Smile and Nod: Few-Shot Detection of Social Influence-Based Disingenuity in Discussions

Emily Hu
Symbolic Systems Program
xehu@stanford.edu

Michael Cooper
Dept. of Computer Science
coopermj@stanford.edu

Makena Low
Dept. of Computer Science
makenal@stanford.edu

Abstract

Deception detection is a difficult task, both for humans and machine learning models. In this paper, we contribute the novel NLU task of detecting social influence-based *disingenuity* in group discussions. We present a novel dataset for this task, on which we test three state-of-the-art machine learning models with two optimization schemes. We additionally benchmark our performance against humans. The task we propose is challenging; although several of our models outperform statistical and human baselines, overall performance leaves room to be desired. This task and dataset, therefore, remain an open challenge for NLU research.

1 Introduction

Social influence is a well-studied phenomenon in social science; over the years, numerous studies have demonstrated that a group's majority strongly determines outcomes (Sunstein, 2008; Hackman and Katz, 2010; Hastie et al., 2013), even dominating individuals' private doubts (Asch, 1961; Son et al., 2019). The phenomenon is powerful, but often implicit and subtle — one's peers in a group may seem outwardly agreeable, but may leave important countervailing perspectives unsaid.

The ability to detect when a conversation is not as unanimous as it appears is significant, with broader implications for deliberative democracy and group decision-making. When we rely on groups to decide, we expect that deliberation enables the decision-making body to thoroughly consider diverse perspectives. Social influence-based disingenuity breaks that assumption.

1.1 Task Definition

Our central research question asks, *is it possible to capture disingenuity using NLU techniques*? To answer this question, we define the novel task of disingenuity detection as:

> a (set of) utterance(s) that is: (1) disingenuous to the speaker, but is perceived to be not disingenuous by the listener; and (2) uttered in a social context in which the majority of listeners agree with the utterance.

The first part of this definition draws from definitions of deception (Salvetti et al., 2016; Siegler, 1966; Peskov and Cheng, 2020). The second part establishes the aspect of social influence.

Importantly, individual utterances may not explicitly express the opinion, but could imply it collectively. For example, if a person is merely nodding along in feigned agreement, they may say "sure," and "uh-huh" repeatedly, but may not explicitly make a deceptive statement. Thus — as we formalize in Hypothesis 1 — disingenuity may need to be inferred through the context of a conversation, rather than through sentence-level classification.

In this paper, we contribute a novel task in NLU that seeks to detect social influence-based disingenuity in group discussions. We also contribute a labeled dataset (the JUror Disingenuity in Group Environments, or JUDGE dataset) with which to test model performance on this task. We test three state-of-the-art machine learning models (Context-Free Classifier, Contextual Classifier, and Augmented Contextual Classifier, all defined below) with two optimization schemes (transfer learning and model-agnostic meta learning) on this dataset. Several approaches outperform statistical and human baselines in this task, suggesting promising research avenues for detecting subtle social indications of disagreement that even humans struggle to identify.

1.2 Hypotheses

Broadly, we hypothesize that the answer to our research question is *yes*: NLU will be successful in capturing disingenuity. We divide this hypothesis into three parts, each of which makes specific predictions about our NLU experiments. The hypotheses are summarized in Figure 1.

Hypothesis 1 (H1): A Context-Free Classifier trained on sentence-level inputs will underperform a Contextual Classifier trained with participant-level inputs.

Hypothesis 2 (H2): Within models trained on contextual (participant-level) inputs, those augmented with human-selected features will outperform those without augmentation. These human-selected features enable models to incorporate human insight about the "tells" of disingenuity — such as using more positive sentiment, or speaking in brief sentences (Niculae et al., 2015).

Hypothesis 3 (H3): The JUDGE dataset is very small in size (fewer than 200 disingenuous statements); therefore, we hypothesize that transfer-learning techniques will perform poorly. We hypothesize that model-agnostic meta learning (Finn et al., 2017), which has been shown to significantly improve performance on few-shot tasks in NLU (Dou et al., 2019), will outperform Transfer Learning in detecting disingenuity on this dataset.

2 Related Work

Given that disingenuity detection is (to our knowledge) a novel task, in this section we will address such past research on deception detection, which is a close parallel to our main task. We will also discuss a modeling technique rarely explored within this domain: few-shot learning via meta-learning.

2.1 Deception Detection

Many deception papers reveal that humans are not skilled at detecting lies, while machines are able to determine common "tells" of deception.

Peskov and Cheng (2020), for instance, is a very recent paper introducing a large labeled dataset on deception via player data from the game Diplomacy. The authors then investigated detection of intended and perceived lies using logistic regression and neural models. Notably, both the models and humans generally performed poorly.

Another paper by Chen et al. (2020) explores the physical and linguistic cues that people use to detect lies using a game called LieCatcher (text and audio clips only). Consistent with previous literature, they find that, on average, participants deception dection is close to chance. Via logistic regression, the authors find that speech hesitations and errors were both strong indicators of deception for humans and truly an accurate predictor.

Chen et al. (2020), Porter and ten Brinke (2010), and Pérez-Rosas et al. (2015) also find behavioral traits important. This underscores the challenge of our task: we flatten a multi-modal problem (e.g., facial expression, tone) into a unimodal one. In this way, we make an already-tough problem harder.

The mere diversity of "tells" across each paper underscores the complexity. Moreover, humans and models have both historically struggled with performing well on text-based deception detection (Peskov and Cheng, 2020; Chen et al., 2020; Pérez-Rosas et al., 2015).

We do, however, draw inspiration from the multi-modal literature to lighten our (admittedly quite heavy) load: for instance, we use social features such as talkativeness and sentiment as possible indicators for deception.

2.2 Few-Shot Learning via Meta-Learning

Finn et al. (2017) present model-agnostic meta-learning (MAML), a meta-learning approach designed to enable the fast adaptation of models to novel unseen tasks. The task of MAML is presented as the problem of adapting a model to new tasks drawn from a task distribution. Building on the work of Finn et. al, Nooralahzadeh et al. (2020) apply meta-learning to the problem of information sharing across different languages. They propose X-MAML, a three-step algorithm, consisting of first pre-training a model on a high resource language (English), then meta-training according to the MAML algorithm on low-resource languages, and finally performing zero- or few-shot learning on the low-resource target languages. X-MAML outperforms external and internal baselines pre-trained on SQuAD (Rajpurkar et al., 2016). Lastly, a recent investigation by Dou et al. (2019) explores whether transfer learning — the current trend in training deep-learning based natural understanding models — is optimal in light of recent advances in meta-learning for few-shot domain adaptation. They investigate how well meta-learned representations transfer to new tasks in the limited-data regime: to do so, they compare the performance

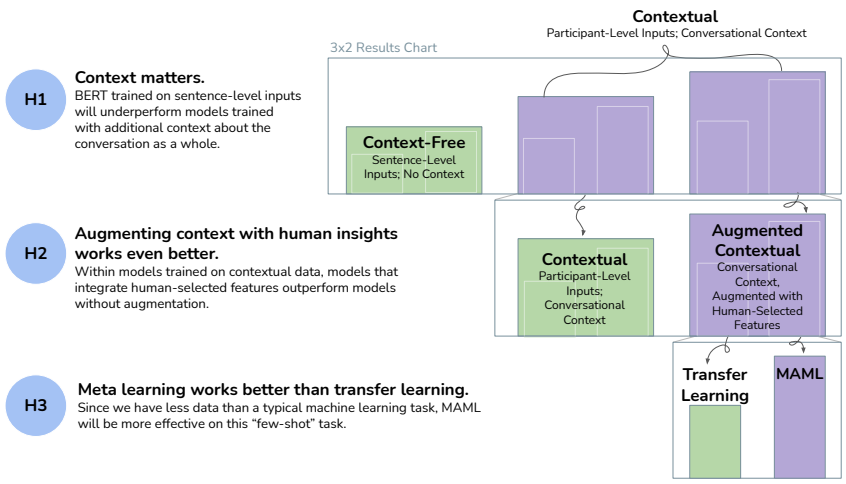


Figure 1: Hypotheses 1, 2, and 3, juxtaposed against an illustration of our results chart. We demonstrate how each hypothesis examines a different comparison by examining increasingly narrower slices of our chart.

of Reptile (a variant of the meta-learning update step) (Nichol et al., 2018) against two baselines, BERT (Devlin et al., 2018) and MT-DNN (Liu et al., 2019), on the SciTail dataset (Khot et al., 2018). They trained on a randomly subsampled percentage of 0.1%, 1%, 10%, and 100% of the training data, and found that, unsurprisingly, the performance of each model improves as more training data is used; however, Reptile outperforms MT-DNN and BERT in each instance (and, in the cases where 0.1% or 1% of the training data is used, Reptile and MT-DNN significantly outperform BERT).

We draw upon the meta-learning papers the the following way: we use Finn et al., to establish a base understanding of the MAML algorithm, Nooralazadeh et. al to approach resource-limited tasks in the NLU space, and Dou et al. to frame the relative strengths and weaknesses of meta-learning approaches to data-limited tasks in NLU.

3 Data

We used two datasets for the training of our models: the JUDGE Dataset, which we introduce, and the Deception in Diplomacy Datset, sourced from (Peskov and Cheng, 2020). In both datasets, we calibrate the labels such that the *positive class* indicates disingenuity or deception, since this is our phenomenon of interest.

3.1 JUDGE Dataset

This dataset originates from Hu et al. (2021), a study investigating the consistency of online jury deliberation. The data involve a contextually rich setting (jury deliberation), in which participants engaged in deliberations about topics sourced from Reddit. Participants were randomly assigned pseudonyms on the platform in the form of *adjective-animal*, such as *happyCow* or *excitedLion*. Before and after each deliberation, jurors submitted a private survey about their opinion of the case; during the round, they submitted an in-round public vote. Thus, we labeled participants as "disingenuous" by comparing in-round voting patterns to out-of-round private expressions of belief.

Of 356 team deliberations, we identified 31 jurors whose post-deliberation survey results did not match their in-round votes. After removing participants who were clearly entering spam, 29 jurors collectively made 181 statements. The remaining deliberations totaled 2,348 jurors, who collectively made 14,798 statements; we generated one dataset for the utterance level (JUDGE-utt) and one for the discussant level (JUDGE-dis). Table 1 summarizes the metrics:

	Statements	**Participants**
Disingenuous	181	29
Not Disingenuous	14798	2348

Table 1: Data metrics for the jury deliberation dataset. Note that the dataset does not contain statement-level information about whether a given statement is true or false. Therefore, we use a coarse heuristic for JUDGE-utt, in which we assume that everything uttered by a disingenuous juror is a 'Disingenuous,' although we recognize that, in reality, those jurors have likely made some true statements.

We split the data into training and testing (omitting validation due to the low number of positive

labels). At the statement level, 61 disingenuous statements were included in the training set, and 120 in the test set; at the participant level, 10 disingenuous jurors were included in the training set, and 19 in the test set. The negative class was divided evenly between train and test sets.

3.2 Deception in Diplomacy Dataset

In addition to the JUDGE Dataset, we leveraged the Deception in Diplomacy Dataset from Peskov and Cheng (2020) for pre-training and meta training, since we reasoned that the similar task would transfer to disingenuity detection. This dataset contains 17,289 messages exchanged between players of the negotiation and strategy-based game, Diplomacy. Each message is annotated by the sender, with a score of intended truthfulness, and by the receiver, with a score of perceived truthfulness. This dataset is split into training, validation, and test sets, respectively; metrics for each split are provided in Appendix A.4.

Finally, while we requested access to the Boulder Lies and Truth dataset from Salvetti et al. (2016) (which we planned use the same way as the Deception in Diplomacy Dataset), we were unable to obtain access via the Stanford channels.

4 Models

In this section, we describe models used to execute our task, beginning with baselines, proceeding to parameterized models, and closing with metrics.

4.1 Baselines

4.1.1 Random Baselines

We define the following three random baselines, each of which is designed to test a different assumption over the JUDGE dataset.

The first random baseline is a Naive Random Baseline, in which the baseline model returns 'Disingenuous' or 'Not Disingenuous' with 50% probability. This baseline tests the performance of random guessing without prior knowledge over the label distribution.

Additionally, we propose a Weighted Random Baseline that will return 'Disingenuous' or 'Not Disingenuous' based on the prevalence of each class in the training set (e.g., the training has 90% 'Disingenuous' instances, and 10% 'Not Disingenuous' instances, then the baseline returns 'Not Disingenuous' 90% of the time, and 'Disingenuous' 10% of the time), and a Frequentist Baseline that will always return the most frequent class present in the training set. These two baselines test the performance of knowledge of the label distribution without knowledge of linguistic features: to outperform these baselines, our models will need to learn actual linguistic features corresponding to a participant being 'Disingenuous'.

4.1.2 Human Baseline

Moreover, we test the performance of human annotators on our dataset. We constructed a web-based data annotation platform (Appendix A.2) in which users are presented with the transcript of a group discussion, and are asked to select which discussants - if any - are 'Disingenuous' in the context of the discussion. Thus, we attain human level annotations over the JUDGE-dis dataset. Due to limitations on the number of annotators we were able to recruit, we annotated only 93 of the 356 team deliberations, corresponding to 603 of the 2,348 jurors, and 3,871 of the 14,798 messages exchanged. Although the annotation is incomplete, given that the annotated samples were drawn I.I.D from the full dataset, we argue it is likely that human performance on this subsample closely resembles the results which would have obtained had participants analyzed the entire dataset.

4.2 Parameterized Models

4.2.1 Context-Free Classifier

To test the performance of context-free embeddings on this task (as part of H1), we first implement a Context-Free Classifier, using a BERT backbone to accept one tokenized statement as input, obtain textual embeddings, and pass those embeddings through a linear layer with a sigmoid activation in order to obtain a likelihood score that the statement is 'Disingenuous'.

4.2.2 Contextual Classifier

To compare the performance of the Context-Free Classifier against a contextual model (as part of H1), we next implement a Contextual Classifier. The Contextual Classifier has two input heads: (1) a *contextual embedder* accepts as input the tokenized statements from all jurors except the one in question, and (2) an *expressive embedder* accepts as input the tokenized statements from the juror in question. Each embedding head leverages BERT to produce return vectors corresponding to embeddings of the contextual text and statement text, respectively. These embeddings are then fused

and passed into a linear layer with a sigmoid activation in order to obtain a likelihood score that the juror is being 'Disingenuous' in the context of their discussion. Appendix A.1.1 contains an illustration of the structure of this class of model.

4.2.3 Augmented Contextual Classifier

Next, we implement the Augmented Contextual Classifier, designed to incorporate insights from the literature on deception. The Augmented Contextual Classifier, like the Contextual Classifier, contains both a *contextual embedder* and an *expressive embedder*; however, unlike the Contextual Classifier, it additionally contains four additional feature extractors corresponding to Niculae et al. (2015)'s *Linguistic Harbingers of Betrayal*:

First, disingenuous speakers are likely to be more positive, we include a *sentimental embedder* (a BERT model pre-trained on SST-2).

Second, because disingenuous speakers are more polite than others in the discussion, we include a *politeness differential feature* (which computes politeness scores of the contextual and speaker expressions according to Zhang et al. (2018) and calculates $Politeness(context) - Politeness(speaker)$ as a feature).

Third, because the talkativeness of the speaker can suggest disingenuity, we include a *talkativeness feature*, (which consists of the number of words of all expressions made by the speaker over the course of the discussion).

Regrettably, due to dataset access limitations, we could not include all features from Niculae et al. (2015): we've therefore omitted Subjectivity and Argumentativeness/Discourse as features, even though they are relevant features within Niculae et al.'s paper.

All of these features are fused, and, as in the Contextual Classifier, passed into a linear layer with a sigmoid activation in order to obtain a likelihood score that the juror is being 'Disingenuous' in their discussion. Appendix A.1.2 contains an illustration of the structure of this class of model.

4.3 Evaluation Metrics

People are largely truthful; with just 1.22% of statements classified as 'Disingenuous' and 1.25% of jurors classified as being 'Disingenuous', the JUDGE dataset is heavily imbalanced at both the utterance- and discussant-level. We therefore selected metrics by paying close attention to how metrics handle imbalanced datasets; we excluded ROC-AUC because the larger class has "dominant influence on the value of AUC" (Brabec and Machlica, 2018).

A key practice we wanted to avoid was giving high performance to a model that produces 'Not Disingenuous' 100% of the time — which could achieve high accuracy given the extreme imbalance. We thus chose macro-averaged F-2 scores as a primary evaluation metric. Macro-averaging enables us to weigh the classes equally instead of rewarding cases in which the 'Not Disingenuous' label dominates. Additionally, we choose F-2 to prioritize recall (and thereby penalize the practice of always returning the negative class). In addition to macro-averaged F-2, we also report F-1, Precision, Recall, and Accuracy to give further context.

Finally, we also apply Laplace smoothing, adding 1 to each entry in the confusion matrix before calculating precision and recall. We do this because, in some cases (e.g., the baseline models), returning only the negative class results in undefined metrics and makes bootstrapping and comparison impossible. A drawback is that Laplace smoothing creates a small distortion, and we explicitly note areas where this occurs. Since accuracy is unaffected by the problem of not having positives, we report Accuracy without Laplace smoothing.

5 Experiments

5.1 Transfer Learning Experimentation

5.1.1 Experiment 1: Does Context Matter?

As per H1, to determine the influence of context on the performance of the model, we first compare the performance of a Context-Free Classifier on the JUDGE-utt test set against that of a Contextual Classifier on the JUDGE-dis test set.

Both the Context-Free Classifier and Contextual Classifier were pre-trained for 100 epochs on the Peskov dataset. The Context-Free Classifier was then fine-tuned for 1 epoch on the JUDGE-utt training set, while the Contextual Classifier was fine-tuned for 1 epoch on the JUDGE-dis training set. Training and tuning for both models used binary cross entropy (BCE) loss, an AdamW optimizer with a learning rate $\alpha = 1\text{e}{-4}$, and early stopping (with a patience of 5). Due to the imbalance of the dataset, a balanced sampler (with replacement) was implemented, so that the each example within each batch was sampled with a likelihood inversely proportional to the frequency of its label in the dataset. In this way, we could ensure that, even though the Peskov dataset is mildly imbalanced,

Experimental Results							
Model	Test Set	Training Scheme	F-2	F-1	Acc.	Prec.	Rec.
Naive Random Baseline	JUDGE-utt	N/A	0.068	0.029	0.447	0.015	0.516
Weighted Random Baseline	JUDGE-utt	N/A	0.010	0.016	**0.984**	0.500	0.008
Frequentist Baseline	JUDGE-utt	N/A	0.010	0.016	**0.984**	0.500	0.008
Naive Random Baseline	JUDGE-dis	N/A	0.069	0.030	0.458	0.016	0.476
Weighted Random Baseline	JUDGE-dis	N/A	0.058	**0.087**	**0.984**	0.500	0.048
Frequentist Baseline	JUDGE-dis	N/A	0.058	**0.087**	**0.984**	0.500	0.048
Human Baseline	JUDGE-dis	N/A	0.074	0.039	0.839	**0.022**	0.182
Context-Free Classifier	JUDGE-utt	Transfer Learn.	0.054	0.024	0.576	0.012	0.320
Contextual Classifier	JUDGE-dis	Transfer Learn.	0.082	0.040	0.755	0.021	0.286
Aug. Contextual Classifier	JUDGE-dis	Transfer Learn.	**0.084**	0.036	0.326	0.019	0.714
Context-Free Classifier	JUDGE-utt	MAML	0.072	0.034	0.751	0.018	0.270
Contextual Classifier	JUDGE-dis	MAML	0.057	0.026	0.689	0.014	0.238
Aug. Contextual Classifier	JUDGE-dis	MAML	0.079	0.033	0.016	0.017	**0.952**

Table 2: Tabular summary of experimental results, with highest results bolded. All columns except for Accuracy reflect macro-averaged values with Laplace smoothing. Note that the 0.5 precision for the Weighted and Frequentist Baselines are inflated due to Laplace Smoothing $(1/(1+1) = 0.5)$, since the baselines otherwise returned no positive values. The next highest precision value (which we have bolded instead) is that of the Human Baseline.

and the JUDGE dataset is significantly imbalanced, our model still learns from both the positive and negative examples in equal proportion.

5.1.2 Experiment 2: Does Human-Level Augmentation Improve Performance?

To test H2, we compare the performance of a Contextual Classifier against that of an Augmented Contextual Classifier on the JUDGE-dis test set. As in Experiment 1, the Augmented Contextual Classifier was trained for 100 epochs on the Peskov dataset, then fine-tuned for 1 epoch on the JUDGE-dis training set. Training and tuning used BCE loss, an AdamW optimizer with a learning rate of $\alpha = 1\text{e}{-4}$, early stopping (patience of 5), and balanced batch sampling. The performance of the Augmented Contextual Classifier on the JUDGE-utt test set was then compared against that of the Contextual Classifier from Experiment 1.

5.1.3 Experiment 3: Does Meta Learning Improve Performance in the Few-Shot Setting?

To test H3, we train a Context-Free Classifier, a Contextual Classifier, and an Augmented Contextual Classifier using MAML (Finn et al., 2017), and compare the results against those from Experiment 2, when we trained these same models using Transfer Learning. Unlike in the prior two experiments, in which the models were pre-trained on the Peskov dataset, in this experiment, we instantiated both the JUDGE(-dis for the Context-Free Classifier, -utt for the Contextual and Augmented Contextual Classifiers) and Peskov datasets as meta tasks from which samples would be drawn during each iteration of the MAML outer training loop. Each iteration of MAML, we sample a batch of tasks, where each task $\mathcal{T}_i$ is defined by dataset $\mathcal{D}_i$ and consists of m samples drawn from $\mathcal{D}_i$ ($x_{\mathcal{T}_i} \in R^{m \times d}$, $y_{\mathcal{T}_i} \in [0,1]^m$), and a loss function $\mathcal{L}_{\mathcal{T}_i}$ (which, in our case, is always the BCE loss). Formally, we have:

$$\mathcal{T}_i = (x_{\mathcal{T}_i}, y_{\mathcal{T}_i}, \mathcal{L}_{\mathcal{T}_i})$$

We perform *stratified task sampling*: each iteration, we sample two tasks, $\mathcal{T}_1, \mathcal{T}_2$, where, without loss of generality, $\mathcal{T}_1$ is defined over the JUDGE dataset, and $\mathcal{T}_2$ is defined over the Peskov dataset.

For all models, our MAML implementation was trained for 25 steps using BCE loss, an AdamW outer optimizer (learning rate $\beta = 1\text{e}{-3}$) and an SGD inner optimizer (learning rate $\alpha = 0.05$; with 5 adaptation steps per inner loop). In our implementation, each task consists of 10 examples, 3 of which are allocated into the support (meta-training) set, 7 of which are allocated into the query (meta-testing) set. Each example is sampled into a given task $\mathcal{T}_i$ with a likelihood inversely proportional to the frequency of its label within $\mathcal{D}_i$, to ensure that the model learns from both positive and negative examples within each task. Finally, we applied a weighting to the loss over each task when computing the total loss in each outer step: given that the task in question was the JUDGE dataset (while Peskov was primarily used to provide additional background data from a similar - though not identical - domain), we added weights to each task loss when computing the outer MAML loss to emphasize performance on the JUDGE dataset. Specifically, we modify the outer gradient update step as

follows:

$$\theta \leftarrow \theta - \beta \nabla_\theta \big(\mathcal{L}_{\mathcal{T}_1}(f_{\theta'_1}) + \gamma \mathcal{L}_{\mathcal{T}_2}(f_{\theta'_2}) \big)$$

Here, f, θ, θ' are as in Finn et al. (2017), $\mathcal{T}_1, \mathcal{T}_2$ are as above, and $\gamma = 0.25$ is a discounting factor.

6 Experimental Results

Table 2 includes a summary of our experimental results, which are also plotted in Figure 2 (additional plots are included in Appendix A.3). Overall, we observe that, across the baselines and our models, performance is quite poor. The maximum F-2 score attained is only 0.084 — highlighting the difficulty of our task. Therefore, our ability to out-perform humans is a noteworthy contribution.

Our models and baselines fail in telling ways: in general, *the machine learning models tended to have lower precision*, as they erred on the side of predicting the positive class more frequently; likely a result of the balanced sampling during training. Excluding the Weighted and Frequentist baselines (see note in Table 2's caption), the highest precision was that of the the human baseline, at only 0.022. On the other hand, *the machine learning models tended to do better on recall*; the Augmented Contextual Classifier with MAML had a recall of 0.952, and the Augmented Contextual Classifier with transfer learning had a recall of 0.714. These strongly out-perform the baselines, in which the highest, the Naive Random baseline, had recall of 0.476, but come at the expense of a large number of false positives.

6.1 H1: Context Matters

Ultimately, we find that H1 is supported. We observe that the Context-Free Classifier (F-2 = 0.054) underperforms the Contextual Classifier (F-2 = 0.082), suggesting that predictions are far better with added conversational context. We additionally note that, while the Context-Free Classifier underperformed the statement-level Naive Random Baseline (F-2 = 0.068), the Contextual Classifier outperformed not only the random baselines, but also humans on the task defined in JUDGE-dis, further supporting H1.

6.2 H2: Augmenting with Human Insights is Useful.

We also find support for H2. Within transfer learning-trained models, the model with human augmentation outperforms the model without: the

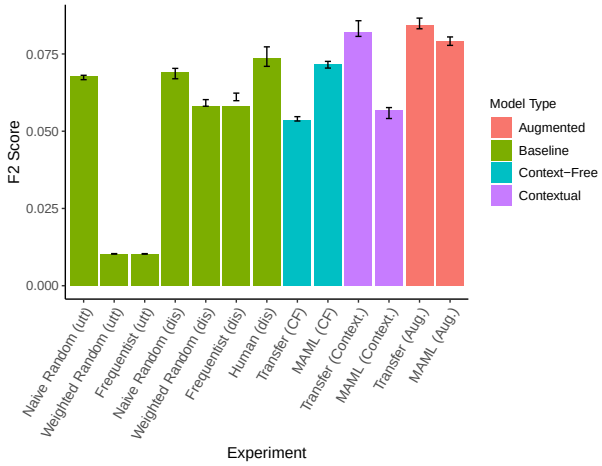


Figure 2: A bar chart showing the results from seven baselines and five models, with bootstrapped 95% confidence intervals with 500 samples. The primary metric is the macro-averaged F-2 score with Laplace smoothing. Baselines for JUDGE-utt are marked (utt), and baselines for JUDGE-dis are marked (dis).

Contextual Classifier had an F-2 of 0.082; the Augmented Classifier performs slightly better, with F-2 of 0.084. Within MAML-trained models, the model with human augmentation very significantly outperformed the one without; while the augmented MAML model attained an F-2 of 0.079, which beat the human baseline of 0.074, the non-augmented model attained an F-2 of only 0.057 — substantially underperforming both the Naive Random Baseline and the Human Baseline. Thus, H2 is supported.

6.3 H3: Meta-Learning Improves Performance Only When Context Isn't Given.

H3 was not supported. We hypothesized that MAML models would greatly outperform their Transfer Learning counterparts, but instead saw mixed results. While the Context-Free Classifier with MAML (F-2 = 0.072) outperformed its transfer learning counterpart (F-2 = 0.054), MAML performed very poorly for the Contextual Classifier. The ostensibly strong performance for the Augmented Contextual Classifier (recall = 0.952) appears to be due only to the model predicting 'Disingenuous' *100%* of the time (the reason recall is not 1.0 is because of Laplace smoothing).

7 Analysis

In this section, we analyze our models' performance and acknowledge limitations of this work.

7.1 H1: Why does BERT Underperform Against Baselines?

We suggest two reasons that BERT underperformed its Human, Contextual, and Augmented Contextual counterparts. First, the labels of JUDGE-utt were coarser — and therefore, noisier — than those of JUDGE-dis: recall that JUDGE-utt labels *all* statements made by a disingenuous speaker as disingenuous, but this assumption is flawed: disingenuous speakers make some true statements. This method of labeling reduced the ability of BERT to cleanly delineate between the two classes. Second, as stated in H1, social influence-based disingenuity is, eponymously, a *social* phenomenon: understanding and detecting it requires additional contextual insight which was denied to the BERT model which was tuned/evaluated on JUDGE-utt.

7.2 H2: Why Did Human Input Boost Performance?

To better understand this result, we investigate the weights in the final linear layer of the Augmented Contextual Classifier (Transfer Learning) corresponding to two of the augmentation features. We found that the talkativeness weight is $-8.71\text{e}{-8}$; more talkative participants were considered more 'Not Disingenuous.' These findings align with our expectations from Niculae et al. (2015)'s *Linguistic Harbingers of Betrayal*, as well as Chen et al. (2020)'s findings on hesitation as a deception tell. Since Augmented Contextual Classifier (Transfer Learning) correctly identifies a trait of deception cited in the literature, the talkativeness weight could have contributed to high performance. However, the politeness weight ($-1.15\text{e}{-2}$) is unexpected; we believed, in accordance with Niculae et al. (2015) literature, that 'Disingenuous' participants would be more polite. This leaves an open question as to whether this is weight is an artefact of our training process, or whether it implies that politeness is less of an indicator of disingenuity than prior literature would imply.

We omit analysis of the features weights of the Augmented Contextual Classifier (MAML), as the model failed to show discretion (it predicted 'Disingenuous' every time).

7.3 H3: Why does MAML Underperform on Contextual Models?

We next consider why MAML underperforms on contextual models. We hypothesize that it is a function of the task sampling process: given that there are 10 disingenuous training examples in JUDGE-dis, and given that we sampled $\mathcal{T}_2$ (which contained 5 disingenuous training examples in expectation) each iteration, we suspect that the process of showing these same few examples to the model over many MAML iterations led the contextual architectures to overfit to the JUDGE-dis training set. If we had additional data, we could test this hypothesis by defining a validation set over JUDGE-dis, and observing the differences in training/validation loss on JUDGE-dis during the meta-training process.

7.4 General Limitations

A fundamental limitation of this task is searching for already-subtle signals in a sparse field. To improve disingenuity detection on textual data, future work will require *far* more data, with *far* more annotations. Another avenue is improving the metrics; though we designed our metrics to avoid a "only return negative" scenario, the Augmented MAML model returned only positives and performed well — suggesting that the metrics require improvement. We also suggest deeply investigating layers of our models to understand whether they have truly learned to detect the subtleties of disingenuity, or merely memorized examples. Section 7.2 outlines a few ways in which our model appears to have learned socially interpretable patterns, but much work remains.

8 Conclusion

We make the following contributions: to our knowledge, we are the first to approach the task of detecting social influence-based disingenuity using NLU techniques. We present the JUDGE dataset, over which we define two tasks (JUDGE-utt, JUDGE-dis), and we benchmark performance on those tasks using state-of-the-art NLU models, and (in the case of JUDGE-dis) human performance. Overall, we observe low performance on these tasks by both models and humans, although our top-performing models outperform humans. Our results confirm that we have defined a challenging task, and further research in NLU will be required to obtain strong performance on this task.

9 Acknowledgements

We would like to give thanks to Dora Demszky and Dr. Christopher Potts, who advised this project.

10 Authorship Statement

Michael Cooper implemented the Context-Free Classifier, Contextual Classifier, Augmented Contextual Classifier, transfer learning training procedure, MAML training procedure, and online annotation platform.

Emily Hu implemented the data scraping/cleaning, and train/test split for the JUDGE dataset; the baseline models; the evaluator (which executes fine-tuning and testing); and plotting of figures for the paper.

Makena Low handled porting the models written by Emily and Michael to a GPU on AWS. She ran the training and evaluation experiments. She also performed feature analysis.

The final paper was jointly authored.

References

Solomon E Asch. 1961. Effects of group pressure upon the modification and distortion of judgments. In *Documents of gestalt psychology*, pages 222–236. University of California Press.

Jan Brabec and Lukas Machlica. 2018. Bad practices in evaluation methodology relevant to class-imbalanced problems. *arXiv preprint arXiv:1812.01388*.

Xi Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic, and Julia Hirschberg. 2020. Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, 8:199–214.

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*.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

J Richard Hackman and Nancy Katz. 2010. Group behavior and performance.

Reid Hastie, Steven D Penrod, and Nancy Pennington. 2013. *Inside the jury*. Harvard University Press.

Xinlan Emily Hu, Mark E Whiting, and Michael S Bernstein. 2021. Can online juries make consistent, repeatable decisions? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

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. *arXiv preprint arXiv:1506.04744*.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. *arXiv preprint arXiv:2003.02739*.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2336–2346.

Denis Peskov and Benny Cheng. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL*.

Stephen Porter and Leanne ten Brinke. 2010. The truth about lies: What works in detecting high-stakes deception? *Legal and criminological Psychology*, 15(1):57–75.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Franco Salvetti, John B Lowe, and James H Martin. 2016. A tangled web: The faint signals of deception in text-boulder lies and truth corpus (blt-c). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3510–3517.

Frederick A Siegler. 1966. Lying. *American Philosophical Quarterly*, 3(2):128–136.

Jae-Young Son, Apoorva Bhandari, and Oriel Feldman-Hall. 2019. Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific reports*, 9(1):1–15.

Cass R Sunstein. 2008. Why groups go to extremes. *Yale Law Journal (*, 110:2000.

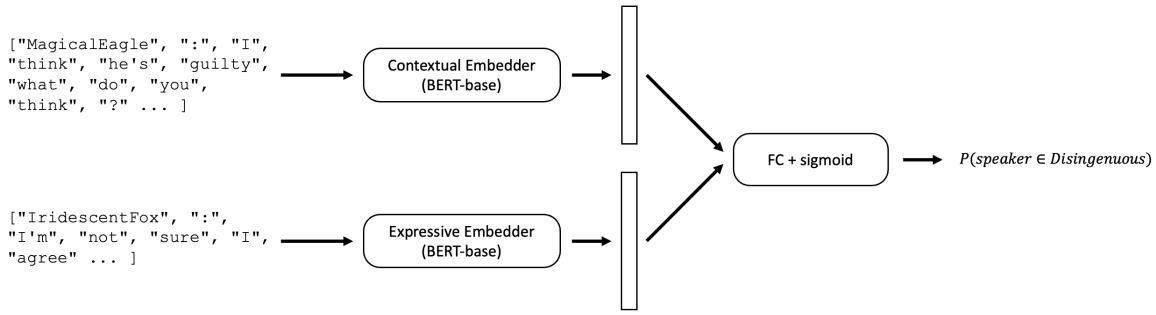
Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

A Appendix

A.1 Model Architectures

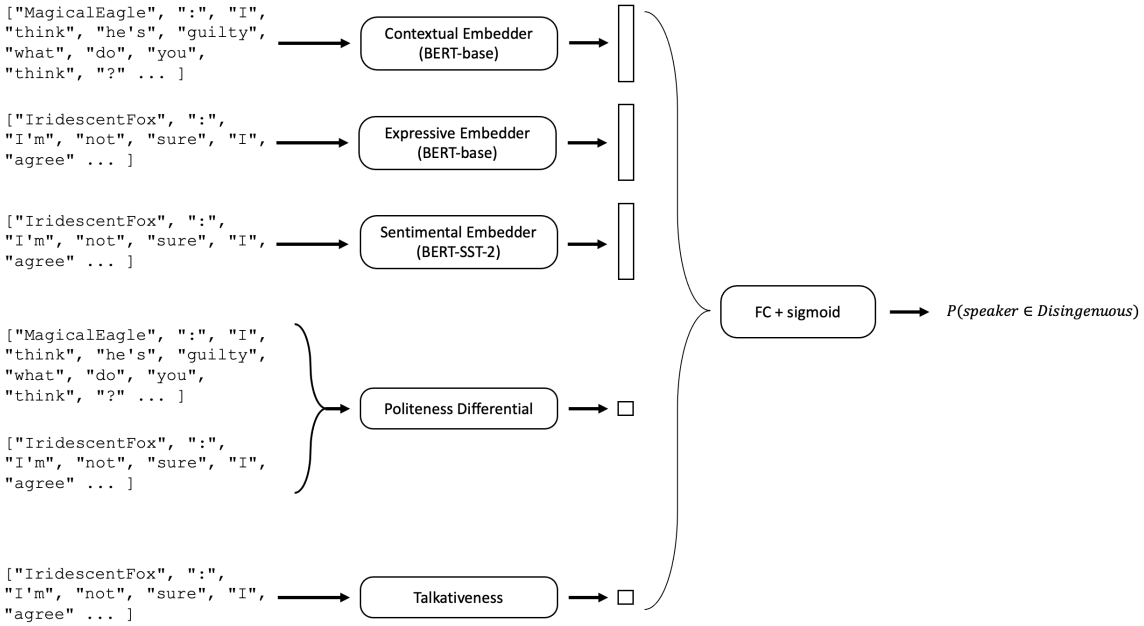
Below, we present architectural diagrams for the Contextual Classifier (Section A.1.1) and for Augmented Contextual Classifier (Section A.1.2).

A.1.1 Contextual Classifier



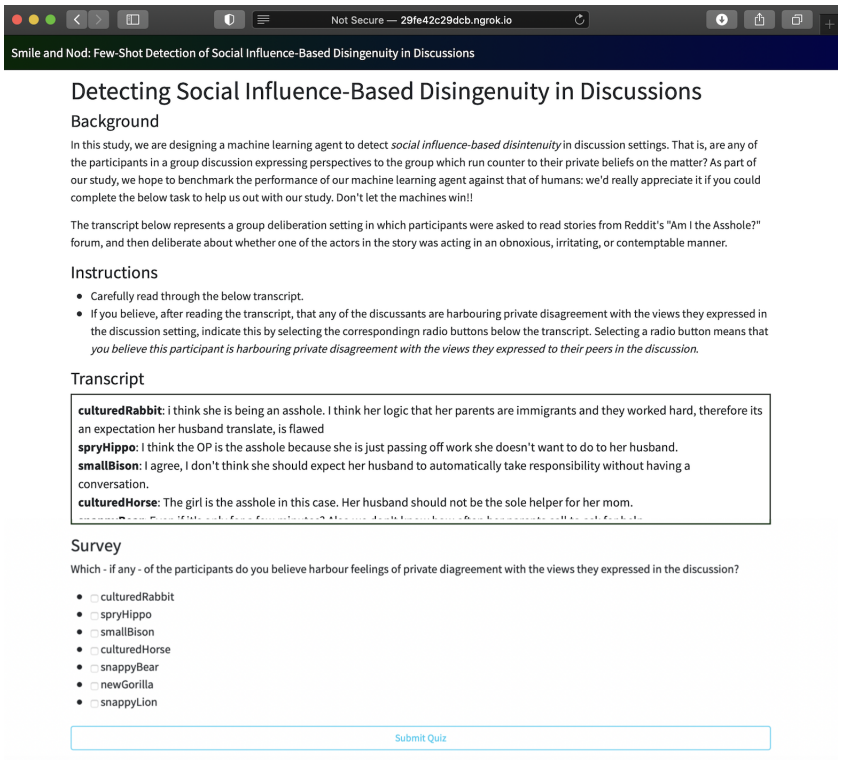
Simplified diagram of the Contextual Classifier model architecture. Here, the speaker is `IridescentFox`.

A.1.2 Augmented Contextual Classifier



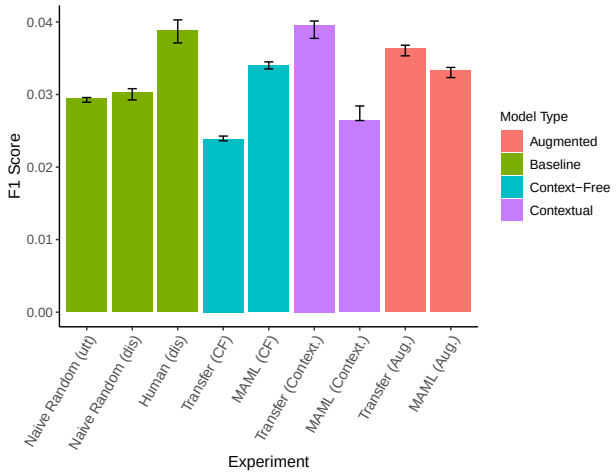
Simplified diagram of the Augmented Contextual Classifier architecture. Here, the speaker is `IridescentFox`.

A.2 Web Annotation Platform

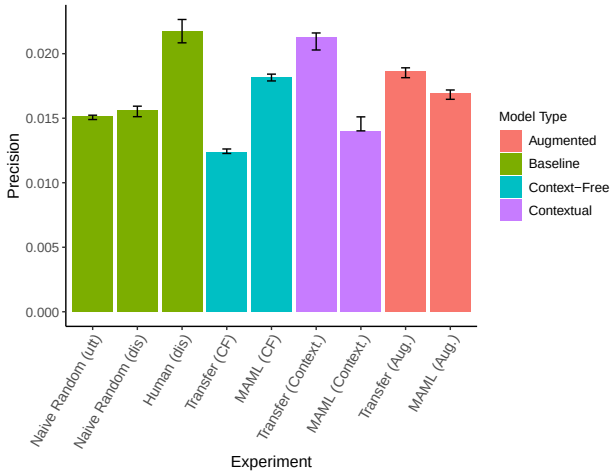


A screenshot of the web annotation platform that we built in order to obtain human annotations over our dataset.

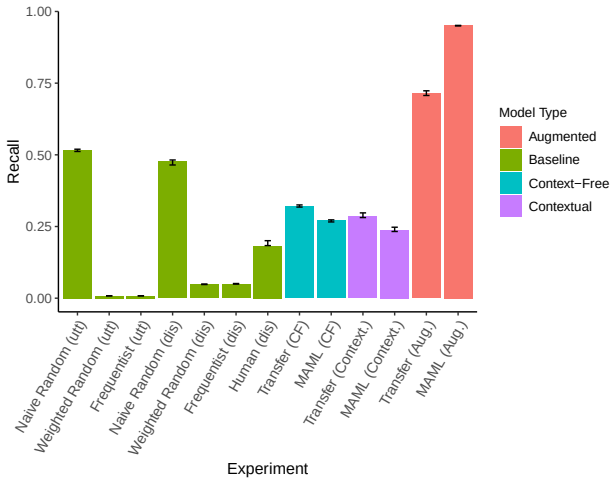
A.3 Additional Results Plots



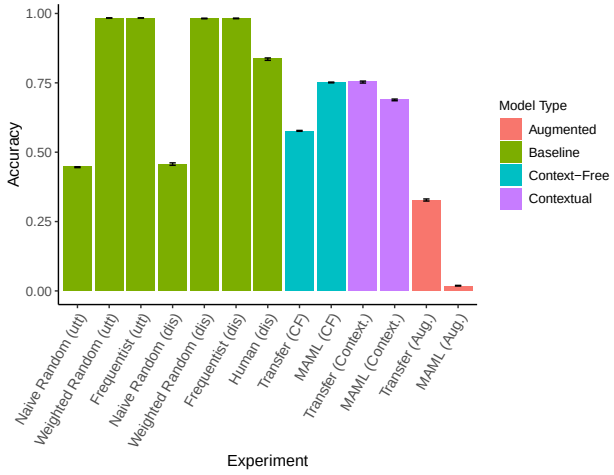
A bar chart showing the results from three baselines and five models, with bootstrapped 95% confidence intervals with 500 samples. The metric shown is the **F-1**. The Frequentist and Weighted Baselines have been removed, because bootstrapping makes the F1 higher than reality.



A bar chart showing the results from three baselines and five models, with bootstrapped 95% confidence intervals with 500 samples. The metric shown is **Precision**. The Frequentist and Weighted Baselines have been removed, because bootstrapping makes the Precision 0.5 when it is in fact undefined.



A bar chart showing the results from seven baselines and five models, with bootstrapped 95% confidence intervals with 500 samples. The metric shown is **Recall**.



A bar chart showing the results from seven baselines and five models, with bootstrapped 95% confidence intervals with just 50 samples. The metric shown is **Accuracy**.

A.4 Metrics for the Deception in Diplomacy Dataset

	# of statements	# of players
Lie	591	196
Non-lie	12541	165

Table 3: Data metrics for the Deception in Diplomacy training set.

	# of statements	# of players
Lie	56	19
Non-lie	1360	21

Table 4: Data metrics for the Deception in Diplomacy validation set.

	# of statements	# of players
Lie	240	38
Non-lie	2501	45

Table 5: Data metrics for the Deception in Diplomacy test set.