Don't Let Me Be Misunderstood: Comparing Intentions and Perceptions in Online Discussions

Jonathan P. Chang
Cornell University
jpc362@cornell.edu

Justin Cheng
Facebook
jcheng@fb.com

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

ABSTRACT

Discourse involves two perspectives: a person's intention in making an utterance and others' perception of that utterance. The misalignment between these perspectives can lead to undesirable outcomes, such as misunderstandings, low productivity and even overt strife. In this work, we present a computational framework for exploring and comparing both perspectives in online public discussions.

We combine logged data about public comments on Facebook with a survey of over 16,000 people about their intentions in writing these comments or about their perceptions of comments that others had written. Unlike previous studies of online discussions that have largely relied on third-party labels to quantify properties such as sentiment and subjectivity, our approach also directly captures what the speakers actually intended when writing their comments. In particular, our analysis focuses on judgments of whether a comment is stating a fact or an opinion, since these concepts were shown to be often confused.

We show that intentions and perceptions diverge in consequential ways. People are more likely to perceive opinions than to intend them, and linguistic cues that signal how an utterance is intended can differ from those that signal how it will be perceived. Further, this misalignment between intentions and perceptions can be linked to the future health of a conversation: when a comment whose author intended to share a fact is misperceived as sharing an opinion, the subsequent conversation is more likely to derail into uncivil behavior than when the comment is perceived as intended. Altogether, these findings may inform the design of discussion platforms that better promote positive interactions.

ACM Reference Format:
Jonathan P. Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. 2020. Don't Let Me Be Misunderstood: Comparing Intentions and Perceptions in Online Discussions. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380273

1 INTRODUCTION

> "I'm just a soul whose intentions are good..."
>
> – Nina Simone, "Don't Let Me Be Misunderstood"

Conversations, both online and offline, fundamentally involve two perspectives: a speaker's *intention*—that is, the goals they seek to achieve through their utterance—and others' *perception* of the speaker's words [26]. When the intentions and perceptions of participants in a conversation are misaligned [14, 59], undesirable outcomes ranging from low productivity to overt strife can occur [46, 58].

One important type of misalignment occurs when people confuse *facts*[1] and *opinions* [53]. A Pew survey on online news consumption found that 65% of Americans mistakenly perceived opinions extracted from online news as facts, while 75% took facts to be opinions [47]. In this work, we investigate how this type of misalignment plays out in online public discussions where people engage with each other rather than passively consuming content. How often and under what circumstances does a speaker's intended statement of a fact get misperceived as an opinion? How does such misalignment tie into the quality of online discourse?

Answering such questions requires ground truth data both on what the speaker's intention was in crafting an utterance and on how that utterance was perceived by others. While perceptions can be approximated through third-party annotation [38, 54, 64], only the speakers themselves know what their actual intentions were. To obtain ground truth about both intention and perception, we surveyed over 16,000 people about their intention in writing public comments on Facebook or about how they perceived comments to which they had replied. Combining these surveys with data about these conversations then allows us to compare how facts and opinions are intended and perceived in different contexts.

We start with a high-level approach exploring differences in the distributions of intentions and perceptions. We find, for example, that in online discussions people perceive opinions at a higher rate than they are intended. Next, we investigate this apparent incongruency at a linguistic level. While linguistic cues developed to capture subjectivity can generally distinguish between facts and opinions, salient differences arise when considering how utterances are intended rather than how they are perceived. For instance, the explicit use of factual language (e.g., "In fact, ...") signals that the speaker intended to make a factual claim, but not that others will perceive it as such. Starting from these insights, we assess the extent to which third-party perception labels (as used in prior work on subjectivity detection) are interchangeable with author-sourced intention labels when predicting intentions from text.

Finally, we examine how differences between intentions and perceptions relate to the outcome and quality of online public discussions. The trajectory of a conversation likely depends on its intended starting point as well as on how others perceive it. For example, past work on Wikipedia discussions found that conversations starting with factual checks are more likely to turn uncivil than those appearing to share or seek opinions [69], and qualitative

[1] Following prior literature, we consider facts to be statements that could in principle be conclusively proven or disproven based on evidence regardless of their veracity [16, 47].

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380273

studies have suggested that the alignment between the intentions and perceptions of participants in a conversation is key to keeping it on track [14, 58].

We find that both the intended and perceived *goals* of the initial comment are indicative—in potentially different ways—of a conversation's future trajectory. While certain intended goals (e.g., sharing an opinion) are more likely to lead to uncivil behavior in the subsequent discussion, even more significant is whether other participants in the conversation perceive those goals as they were intended. For instance, when the initial commenter intends to share a fact but this goal is *misperceived* by others, the conversation is more likely to turn uncivil than when that initial intention is correctly perceived.

Taken together, these findings have potential implications for designing public discussion platforms that can better promote healthier conversations. They show that when assessing the likely trajectory of a conversation, both a speaker's intention and how that intention might be (mis)perceived by others are important factors to consider. More generally, by exposing differences between ground-truth labels coming from a first-person perspective and those coming from a (more easily accessible) third-person perspective, these results also signal potential biases that systems relying only on one of the two types of labels might incur.

To summarize, we:

- Present a large-scale study comparing intentions and perceptions of facts and opinions in online conversations and their relation to conversational outcomes;
- Identify and compare linguistic cues that signal intentions and perceptions, showing that linguistic differences between the two translate to differences in classifier behavior when training on one versus the other; and
- Show that conversations in which fact-sharing intentions are misperceived as opinion-sharing are more likely to turn uncivil later on.

2 GROUND TRUTH INTENTIONS AND PERCEPTIONS

Though studies of online discussions often use third-party annotation to identify properties of interest (e.g., opinions [65]), this can only capture perceptions and not intentions, as only the original author of a comment knows with certainty what they intended. As such, we instead surveyed comment authors directly to gather intention labels. In this section, we describe our conversational data and survey methodology, and give high-level descriptive statistics of the survey responses. All data was de-identified and analyzed on Facebook's servers, and an internal research board reviewed the study design ethics and privacy practices prior to its start.

2.1 Conversational data

This work focuses on public discussions taking place in the comments sections of posts on Facebook Pages, which typically represent brands, media outlets (including but not limited to news), public figures, or communities. Anyone can view or join these discussions, so they offer a diverse sample of data for comparing intentions and perceptions of facts and opinions.

But as replies are rare on social media [3], most comments are unlikely to be part of conversations. As a heuristic for finding conversations in comment sections, we searched for a reciprocity pattern: one person (the *initiator*) makes a comment, a different person (the *replier*) replies to the initiator's comment, and then the initiator follows up by either reacting to or replying to the replier.

We constructed our conversational dataset by finding comment threads on English-language Page posts that begin with this reciprocity pattern. Data was collected from a 1.5 month window spanning mid-May through June 2019, resulting in approximately 22 million candidate conversations taking place across 3 million posts on nearly 800,000 pages.

2.2 Intentions and perceptions surveys

Selecting survey participants. Starting from this conversational dataset, we created two survey participant pools: a pool of initiators who would receive a survey asking about their intentions, and a pool of repliers who would receive a survey asking about what they perceived to be the intention underlying the comment they replied to. To minimize demand effects, we filtered out conversations where the initiator and replier were friends on Facebook. We additionally filtered out any cases where at least one comment was no longer publicly viewable. To ensure diversity in the participants and types of conversations we asked about in the surveys, we imposed a limit on how many participants could be selected from any given Page: up to 1% of a Page's followers, capped at 10.

Survey design. The surveys for initiators and for repliers both asked about facts and opinions in the initiator's opening comment (i.e., the first comment of the reciprocal chain). While in the context of monologic text—such as news articles and reviews—subjectivity mainly concerns the *sharing* of facts versus opinions [38, 65, 68], in a conversational setting, participants can also *seek* factual information or others' opinions [49, 51]. As such, both surveys distinguished between sharing and seeking of opinions and facts. Finally, prior work identifies humor as a prominent axis that is orthogonal to opinions and facts [48], so we additionally included it in our survey for completeness.[2] These considerations result in five *goals* that could be intended or perceived: (1) fact sharing, (2) fact seeking, (3) opinion sharing, (4) opinion seeking, and (5) humor.

The *initiator survey* asked initiators to rate their opening comment along each goal using a five-point Likert scale. Analogously, the *replier survey* asked repliers to rate their interpretation of the initiator's comment along each goal. Some subsequent analyses will simplify the responses by binarizing them: we will say that an initiator *intended (or perceived) a goal* if they responded with "mostly" or "definitely", and *did not intend (or perceive) a goal* otherwise.

To better understand the relationship between intentions and conversational outcomes, both surveys also asked participants to rate the conversation along two axes: whether it was worth their time, and whether they felt understood (initiators) or found the other person easy to understand (repliers). This results in a total of seven questions per survey (Table 1). Although both surveys asked about the initiator's opening comment, for context the survey

[2] In our findings, humor is nevertheless rare, being both intended and perceived in only about 10% of cases, and so is excluded from most subsequent analyses.

Question	*Initiator survey*	*Replier survey*
	Goals	
Opinion sharing	When you started the interaction, were you trying to express an opinion?	Do you think the other person was trying to express an opinion?
Opinion seeking	When you started the interaction, were you looking for other people's opinions?	Do you think the other person was looking for opinions?
Fact sharing	When you started the interaction, were you trying to provide information (for example, sharing a fact)?	Do you think the other person was trying to provide information (for example, sharing a fact)?
Fact seeking	When you started the interaction, were you looking for information?	Do you think the other person was looking for information?
Humor	When you started the interaction, were you trying to make a joke? Not at all / Mostly not / Somewhat / Mostly / Definitely	Do you think the other person was joking? Not at all / Mostly not / Somewhat / Mostly / Definitely
	Outcomes	
Time-worthiness	Looking back on this interaction, do you think it was worth your time? Not at all / Mostly not / Somewhat / Mostly / Definitely	Looking back on this interaction, do you think it was worth your time? Not at all / Mostly not / Somewhat / Mostly / Definitely
Understanding	How well do you feel your goals and intentions in this interaction were understood? Not understood at all / Not very well understood / Somewhat understood / Mostly understood / Very well understood	Overall, how difficult was it for you to guess the other person's goals and intentions? Not difficult at all / Not very difficult / Moderately difficult / Very difficult / Extremely difficult

Table 1: The initiator survey asked participants about their intentions with respect to a comment they had written, while the replier survey asked participants about their perceptions of a comment that they had replied to. Answers are shown in gray.

participants were also shown the reply and the Page post on which the conversation took place (Figure 1).

Low response rates made it infeasible to obtain paired initiator-replier responses (i.e., responses from both the initiator and replier on each conversation). As a result, we ran the initiator and replier surveys on disjoint conversations. While we conducted third-party annotation to study the relationship between intentions and perceptions in the same conversation (Section 4), exploring other ways to address this limitation would be valuable future work.

Running the survey. Participants were recruited for both surveys via an ad on Facebook targeted at a random sample of English-speaking people, which ran for two weeks in early July 2019. Each survey was opt-in, and participants could choose to stop at any time. Other than the ad and survey, participants' Facebook experience was not altered or manipulated in any way.

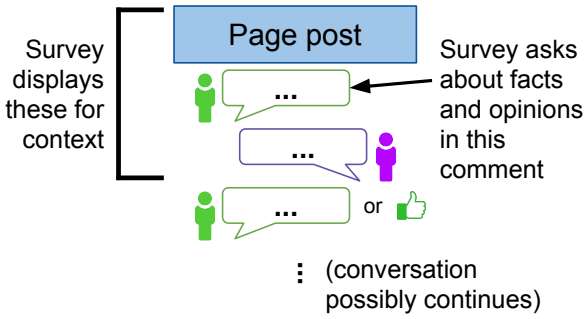

Figure 1: We surveyed on conversations containing *reciprocity*: an initiator (green) makes a comment, a replier (purple) replies to the initiator, and the initiator follows up with another comment or a reaction. Surveys asked about facts and opinions in the initiator's opening comment, though for context the survey participant was additionally shown the reply and the Page post on which the exchange took place.

2.3 Participant statistics

9,174 people completed the initiator survey, while 7,129 people completed the replier survey. On average, participants were 5.6 years older and 1.6% more likely to be female compared to the average Facebook Page commenter. To test if the amount of time between when someone commented or replied and when they took the survey affected their responses, we calculated correlations with this time gap, finding that they are negligible ($r \leq 0.02$, n.s.).

The lower response rate for the replier survey suggests that questions about one's own intentions are easier to answer than questions about perceiving others' intentions. To test for a response bias, we examined demographic differences between the two surveys. We find small but significant differences for age ($D = 0.03$, $p < 0.001$ using a K-S test) and gender ($\chi^2 = 3.97$, $p = 0.05$). Though these differences are small, they may have an effect on results if responses vary significantly across demographics. Thus, we next examine how age and gender may affect intentions and perceptions.

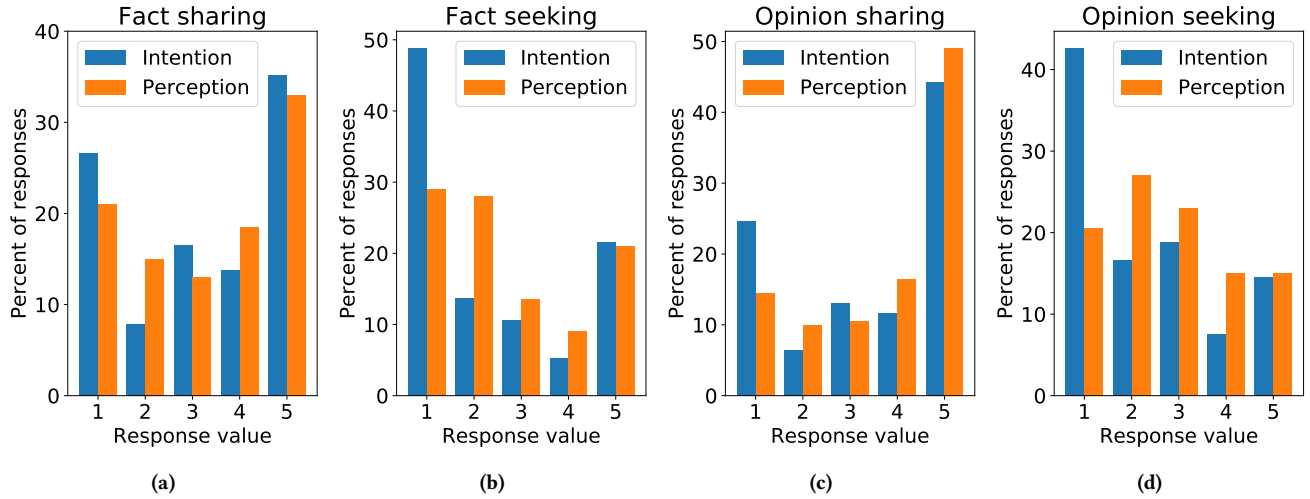

Figure 2: Distributional differences in how people intend their own comments and perceive others' comments. For example, people perceive an opinion sharing goal more often than it is intended (c). A response value of 1 corresponds to "did not intend at all", while a response value of 5 corresponds to "definitely had this intent".

Older people are less likely to intend to seek facts (Spearman's R = -0.08, $p < 0.001$) or opinions (Spearman's R = -0.09, $p < 0.001$), which may be partly explained by previous work showing that older people tend to prefer passive learning (e.g., through reading) over learning through direct interaction [24]. They are also more likely to perceive others as sharing facts (Spearman's R = 0.12, $p < 0.001$), echoing prior research that showed older people may be more inclined to treat a statement as factual [27]. We also find that men were less likely to intend to seek opinions from others (Mann-Whitney U = 10820738.5, $p < 0.001$).

To address these potentially confounding effects, in subsequent analyses we control for demographic differences (as well as descriptive properties of the Page, namely size and category) as appropriate. We further note that although the demographic effects are statistically significant, demographic and Page features are nonetheless poor predictors of intention and perception,[3] and in practice we find that uncontrolled versions of each analysis yield similar results.

3 INTENTIONS VERSUS PERCEPTIONS

To understand if there is a systematic misalignment between intentions and perceptions in the context of online public discussions, we (a) compare response frequencies among the initiator survey responses and replier survey responses (i.e., how often a goal is actually intended versus how often it is perceived), (b) consider linguistic cues that are indicative of a goal and explore whether these are different for intended versus perceived goals, and (c) examine how intentions and perceptions may differ in their relationship to the trajectory of the conversations in which they are observed.

[3] $R^2 \leq 0.04$ in regressions predicting intention or perception using gender, age, Page size (number of followers), and Page category (e.g., "sports").

3.1 Distributional differences

If perceptions perfectly captured intentions, the overall distribution of responses for intentions and perceptions of each goal would be nearly identical. But if intentions and perceptions are misaligned, then we may observe systematic differences between the two response distributions. As such, our first analysis compares response distributions of intention and perception for each goal. We controlled for demographic differences between surveys by reweighting the perception survey responses to match the age and gender distribution of the intention survey via post-stratification [60].

These distribution comparisons are visualized in Figure 2. This data exposes two types of distributional differences: (a) systematic overestimation, in which perceivers judge a particular goal to occur more frequently than it is actually intended, and (b) uncertainty, in which perceivers are unsure of people's intentions and hence tend to pick less definite response choices.

Systematic overestimation. Systematic overestimation occurs when a goal is perceived to occur more frequently than it is actually intended. This can be formalized as the mean response for perception being significantly larger than the mean response for intention. Under this definition, overestimation occurs for opinion sharing (mean perception response = 3.8, mean intention response = 3.4), corroborating prior work which found that people were more likely to misidentify factual statements as opinions than vice versa [47, 53]. Overestimation also occurs for fact seeking (2.7 vs 2.4) and opinion seeking (2.8 vs 2.4), but not fact sharing. These differences are significant at $p < 0.001$ via Mann-Whitney U test.

Difference in certainty. In several of the distributions in Figure 2, intention survey responses are often more likely to take on one of the extreme rating options ("not at all" or "definitely") compared to perception responses. Conversely, perception responses are often more likely to display some degree of hedging by giving ratings of "mostly not", "somewhat", or "mostly". We refer to this as a difference in certainty, and formalize it as follows: for each goal, we compute

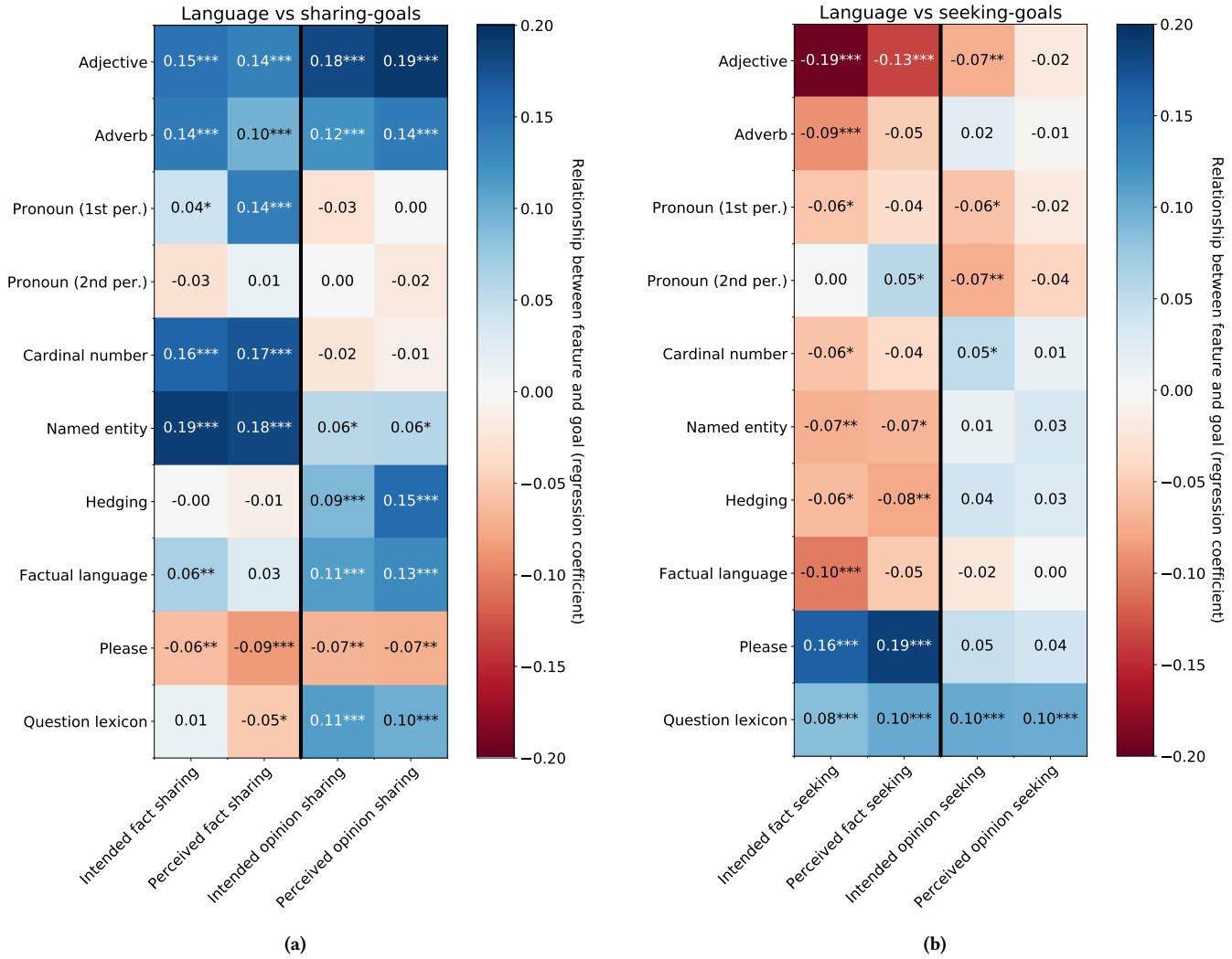

Figure 3: Regression analysis comparing the rates of linguistic features in intention and perception of sharing-goals (a) and in intention and perception of seeking-goals (b). Observed correlations are generally consistent with results from literature on subjectivity detection. While most features correlate similarly between intents and perceptions, there are some that differ, including factual language and use of first or second person pronouns.

the proportion of uncertain ratings ("mostly not", "somewhat", or "mostly") among intention and perception responses, and compare the proportions via chi-squared test. We take the chi-squared test statistic as the *relative uncertainty score* for that goal. A higher score means that the perception responses are more inclined towards uncertain ratings relative to the intention responses. Ranking the goals by uncertainty score, we find that opinion sharing ranks lowest (uncertainty score = 7.8), which reflects the relatively strong lean towards 5 ("definitely") among reported perceptions of that goal (Figure 2c). Fact seeking and opinion seeking are nearly tied for the highest relative uncertainty scores (26.9 and 27.8, Figures 2b and 2d, respectively). Such seeking-goals may be harder to perceive with high confidence because they are sometimes implicit [35].

3.2 Linguistic cues

Past work found that linguistic features such as part of speech, named entities, and hedging can distinguish between subjective [66, 68] and objective [38, 54] statements (corresponding to sharing opinions and facts), and that lexicon-based features can distinguish information seeking questions (which roughly correspond to fact seeking) from other types of questions such as social coordination [28, 30, 45]. But because these results relied exclusively on third-party labels, they only reflect perceptions. Here, we explore whether these linguistic features are also indicative of intentions.

Selecting linguistic features. We began with a basic set of linguistic features [66]: the usage of pronouns, adjectives, cardinal numbers, modals, and adverbs. We then refined the pronoun feature by distinguishing the use of first-person and second-person

pronouns [30, 38]. Mentions of named entities are characteristic of objective statements while hedging language (e.g., "I believe...") tends to signal subjectivity [54], so we incorporated these as additional features, alongside the explicit use of factual language (e.g., "In fact...") which can be regarded as the opposite of hedging [10]. Finally, we added features associated with information seeking questions: the use of please [28] and a question lexicon based on prior work [45]. For simplicity, all features were treated as binary (a comment either exhibits at least one instance of the linguistic feature or it does not). All features were extracted from the opening comment of the conversation as that was the comment the surveys asked about.

Comparing linguistic features. To compare how linguistic features are tied to intentions versus perceptions, for each pair of linguistic feature and goal (binarized, as described in Section 2), we separately regressed the intended goal on the feature, as well as the perceived goal on the feature, controlling for age, gender, Page size, and Page category. Regression coefficients are shown in Figure 3, where all variables were standardized for ease of comparison.

Several linguistic features correlate similarly with intentions and perceptions. For example, a one-standard-deviation increase in hedging corresponds to an increase in opinion sharing intent by 0.09 standard deviations and to an increase in opinion sharing intent by 0.15 standard deviations ($p < 0.001$). Similarly, adjectives signal both intended and perceived fact sharing (regression coefficients 0.15 and 0.14, respectively, $p < 0.001$).

Furthermore, the perception correlations generally corroborate prior results on subjectivity and information seeking. Consistent with findings in subjectivity detection, mentions of named entities are more correlated with fact sharing (0.18, $p < 0.001$) than with opinion sharing (0.06, $p < 0.05$), and use of cardinal numbers (intuitively, a heuristic capturing mentions of specific values) is correlated with fact sharing (0.17, $p < 0.001$) and not opinion sharing. Conversely, hedging is correlated with opinion sharing (0.15, $p < 0.001$) but not fact sharing. Consistent with prior work on information seeking questions, the use of please is associated with fact seeking (0.19, $p < 0.001$) and not opinion seeking; a similar but weaker effect holds for second person pronouns (0.05, $p < 0.05$).

Although many of the observed correlations are largely similar between intentions and perceptions, there are also some notable differences. For instance, the use of factual language is significantly correlated with intended fact sharing (0.06, $p < 0.01$) but not with perceived fact sharing. This may relate to the previously observed bias towards perceiving statements as opinions: even if the initiator tries to "double down" on the intended factuality of their comment through the explicit use of factual language, this might not have any effect on the replier, who is inclined towards perceiving opinions. Other examples include question words being negatively correlated with perceived fact sharing (-0.05, $p < 0.05$) but uncorrelated with intended fact sharing, and second person pronouns being correlated with perceived (0.05, $p < 0.05$) but not with intended fact seeking. Future work could examine in greater detail why these differences occur, and also consider more sophisticated language features.

Predicting conversational intentions and perceptions. Linguistic features, such as the ones described in the preceding analysis, have previously been successfully used in models for predicting subjectivity from text [38, 64, 68]. However, these models' reliance on third-party labels means that they must be understood as predicting *perceived* subjectivity. We now leverage our unique access to ground truth intentions to address the following question: Can intentions also be predicted from text, and if so, how different are intention prediction models from perception prediction models?

	Perception prediction		*Intention prediction*		*Label swap*	
Goal	**IC**	**+R**	**IC**	**+R**	**IC**	**+R**
Fact sharing	.64	.64	.68	.68	.65	.64
Fact seeking	.81	.83	.82	.83	.80	.80
Opinion sharing	.72	.74	.75	.78	.75	.77
Opinion seeking	.64	.65	.72	.71	.68	.68

Table 2: BERT-based classifiers using either the text of the initiator's comment only (IC) or the text of both the initiator's comment and the reply (+R) achieve reasonable performance in predicting both perceptions (left) and intentions (middle). Furthermore, using perception labels to predict intentions ("label swap") results in performance drops compared to using intention labels (compare middle and rightmost columns). All results are reported as area under the ROC curve (AUC) to account for class imbalance.

To evaluate the feasibility of intention prediction, we fine-tuned a BERT-based classifier [19] on the task of predicting the initiator's intention based on the text of their initial comment. Since prior work has found that incorporating context can improve predictive performance in conversational settings [21, 23], for completeness we additionally considered a version of this model that also looks at the text of the reply. Both models were trained on about 5,000 samples from the intention survey data, using binarized intention responses as the labels. They were then evaluated on 1,000 held out test samples from the same survey. Finally, as a point of comparison, we also trained models for the more traditional task of perception prediction by using the same setup on the perception survey data.

Table 2 (leftmost two columns) compares the performance of the intention and perception classifiers, measured in terms of area under the ROC curve (AUC) to account for class imbalance. We find that both intentions and perceptions can be predicted with similar performance, thus establishing the feasibility of the intention prediction task.

Beyond demonstrating feasibility, we also want to understand if predicting intention differs from predicting perception. In other words, are intention labels provided by the authors themselves interchangeable with "third-party" labels, not unlike those used in prior work on subjectivity detection? To test this, we applied the model trained on perception labels to the intention-labeled test set.

We find that using perception labels to predict intentions (Table 2, rightmost two columns) results in reduced performance for all four goals compared to using intention labels, with a 3.4% average decrease in AUC. One possible explanation for the difference is that models trained on perception labels may be learning the (distributional and linguistic) perception biases described earlier. If so, the use of such models should account for this limitation, especially in production settings.

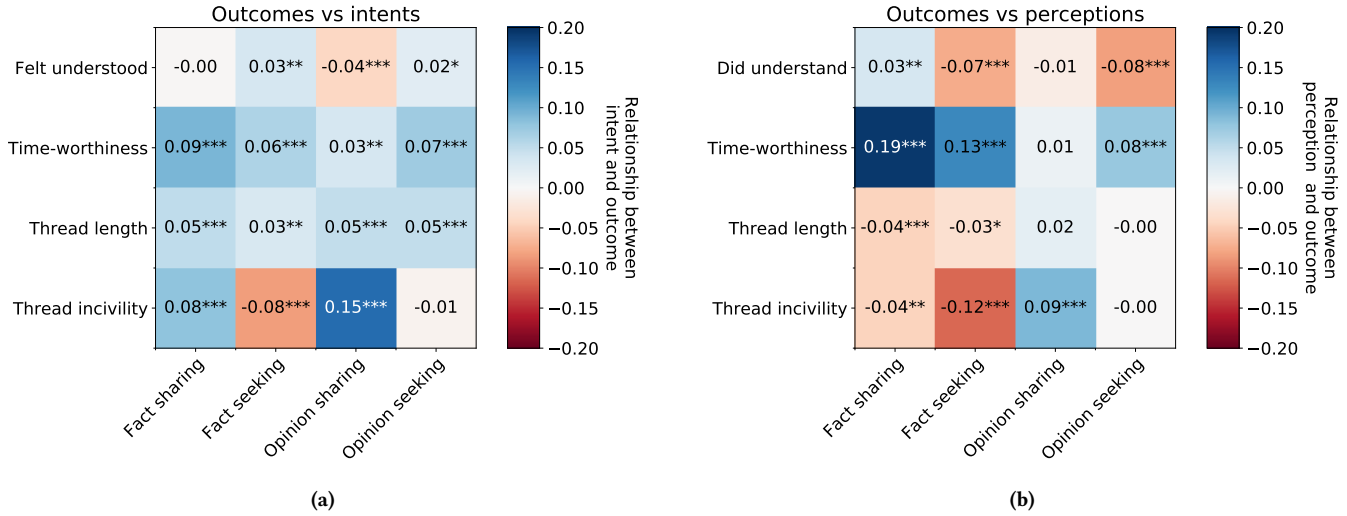

Figure 4: Conversational outcomes can relate differently to intentions (a) and perceptions (b). For example, while an intention to share a fact is positively correlated with greater future incivility in the conversation (regression coefficient 0.08), a perception that a comment is sharing a fact is instead negatively correlated (-0.04).

3.3 Relationship to conversational outcomes

Intentions and perceptions can also differ in their relationship to the outcomes and trajectories of the conversations in which they occur. We considered several conversational properties, three of which rely on the survey responses: whether the initiator or replier felt the discussion was worth their time, whether the initiator felt understood, or whether the replier felt they understood the initiator. We also considered two conversational trajectories proposed in prior work: the eventual length of the thread [2, 5, 37] and whether the discussion eventually turns uncivil [44].

To formalize the latter outcome at scale, we defined *thread incivility* as the maximum incivility score of all comments in the conversation following the initial comment-reply pair, where incivility score is computed by a production DeepText DocNN classifier [15, 36, 71] trained on manually-labeled content that violates Facebook's Community Standards on Hate Speech.[4] We verified the reliability of these scores by manually annotating a random sample of 200 comments using prior guidelines [69]. Substantial agreement between the manual and automated labels (Cohen's $\kappa = 0.73$) suggests that this automated score is a reasonable measure of incivility.

Intentions and outcomes. As before, we compared each pair of goal and outcome using a controlled regression analysis, regressing outcome on goal. Results are shown in Figure 4a.

Opinion sharing. An intention to share an opinion is correlated with higher likelihood of future incivility (regression coefficient 0.15, $p < 0.001$). This corroborates past work suggesting that opinion sharing is correlated with flaming [48]. The intention to share opinions is also correlated with stronger feelings of being misunderstood (-0.04, $p < 0.001$) and with longer threads (0.05, $p < 0.001$); the latter may be the result of opinion sharing triggering extended arguments or debates. However, initiators also tend to rate conversations started with opinion sharing intent as being worth their time (0.03, $p < 0.01$). One possible explanation is that initiators perceive the act of sharing their opinion as inherently valuable, regardless of downstream interactional outcomes. Together, these observations suggest a potential reason for why incivility continues to be prevalent on many online discussion platforms: people feel that sharing their opinion is worth their time despite the increased likelihood that doing so will lead to undesirable outcomes. These observations motivate further work on better understanding why conversation participants rate interactions as worth their time and on the design of platforms that can offer an outlet for expressing personal opinions while also encouraging healthy conversations around those opinions.

Seeking versus sharing facts. Among all the goals, fact seeking appears to be the most unambiguously positive, being significantly correlated with lower thread incivility (-0.06, $p < 0.001$), feeling understood (0.03, $p < 0.01$), and considering the conversation to be worth the time (0.06, $p < 0.001$). This could suggest that in many of these cases the initiator ends up getting the information they sought, leading them to view the interaction positively. On the other hand, fact sharing is slightly associated with negative outcomes: like opinion sharing, it is positively correlated with thread incivility (0.06, $p < 0.001$), although, unlike opinion sharing, it is not related to feeling understood.

Perceptions and outcomes. Several of the correlations we observed for intentions also hold for perceptions (Figure 4b). Notably, perceived opinion sharing remains correlated with higher thread incivility (0.09, $p < 0.001$) while perceived fact seeking remains correlated with lower thread incivility (-0.12, $p < 0.001$).

However, perceptions and intentions relate differently to outcomes in some key ways. In particular, for fact sharing, the direction of the correlation flips for both thread length and thread incivility. When an initial comment is *intended* to share a fact, the resulting conversation is more likely to turn uncivil (0.08, $p < 0.001$) and

[4] https://www.facebook.com/communitystandards/hate_speech

tends to run longer (0.05, $p < 0.001$). When an initial comment is *perceived* to be sharing a fact, the resulting conversation is less likely to turn uncivil (-0.05, $p < 0.01$) and tends to run shorter (-0.04, $p < 0.001$). Since this contrast might provide new insights into why some online public discussions turn uncivil, we examine it in more detail in the following section.

4 MISPERCEPTION OF FACT SHARING: A CASE STUDY

So far, we revealed systematic differences between intentions and perceptions at an aggregate level. In this section, we investigate the effect of *misperception* at the discussion level, or what happens in a conversation in which a replier perceives an initiator's comment differently from how it was intended. Particularly, we examine the relationship between misperception and undesirable conversational outcomes such as future incivility in a discussion.

As discussed above (Section 3.3) intended fact sharing in the initial comment correlates with greater incivility later in the discussion, but perceived fact-sharing instead correlates with less incivility later on—could misperception explain this contrast?

As facts are often misperceived as opinions [47, 53], we suspect perceptions of opinion-sharing may play a role. This, combined with the additional observation that perceived opinion sharing correlates with greater incivility, leads us to hypothesize: *the observed positive correlation between intended fact sharing and thread incivility can be attributed to comments that, while intended to share a fact, get misperceived as sharing an opinion.*

Testing this hypothesis requires labels for both intention and perception on the same conversations, but obtaining such paired ground truth at scale is infeasible (see Section 2). To circumvent this limitation, we explored two alternative approaches for obtaining paired labels. Our first approach supplements ground truth intention labels with automatically inferred perception labels, exploiting the relatively good performance of our perception classifiers (Section 3.2). Still, these classifiers may not (mis)perceive comments the same way that humans do, so any findings may simply reflect classifier error rather than human misperception. As such, our second approach combines the ground truth intention labels with third-party human-annotated perception labels, albeit only on a random subset of data due to platform limitations. Each approach has its drawbacks, but both lead to the same qualitative conclusion that supports our hypothesis.

4.1 Automatically inferred perceptions

Labeling procedure. We started by finding all comments whose ground truth intentions were, according to the initiator survey responses, to share a fact and *not* an opinion. Cases of mixed intention were excluded as they make misperception ambiguous: if a comment intended as sharing both fact and opinion is perceived as only sharing opinion, is it correctly perceived (as the perceiver correctly inferred the opinion sharing intent) or misperceived (as the perceiver failed to infer the fact sharing intent)?

We then ran the perception classifier (Section 3.2) on these comments.[5] For each comment, the classifier returns a confidence score between 0 and 1 for each goal, representing the estimated likelihood that the initiator's comment in the conversation was perceived as having that goal. The result is a dataset of comments where each comment has a fixed ground truth intention (fact sharing and not opinion sharing) and two classifier-generated perception scores, one for perceived opinion sharing and one for perceived fact sharing (the latter is included as a control). The classifier can be thought of as an imperfect proxy for a human perceiver.

	Future incivility		
Predictor	β	*SE*	
(Intercept)	0.37	0.63	
Perceived fact sharing (predicted)	-0.12	0.05	*
Perceived opinion sharing (predicted)	**0.30**	**0.05**	***
Perceived fact sharing × opinion sharing	0.04	0.05	
Initial comment incivility	0.07	0.04	ˆ
Reply incivility	0.19	0.05	***
Initiator is female	-0.31	0.09	***
Replier is female	-0.01	0.08	
Initiator age	-0.03	0.04	
Replier age	0.10	0.04	*
Page size (logged)	0.03	0.04	
Page category (not shown for space)			

Table 3: A regression analysis reveals the relationship between misperception and incivility in discussions where the initial comment was intended to share a fact ($R^2 = 0.23$). (* $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ˆ $p < 0.10$).**

Method. To verify our hypothesis, we regressed future incivility on perceived opinion sharing while accounting for several possible confounds. First, we controlled for perceived fact sharing (and included an interaction effect with perceived opinion sharing) to test the alternate hypothesis that perceived fact sharing alone fully explains differences in incivility. We also controlled for the incivility of the initial comment and that of the reply, as prior work found that incivility in the opening exchange of a conversation is a relatively strong indicator of future incivility [69]. Finally, we also included the gender and age of both the initiator and the replier, Page size, and Page category. All continuous variables were standardized.

If our hypothesis holds, perceived opinion sharing would be positively associated with future incivility.

Results. Consistent with our hypothesis, we find a significant positive effect of perceived opinion sharing (β= 0.30, $p < 0.001$): a one-standard-deviation increase in perceived opinion sharing results in a 0.30 standard-deviation increase in future incivility (Table 3). This effect dominates the effect of other variables in the regression except that of the initiator's gender, which is about equal in magnitude.[6] Consistent with our previous findings, we also find a significant but weaker negative effect of perceived fact sharing

[5] We use the version that uses text from both the initial comment and the reply as it performs best.

[6] Analysis of the relationship between gender and incivility lies outside the scope of the present work; see [17] for some additional discussion.

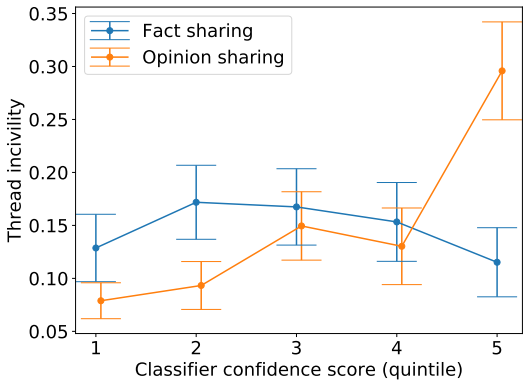

Figure 5: Visualization of the relationship between thread incivility and perception scores for fact sharing (blue) and opinion sharing (orange), conditioned on the ground-truth intent being fact sharing. Each point indicates mean thread incivility among all comments in the specified quintile of classifier scores; error bars indicate 95% CIs.

(β= -0.12, $p < 0.05$), indicating that correctly perceiving fact sharing translates into more civil conversations.

These relationships can also be visualized (in an uncontrolled setting) by binning conversations by quintiles of the perception classifier scores for fact sharing and opinion sharing, then plotting mean incivility per bin (Figure 5). Again, we observe a positive relationship between perceived opinion sharing and thread incivility and a weak negative relationship between perceived fact sharing and thread incivility.

4.2 Human annotated perceptions

Automatically inferred perceptions have the advantage of scalability but the findings made using those labels may only apply to cases where a classifier misperceives fact sharing as opinion sharing, and may not necessarily generalize to cases where a person does the same. As such, we also obtained third-party human annotations of perception to validate the previous results.

Labeling procedure. We sent a random sample of conversations from the initiator survey responses to expert annotators (though platform limitations restricted the scale of annotation). To avoid potentially biasing the annotators, comments were sampled such that half were intended as fact sharing and half were intended as opinion sharing. As before, the two intentions were held to be mutually exclusive. For each conversation, annotators were shown the initial comment and asked to rate whether they thought it was sharing a fact and, separately, whether they thought it was expressing an opinion (such that they had the option of marking both or neither). The terms "opinion" and "fact" were left purposely vague and annotators were encouraged to exercise personal judgment, to more accurately simulate how perceptions get formed upon seeing a comment in the wild. In total, we received annotations for 330 comments from 6 annotators.

Method. Due to the smaller data size and coarser-grained labels, we ran a simplified analysis, comparing the future incivility of discussions in which the initial comment was labeled as expressing an opinion to discussions in which the initial comment was labeled as sharing a fact. If our hypothesis holds, we expect future incivility to be higher in the former case.

Results. Supporting this hypothesis, we find that the median thread incivility among cases labeled as expressing opinions is 0.09, compared to 0.07 among cases labeled as sharing facts; this difference is significant via Mann-Whitney test ($U = 1655.5$, $p < 0.05$). This finding also provides additional evidence that the differences we have observed in this section are actually reflective of differences in perception, as opposed to merely reflecting biases of the classifier.

Together, the results from both perception label sources (automated and human) provide support for our hypothesis, suggesting that among conversations intended as fact sharing, the correlation with thread incivility arises largely from cases that were (mis)perceived as opinion sharing. Future work could build upon this result by investigating what factors (linguistic or otherwise) lead to this kind of misperception, and whether similar effects occur for other combinations of goals and outcomes. This line of research could lead to a better understanding of the mechanisms through which incivility arises in well-intentioned online discussions.

5 FURTHER RELATED WORK

Intentions in other domains. While our work focused on conversational intentions, intentionality is also a dimension of other online settings. For instance, prior work studied intent to communicate (offline) commitments in emails [12, 41, 62], search query intent for search result customization [4, 7, 29, 31], and purchasing intent on online shopping platforms [9, 42, 61]. These settings differ from our setting of online discussions, but there is some overlap in goals: fact seeking intent also applies to email [62] and search [8].

Some forms of intention can be measured without the need for surveys. For instance, some work has studied intended humor [55] and sarcasm [6, 25] by treating user-supplied hashtags as natural labels. In the context of news articles, document tags were treated as natural labels of the intended purpose of the tagged article [68]. Linguistic features such as part-of-speech ended up being predictive of both these natural intent labels [6] and our survey-based labels.

Factors influencing perception. Like intention, perception is also relevant in many online settings, and prior work has explored factors influencing how online comments, documents, and actions get perceived. In particular, prior work on the perceived objectivity of online news [57] and the perceived trustworthiness of product reviews [20] or dating profiles [34] closely relate to our work on perception of facts and opinions, while studies of how perceived fairness of community moderation affects future incivility or community loyalty [13, 33] echo our results relating perceptions to conversational outcomes. We add to the existing literature on perception by using our survey-based methodology to relate perceptions to intentions. Most similar is previous work that surveyed 95 conversation initiators and 41 repliers to measure intentions and perceptions in relation to incivility [48]. The present work, in contrast, involved a much larger-scale survey and considered other outcomes such as thread length.

Subjectivity detection. Distinguishing between opinions and facts is closely related to the task of subjectivity detection, for which a number of language-based models have been proposed [40, 50, 64, 66]; see [43] for a more complete survey. However, the two tasks are not identical as subjectivity encompasses more than opinions: [66] defines subjective language as expressing private state, which includes not only opinions but also emotions and speculation [52]. Some work on subjectivity focused explicitly on opinions in the context of news media and wiki articles [54, 63], but largely relied on third-party annotations [65] and hence mainly captured perceptions of opinions. We add to this work by examining opinions and facts in a conversational context (which in turn introduces a distinction between sharing and seeking), considering intentions to share opinions (or facts) in addition to perceptions, and comparing the use of linguistic cues borrowed from the subjectivity detection literature in predicting intentions and perceptions.

Subjectivity detection has also been shown to be helpful in downstream tasks such as information extraction [56], sentiment analysis [67], and document quality measurement [39]. Our work similarly shows that both intention and perception of opinions and facts can be indicative of conversational outcomes such as future incivility.

Forecasting conversational outcomes. This work has shown that intentions and perceptions relate to future conversational outcomes. Prior work has studied other signals of outcomes, such as pragmatic cues [69], similarity between comments [1], and conversation structure [22]. These have been used to forecast outcomes such as success in negotiation [11, 18] and eventual disagreement [32], as well as two of the outcomes examined in our work: thread length [5] and incivility [44, 69]. In particular, [69] indirectly explores the connection between intentions and future incivility by using an unsupervised method [70] to estimate the intended role of a comment; we build on this by obtaining ground truth intents via survey and additionally relating them to perceptions.

6 CONCLUSION

In this work, we presented a large-scale study of how intentions and perceptions can diverge in online public discussions. Using a survey of over 16,000 people, we obtained unprecedented access to ground truth labels for the intentions underlying comments on online public discussions, as well as how such comments were perceived. Using this data, we revealed both distributional and linguistic differences between intentions and perceptions, showed that such differences are reflected in the performance of automated classifiers, and explored how misperceptions can be tied to the future trajectory of a discussion. In particular, when a comment intended to share a fact is misperceived as sharing an opinion, the subsequent conversation is more likely to turn uncivil than when that intention is correctly perceived.

These results point towards several design opportunities for promoting healthier interactions on online discussion platforms. For instance, classifiers that predict intentions and perceptions could signal to users when a comment they are writing may be misperceived by others and suggest concrete strategies for reducing this risk. Nonetheless, user studies would be needed to guide the design of such interventions to reduce the likelihood of unintended negative consequences. Our results further suggest that reducing misperception may improve civility in online discussions, but additional work is needed to better understand this connection and to what extent, if at all, misperception is *predictive* of incivility. We also note that these findings specifically apply to public discussions, and the effect of misperception in other kinds of settings (e.g., private discussions or article-style monologic text) remains a related but separate question.

Limitations of our survey methodology provide opportunities for future work. Low survey response rates prevented the collection of paired survey responses from an initiator and replier on the same conversation, limiting a direct study of misperception. The classifier predictions we used as a substitute for perception labels were generally reliable, and our results were verified via third-party annotation, but it would nonetheless be valuable to replicate these results on paired responses. The retrospective nature of the surveys also adds ambiguity with respect to interpretations of reported perceptions: were responders reporting how they perceived the comment at the time of the conversation, or were they reporting how they perceived it in hindsight at the time of the survey? Though the results varied little with the amount of time between the time a person commented and the time they took the survey, surveying people at the time of conversation may constitute interesting future work. Finally, while we have accounted for demographic and Page features as potential confounds, other confounds may exist.

Our analysis may also be applied to other conversational goals beyond facts and opinions. For instance, community moderators may want to deal differently with a person who intentionally trolled others in a conversation than one who unintentionally did so. Intentions and perceptions may also relate to community-level rather than conversation-level outcomes. For instance, does the way in which a community member tends to perceive others in the community relate to that member's long-term loyalty? Such effects might also end up being specific to certain *kinds* of communities; while our present work looks only at Facebook Pages, future work could apply our methodology to other platforms with different modes and norms of interaction, like Twitter and Reddit. Finally, while our present work has looked at intentions and perceptions at the start of a conversation, goals may change as the conversation progresses. These all constitute promising paths for future exploration building upon the methods and findings presented in this work.[7]

ACKNOWLEDGMENTS

This research was supported in part by an NSF CAREER award IIS-1750615 and by an NSF Grant IIS-1910147. We would like to thank Lada Adamic, Israel Nir, Alex Dow, Ashish Gupta, Karen Jusko, Alex Leavitt, Moira Burke, Caleb Chiam, Liye Fu, Justine Zhang, as well as our reviewers for their valuable feedback (which we hopefully perceived correctly).

REFERENCES

[1] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-Scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* (Dec. 2016).

[7] All analysis code is available on Github at https://github.com/facebookresearch/intentions-perceptions.

[2] Pablo Aragón, Vicenç Gómez, David García, and Andreas Kaltenbrunner. 2017. Generative Models of Online Discussion Threads: State of the Art and Research Challenges. *Journal of Internet Services and Applications* 8, 15 (Dec. 2017).

[3] Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting Responses to Microblog Posts. In *Proceedings of NAACL*.

[4] Azin Ashkan, Charles L. A. Clarke, Eugene Agichtein, and Qi Guo. 2009. Classifying and Characterizing Query Intent. In *Advances in Information Retrieval*, Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy (Eds.). Springer Berlin Heidelberg.

[5] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2013. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-Entry. In *Proceedings of WSDM*.

[6] David Bamman and Noah A. Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of ICWSM*.

[7] David J. Brenes, Daniel Gayo-Avello, and Kilian Pérez-González. 2009. Survey and Evaluation of Query Intent Detection Methods. In *Proceedings of the 2009 Workshop on Web Search Click Data*.

[8] Andrei Broder. 2002. A Taxonomy of Web Search. *ACM SIGIR Forum* 36, 2 (Jan. 2002).

[9] Mark Brown, Nigel Pope, and Kevin Voges. 2003. Buying or Browsing? An Exploration of Shopping Orientations and Online Purchase Intention. *European Journal of Marketing* 37, 11/12 (Dec. 2003).

[10] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

[11] 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 EMNLP*.

[12] Vitor R. Carvalho. 2011. Modeling Intention in Email - Speech Acts, Information Leaks and Recommendation Models. In *Studies in Computational Intelligence*.

[13] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of WWW*.

[14] Herbert H Clark. 1996. *Using Language* (second ed.). Cambridge University Press.

[15] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 12, Aug (2011).

[16] Victor Corral-Verdugo. 1993. The Effect of Examples and Gender on Third Graders' Ability to Distinguish Environmental Facts from Opinions. *The Journal of Environmental Education* 24, 4 (July 1993).

[17] Naomi Craker and Evita March. 2016. The Dark Side of Facebook®: The Dark Tetrad, Negative Social Potency, and Trolling Behaviours. *Personality and Individual Differences* 102 (Nov. 2016).

[18] Jared R. Curhan and Alex Pentland. 2007. Thin Slices of Negotiation: Predicting Outcomes From Conversational Dynamics Within the First 5 Minutes. *Journal of Applied Psychology* 92 (May 2007).

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

[20] Raffaele Filieri. 2016. What Makes an Online Consumer Review Trustworthy? *Annals of Tourism Research* 58 (May 2016).

[21] Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *Proceedings of RANLP*.

[22] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Quantifying Controversy in Social Media. *ACM Transactions on Social Computing* 1, 1 (2017).

[23] Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The Role of Conversation Context for Sarcasm Detection in Online Interactions. In *Proceedings of SIGDIAL*.

[24] Leonard M. Giambra, Cameron J. Camp, and Alicia Grodsky. 1992. Curiosity and Stimulation Seeking across the Adult Life Span: Cross-Sectional and 6- to 8-Year Longitudinal Findings. *Psychology and Aging* 7, 1 (1992).

[25] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying Sarcasm in Twitter: A Closer Look. In *Proceedings of ACL*.

[26] Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12, 3 (July 1986).

[27] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook. *Science Advances* 5, 1 (Jan. 2019).

[28] Ido Guy, Victor Makarenkov, Niva Hazon, Lior Rokach, and Bracha Shapira. 2018. Identifying Informational vs. Conversational Questions on Community Question Answering Archives. In *Proceedings of WSDM*.

[29] Alan Hanjalic, Christoph Kofler, and Martha Larson. 2012. Intent and Its Discontents: The User at the Wheel of the Online Video Search Engine. In *Proceedings of MM*.

[30] F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. 2009. Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proceedings of CHI*.

[31] Homa B Hashemi, Amir Asiaee, and Reiner Kraft. 2016. Query Intent Detection Using Convolutional Neural Networks. In *Proceedings of QRUMS*.

[32] Jack Hessel and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-Causing Posts from Discussion Features. In *Proceedings of NAACL*.

[33] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. In *Proceedings of CSCW*, Vol. 1.

[34] Seunga Venus Jin and Cassie Martin. 2015. "A Match Made…Online?" The Effects of User-Generated Online Dater Profile Types (Free-Spirited Versus Uptight) on Other Users' Perception of Trustworthiness, Interpersonal Attraction, and Personality. *Cyberpsychology, Behavior, and Social Networking* 18, 6 (June 2015).

[35] Andrew Kehler and Hannah Rohde. 2017. Evaluating an Expectation-Driven Question-Under-Discussion Model of Discourse Interpretation. *Discourse Processes* 54, 3 (April 2017).

[36] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of EMNLP*.

[37] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. 2010. Dynamics of Conversations. In *Proceedings of KDD*.

[38] Elisabeth Lex, Andreas Juffinger, and Michael Granitzer. 2010. Objectivity Classification in Online Media. In *Proceedings of HT*.

[39] Elisabeth Lex, Michael Voelske, Marcelo Errecalde, Edgardo Ferretti, Leticia Cagnina, Christopher Horn, Benno Stein, and Michael Granitzer. 2012. Measuring the Quality of Web Content Using Factual Information. In *Proceedings of WebQuality*.

[40] Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence Subjectivity Detection with Weakly-Supervised Learning. In *Proceedings of IJCNLP*.

[41] Chu-Cheng Lin, Dongyeop Kang, Michael Gamon, and Patrick Pantel. 2018. Actionable Email Intent Modeling With Reparametrized RNNs. In *Proceedings of AAAI*.

[42] Kwek Choon Ling, Lau Teck Chai, and Tan Hoi Piew. 2010. The Effects of Shopping Orientations, Online Trust and Prior Online Purchase Experience toward Customers' Online Purchase Intention. *International Business Research* 3, 3 (June 2010).

[43] Bing Liu. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing* (2nd ed.).

[44] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of ICWSM*.

[45] Zhe Liu and Bernard J. Jansen. 2015. A Taxonomy for Classifying Questions Asked in Social Question and Answering. In *Proceedings of CHI Extended Abstracts*.

[46] Heidi McKee. 2002. "YOUR VIEWS SHOWED TRUE IGNORANCE!!!": (Mis)Communication in an Online Interracial Discussion Forum. *Computers and Composition* 19, 4 (Dec. 2002).

[47] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. 2018. Distinguishing Between Factual and Opinion Statements in the News. https://www.journalism.org/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/.

[48] Peter J. Moor, Ard Heuvelman, and Ria Verleur. 2010. Flaming on YouTube. *Computers in Human Behavior* 26, 6 (Nov. 2010).

[49] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&a Behavior. In *Proceedings of CHI*.

[50] Gabriel Murray and Giuseppe Carenini. 2011. Subjectivity Detection in Spoken and Written Conversations. *Natural Language Engineering* 17, 3 (July 2011).

[51] Silvia Quarteroni, Alexei V. Ivanov, and Giuseppe Riccardi. 2011. Simultaneous Dialog Act Segmentation and Classification from Human-Human Spoken Conversations. In *Proceedings of ICASSP*.

[52] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, New York.

[53] Mitchell Rabinowitz, Maria Acevedo, Sara Casen, Myriah Rosengarten, Martha Kowalczyk, and Lindsay Blau Portnoy. 2013. Distinguishing Facts from Beliefs: Fuzzy Categories. *Psychology of Language and Communication* 17, 3 (Dec. 2013).

[54] Santosh Regmi and Bal Krishna Bal. 2015. What Make Facts Stand Out from Opinions? Distinguishing Facts from Opinions in News Media. *Creativity in Intelligent Technologies and Data Science* 535 (2015).

[55] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering* 74 (April 2012).

[56] Ellen Riloff, Janyce Wiebe, and William Phillips. 2005. Exploiting Subjectivity Classification to Improve Information Extraction. In *Proceedings of AAAI*.

[57] S. Shyam Sundar. 1998. Effect of Source Attribution on Perception of Online News Stories. *Journalism & Mass Communication Quarterly* 75, 1 (March 1998).

[58] Deborah Tannen. 2000. Indirectness at Work. In *Language in Action: New Studies of Language in Society, Festschrift for Roger Shuy*.

[59] Deborah Tannen. 2005. *Conversational Style : Analyzing Talk among Friends*. Oxford University Press, New York.

[60] Richard Valliant. 1993. Poststratification and Conditional Variance Estimation. *J. Amer. Statist. Assoc.* 88, 421 (March 1993).

[61] Hans van der Heijden, Tibert Verhagen, and Marcel Creemers. 2001. Predicting Online Purchase Behavior: Replications and Tests of Competing Models. In *Proceedings of HICSS*.
[62] Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N. Bennett, and Chris Quirk. 2019. Context-Aware Intent Identification in Email Conversations. In *Proceedings of SIGIR*.
[63] Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. 2003. Recognizing and Organizing Opinions Expressed in the World Press. In *AAAI Symposium on New Directions in Question Answering*.
[64] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning Subjective Language. *Computational Linguistics* 30, 3 (Sept. 2004).
[65] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39, 2 (May 2005).
[66] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In *Proceedings of ACL*.
[67] Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-Level Structured Models for Document-Level Sentiment Classification. In *Proceedings of EMNLP*. Cambridge, Massachusetts.
[68] Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of EMNLP*.
[69] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.
[70] Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking Too Much? The Rhetorical Role of Questions in Political Discourse. In *Proceedings of EMNLP*.
[71] Xiang Zhang and Yann LeCun. 2015. *Text Understanding from Scratch*. Technical Report.