

Lauren E. Scissors, Alastair J. Gill, and Darren Gergle. 2008. Linguistic mimicry and trust in text-based cmc. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, page 277–280, New York, NY, USA. Association for Computing Machinery.

Ilaria Torre, Emma Carrigan, Killian McCabe, Rachel McDonnell, and Naomi Harte. 2018. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 423–427, New York, NY, USA. Association for Computing Machinery.

Daniel Ullman and Bertram F. Malle. 2018. What does it mean to trust a robot? steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 263–264, New York, NY, USA. Association for Computing Machinery.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. 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.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Ran Zhao, Oscar J. Romero, and Alex Rudnicky. 2018. Sogo: A social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, page 239–246, New York, NY, USA. Association for Computing Machinery.

A Linguistic indicators reference

Tables 7, 8 and 9 has a complete reference of all lexical features used.

Feature	Definition
n_repeated_pos_bigram	number of repeated POS bigrams in consecutive turns
n_repeated_content	number of repeated content words in consecutive turns
agree	whether there is an agreement expression
disagree	whether there is a disagreement expression
n_repeated_stop	number of repeated stop words in consecutive turns.
coordination score	coordination score between the two speakers

Table 7: Collaborative (COLL) indicators reference.

Feature	Definition
n_adopted_w_hedge	number of w ords re-used from hedges lexicon
n_words	number of words per utterance
n_introduced	total number of words re-used
n_adopted_w_certain	number of words re-used from certain lexicon
n_introduced_w_hedge	number of newly introduced words from the hedges lexicon
pron_we	number of usages of words from the we lexicon
geo	number of usages of words from the geographic terms lexicon
hedge	number of words from the hedges lexicon
n_introduced_w_certain	number of newly introduced words from the certain lexicon
pron_you	number of words from the you lexicon
meta	number of words from the meta lexicon
pron_me	number of words from the me lexicon
n_adopted	number of re-used words
pron_3rd	number of words from the

Table 8: LIWC indicators reference.

Feature	Definition
please_start	if utterance starts with please
factuality	if utterance has factuality expressions (e.g. actually)
apologising	if utterance contains apologetic words
2nd_person	if utterance contains second person words
please	if utterance contains please
direct_question	if utterance starts with what, why, who or how
gratitude	if utterance contains gratitude words
has_positive	if utterance as positive words
1st_person_start	if utterance starts with a first person pronoun
1st_person	if utterance has first person pronouns
1st_person_pl.	if utterance contains first person plural pronouns
subjunctive	if utterance includes 'could' or 'would' before 'you'
indicative	if utterance includes 'can' or 'will' before 'you'
direct_start	if utterance has a direct start
indirect_(greeting)	if utterance starts with 'hi', 'hello' and 'hey'
has hedge	if utterance has hedges
indirect_(btw)	if utterance contains expression 'by the way'
has_negative	if utterance has negative words
deference	if utterance has deference words
2nd_person_start	if uttreance starts with a second person pronouns

Table 9: Politeness Linguistic indicators reference.