# Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation

# Should a Chatbot be Sarcastic?
# Understanding User Preferences Towards Sarcasm Generation

**Silviu Vlad Oprea**[1] and **Steven R. Wilson**[3] and **Walid Magdy**[1, 2]
[1] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[2] The Alan Turing Institute, London, United Kingdom
[3] School of Engineering and Computer Science, Oakland University, Rochester, MI, USA
silviu.oprea@ed.ac.uk, stevenwilson@oakland.edu,
wmagdy@inf.ed.ac.uk

## Abstract

Previous sarcasm generation research has focused on *how* to generate text that people perceive as sarcastic to create more human-like interactions. In this paper, we argue that we should first turn our attention to the question of *when* sarcasm should be generated, finding that humans consider sarcastic responses inappropriate to many input utterances. Next, we use a theory-driven framework for generating sarcastic responses, which allows us to control the linguistic devices included during generation. For each device, we investigate how much humans associate it with sarcasm, finding that pragmatic insincerity and emotional markers are devices crucial for making sarcasm recognisable.

## 1 Introduction

The prevalence of sarcasm on the social web (Khodak et al., 2018; Sykora et al., 2020) has motivated computational investigations across the NLP community. Most focus on textual sarcasm detection, the task of classifying whether or not a given text is sarcastic (Riloff et al., 2013; Wallace et al., 2015; Rajadesingan et al., 2015; Bamman and Smith, 2015; Joshi et al., 2016; Amir et al., 2016; Hazarika et al., 2018; Oprea and Magdy, 2019; Abu Farha et al., 2022).

A recent research direction considers sarcasm generation. Approaches to sarcasm generation introduced so far (Joshi et al., 2015; Mishra et al., 2019; Chakrabarty et al., 2020) are mainly motivated by the potential to create more approachable, human-like conversational agents, considering that sarcasm is a natural part of human discourse. We suggest reconsidering this motivation, as a community, for two reasons.

First, in *human* discourse, sarcasm is not a communicative goal in itself. Rather, it can be used to achieve a wide variety of goals. Some of these goals, such as to diminish the impact of criticism (Dews and Winner, 1995), to create humour (Kreuz et al., 1991; Colston and O'Brien, 2000b,a), to praise (Bruntsch and Ruch, 2017), or to strengthen relationships (Jorgensen, 1996; Pexman and Zvaigzne, 2004), might be desirable in human-machine interactions as well. However, other goals, such as criticising, mocking, or expressing dissociation, often with surface contempt or derogation (Wilson, 2006), might not be desirable in human-machine interactions.

Second, the communicative goals mentioned above were observed in *human* interactions. Even when a machine seeks potentially desirable goals, it is unclear whether sarcastic utterances have the same effect on humans when coming from machines.

Therefore, we suggest it is imperative, not least from an ethical perspective, to consider the following **research questions** (RQs):

RQ1. When should a chatbot be sarcastic?
  (a) When do humans consider sarcasm appropriate?
  (b) When do humans prefer sarcasm, over non-sarcasm?

RQ2. How should a chatbot formulate sarcasm?
  (a) What linguistic devices do humans associate with sarcasm?
  (b) What sarcasm flavour do they prefer?

Here, by *flavour*, we mean a specific conjunction of linguistic devices that humans may associate with sarcasm, such as intensifiers and emotional markers, as introduced in Section 3, and expanded upon in Section 4.

To address our research questions, we suggest the following approach. First, given a set of input utterances, generate several sarcastic responses.

*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 7686 - 7700
May 22-27, 2022 ©2022 Association for Computational Linguistics

Each response should be of a specific sarcasm flavour, i.e. should display a specific conjunction of linguistic devices. Next, create a survey that asks human participants: to indicate how appropriate it was to respond sarcastically to the input; to select their preferred response; and to rate the sarcasticness of each response, investigating whether they associate the linguistic devices in the response with sarcasm.

To achieve this, we require a sarcastic response generator that provides control over the linguistic devices used. Most previous generators rely on variants of the traditional theory of sarcasm, which claims that the intended meaning concealed by sarcasm is the opposite of the literal meaning. However, this theory provides a grounding that is neither necessary, nor sufficient, for sarcasm to occur, as discussed in Section 3. To overcome this limitation, we recently introduced Chandler, a novel modular sarcastic response generation framework (Oprea et al., 2021). It is grounded on a formal theory that, from a linguistic-theoretical perspective, specifies devices whose presence is both necessary and sufficient to unambiguously differentiate sarcasm from non-sarcasm. These are allusion to a failed expectation, pragmatic insincerity, and emotional markers. Chandler can generate sarcasm of different flavours, and allows control over flavour its output should reflect. Herein, we also compare Chandler's outputs to those of previous generators, to examine participant preferences toward an even greater range of sarcasm flavours.

Our results indicate that people find sarcastic responses inappropriate for most input utterances. When sarcasm was considered appropriate, the inputs commonly had a positive sentiment, and often had elements of humour. Further, even when considered appropriate, people still did not usually *prefer* sarcastic responses over non-sarcastic ones. Sarcasm was typically preferred when it was also considered funny and not too specific. Finally, we identified pragmatic insincerity and emotional markers (cf. Section 3) as crucial linguistic devices to include in generating recognizable sarcasm.

**Contributions** We summarise our contributions as follows. First, our approach allows us to understand people's preferences about when sarcasm should be used, and how it should be formulated. Using this information, we provide guidelines for future work in sarcasm generation. Second, observing people's preferences also allows us to quantitatively evaluate the practical advantages of the formal linguistic theory that grounds Chandler.

## 2 Related Work

The earliest work on sarcasm generation is that of Joshi et al. (2015), who introduce SarcasmBot, a sarcastic response generation system. SarcasmBot uses one of eight possible generators, each containing a set of predefined patterns, one of which is instantiated as the response. The generators do not in fact account for the meaning of the input, rather, they only focus on aspects such as the overall sentiment or presence of swear words. Further, in our experiments, we noticed that most of the time a fallback generator was employed, returning the simple concatenation of a random positive phrase to a random negative one, from a set of predefined phrases that have no specific connection to the input.

Mishra et al. (2019) suggest a sarcastic paraphrase generator. They assume that the input is always of negative polarity, and suggest an unsupervised pipeline of four modules to convert such an input $u^{(-)}$ to a sarcastic version. In the Sentiment Neutralisation module, they filter out negative sentiment words from $u^{(-)}$ to produce $u^{(0)}$. In the Positive Sentiment Induction module, they modify $u^{(0)}$ to convey positive sentiment, producing $u^{(+)}$. Next, in the Negative Situation Retrieval module, they mine a phrase $v^{(-)}$ that expresses a negative situation. $v^{(-)}$ is selected from a set of predefined phrases, based on the similarity to the original input. Finally, the Sarcasm Synthesis module constructs the sarcastic paraphrase from $u^{(+)}$ and $v^{(-)}$.

Chakrabarty et al. (2020) suggest a similar pipeline. Their $R^3$ system first employs a Reversal of Valence module, which replaces input words of negative valence with their lexical antonyms using WordNet (Miller, 1995) to produce $u^{(+)}$. Next, it builds an utterance $v$ that is incongruous to $u^{(+)}$, and generates sarcasm from $u^{(+)}$ and $v$.

Previous generators share a limitation that make them unfit for our purposes. Mainly, relying on the traditional theory, they identify sarcasm with linguistic incongruity. Thus, they only provide this single device for investigation, device that is not sufficient for sarcasm to occur, as discussed in Section 3. A further limitation, shared by Mishra et al. (2019) and Chakrabarty et al. (2020), is that their generators only work with input utterances of negative sentiment. However, as discussed earlier, sarcastic communication can have many goals, including to praise, or to strengthen friendships.

## 3 Linguistic Grounding

In this section, we describe the Implicit Display Theory, a formal linguistic theory that grounds Chandler.

**Previous Theories** In the *traditional theories*, sarcasm is created by literally saying one thing but figuratively meaning, or conversationally implicating (Grice, 1975), the opposite. However, such incongruity is not necessary for sarcasm. To see this, consider sarcastic understatements such as saying "This was not the best movie ever" to mean the movie was bad. It is also not sufficient. For instance, it also occurs in the construction of certain stylistic devices, such as metaphors, e.g. "Time is money". Further theories have been suggested to address these limitations, including the *echoic mention theory* (Sperber and Wilson, 1981) and its variants (Kreuz and Glucksberg, 1989; Wilson and Sperber, 1992; Sperber and Wilson, 1998), and the *pretense theory* (Clark and Gerrig, 1984) and its variants (Clark, 1996). However they all fail to uniquely identify sarcasm, as argued by Utsumi (2000) and Oprea and Magdy (2020).

**Implicit Display Theory (IDT)** Introduced by Utsumi (1996), the IDT focuses specifically on making the distinction between sarcasm and non-sarcasm. We invite the interested reader to consult (Utsumi, 2000) for an overview of how it overcomes the limitations of previous theories. We chose it as a grounding for our generation system.

The IDT first defines the concept of an ironic environment. We say a situation in which an utterance occurs is surrounded by an ironic environment if the discourse context includes the following components: (1) The speaker has expectation $Q$ at time $t_0$; (2) $Q$ fails at time $t_1 > t_0$; and (3) The speaker has a negative attitude towards the failure of $Q$. Note that the idea of linking sarcasm to an expectation is not new to Utsumi (1996), rather it is supported by previous work (Kreuz and Glucksberg, 1989; Kumon-Nakamura et al., 1995).

Next, according to the IDT, an utterance is sarcastic if and only if it implicitly displays the ironic environment. Implicit display is realised if the following linguistic devices are present in the utterance: (1) allusion to the speaker's failed expectation $Q$; (2) pragmatic insincerity, realised by intentionally violating one of the pragmatic principles, e.g. Grice's maxims (Grice, 1975); and (3) implication (indirect expression) of the speaker's negative attitude towards the failure of $Q$. Finally, the theory claims that the degree of sarcasm of an utterance is proportional to how many of these linguistic devices are present in the utterance.

## 4 Methodology

In this section we look at the methodology employed to address our research questions. Specifically, we first select a set of input utterances. Next, for each input, we generate four sarcastic responses of different flavours using Chandler, and three more responses using other systems. Finally, for each input, in a survey, we ask human participants to rate the responses across several dimensions, to understand their preference towards the appropriateness of sarcasm, and which linguistic devices they associate with sarcasm.

### 4.1 Selecting Input Texts

As inputs, we select texts from the corpus published by Wilson and Mihalcea (2019). The corpus contains short texts (extracted from tweets) where users describe actions they performed. We compute the sentiment polarity of each text using the classifier from Barbieri et al. (2020), a RoBERTa model (Liu et al., 2019) fine-tuned on the tweet sentiment dataset from Rosenthal et al. (2017). Next, we form five partitions of 50 texts each: *very negative* and *very positive*, containing the top 50 texts based on their negative and positive probabilities, respectively; *negative*, containing random texts for which the probability of being negative was higher that the probabilities of being positive or neutral; and *positive* and *neutral*, partitions that we formed analogously to how we formed the *negative* partition. Our final input dataset contains 250 texts.

### 4.2 Generating Sarcastic Responses

For completeness, in this section we describe Chandler, the sarcastic response generator that we introduced in Oprea et al. (2021).

The IDT directly suggests an algorithm for sarcasm generation that identifies an ironic environment, then creates an utterance that implicitly displays it. We now discuss how we implement each step.

**Ironic Environment** As discussed in Section 4.1, each input text $U_{\text{in}}$ describes an action. In this scenario, herein, we assume the expectation

$Q$ that is part of the ironic environment negates that action. For instance, say $U_{\text{in}}$ expresses the event $P$ = [<user> wins the marathon]. We assume $Q = \neg P$ = [<user> does not win the marathon]. As we shall see, the algorithm we suggest will not, in fact, require us to formulate Q, but it relies on the above assumption.

**Allusion to $Q$** Following Utsumi (2000), we define allusion in terms of coherence relations, similar to the relations of rhetorical structure theory (RST) (Mann and Thompson, 1987). That is, if $U_\alpha$ is an utterance that expresses proposition $\alpha$, we say $U_\alpha$ alludes to the expectation $Q$ if and only if there is a chain of coherence relations from $\alpha$ to $Q$[1]. So, we need to first select a proposition $\alpha$ to either start or end the coherence chain, then specify the chain between $\alpha$ and $Q$, and formulate $U_\alpha$ such that it expresses $\alpha$. We suggest defining such $\alpha$ as objects of if-then relations, where the subject is $P$, the proposition expressed by input text $U_{\text{in}}$. That is, relations of the form "if $P$ then $\alpha$" should hold. To infer $\alpha$ given $U_{\text{in}}$, we use COMET (Bosselut et al., 2019), an adaptation framework for constructing commonsense knowledge. Specifically, we use the COMET variant fine-tuned on ATOMIC (Sap et al., 2019), a dataset of typed if-then relations. COMET inputs the subject of the relation, along with the relation type, and outputs the relation object. In our case, the subject is $U_{\text{in}}$, and we set $\alpha$ to the relation object.

In the examples that follow, assume the input text is $U_{\text{in}}$ ='<user> won the marathon'. We leverage four relation types: (1) **xNeed**: the object $\alpha$ of a relation of this type specifies an action that the user needed to perform before the event took place, e.g. "if $U_{\text{in}}$ then $\alpha$ = [*xNeed* to train hard]"; (2) **xAttr**: the object $\alpha$ specifies how a user that would perform such an action is seen, e.g. "if $P$ then $\alpha$ = [*xAttr* competitive]"; (3) **xReact**: the object $\alpha$ specifies how the user could feel as a result of the event, e.g. "if $P$ then $\alpha$ = [*xReact* happy]"; and (4) **xEffect**: the object specifies a possible effect that the action has on the user, e.g. "if $P$ then $\alpha$ = [*xEffect* gets congratulated]". In Table 1 we show, for each relation type, the coherence chains between the relation object $\alpha$ and the failed expectation $Q$. Under these conditions, to generate an utterance $U_\alpha$ that alludes to $Q$, we need to choose any $U_\alpha$ that expresses $\alpha$.

**Algorithm 1:** Generate sarcastic response

```
input: utterance $U_{\text{in}}$;
ironic environment
    Let $Q := \neg P$ be the failed expectation;
implicit display
    Choose an if-then relation type $\tau$ from xNeed,
      xAttr, xReact, and xEffect;
    Let $\alpha = \text{COMET}(U_{\text{in}}, \tau)$;
return response $U_{out}$ that expresses emotion$(\neg\alpha)$;
```

**Pragmatic insincerity** The second requirement for implicit display is that the utterance generated should include pragmatic insincerity. In this paper, we focus on violating Grice's maxim of quality (Grice, 1975), where we aim for the propositional content of the generated utterance to be incongruous to that of $U_{\text{in}}$ (input text). To achieve this, we first choose an if-then relation type, then infer the relation object $\alpha$ from $U_{\text{in}}$ using COMET, and construct an utterance that expresses $\neg\alpha$. For instance, if $U_{\text{in}}$ ='<user> won the marathon', and we have chosen the *xAttr* relation type, the constructed utterance could express $\neg\alpha$ = [<user> is not competitive].

**Negative attitude** To fulfill the last requirement of implicit display, the utterance generated should imply a negative attitude towards the failure of the expectation $Q$. As pointed out by Utsumi (1996), this can be achieved by embedding verbal cues usually associated with such attitudes, including hyperbole and interjections.

**Logical form and explainability** At this point we formulate Algorithm 1 for generating a sarcastic response $U_{out}$, given an input utterance $U_{\text{in}}$ that expresses proposition $P$. We refer to *emotion*$(\neg\alpha)$ as the *logical form* of the sarcastic response we generate. Here, *emotion* is a function that augments $\neg\alpha$ to express a negative attitude. Note that the logical form, together with the coherence chain between $\alpha$ and the failed expectation $Q$, provide a complete explanation for *how* and *why* sarcasm occurs. The explanation is $\epsilon = (\text{emotion}(\neg\alpha), \mathcal{C})$, where is $C$ the coherence chain from $\alpha$ to $Q$. The coherence chain for each relation type can be selected from Table 1. This makes our sarcasm generation process accountable.

**Logical Form to Text** To convert the logical form to text, we rely on predefined patterns for each if-then relation type. As a running example,

[1] Note that a restriction in Utsumi (2000)'s definition of allusion is that $U$ does not directly express the state of affairs that $Q$ is expected via phrases such as "I've expected ...".

| relation type | example relation | coherence chain |
|---|---|---|
| xNeed | if $P$ then $\alpha =$ [xNeed to train hard] | volitional-cause$(\alpha, P)$ and contrast$(P, Q)$ |
| xAttr | if $P$ then $\alpha =$ [xAttr competitive] | condition$(\alpha, I_P) \wedge$ purpose$(I_P, P) \wedge$ contrast$(P, Q)$ |
| xReact | if $P$ then $\alpha =$ [xReact happy] | contrast$(Q, P) \wedge$ volitional-result$(P, \alpha)$ |
| xEffect | if $P$ then $\alpha =$ [xEffect gets congratulated] | contrast$(Q, P) \wedge$ non-volitional-result$(P, \alpha)$ |

Table 1: Coherence chains between the object $\alpha$ of an if-then relation and the failed expectation $Q$, for each relation type, as discussed in Section 4.2. Here, $P$ is the proposition expressed by the input text $U_{\text{in}}$. In the examples, $U_{\text{in}} =$'<user> won the marathon'.

assume the input utterance $U_{\text{in}} =$'<user> won the marathon' and the chosen relation type is *xAttr*. Say $\alpha$ = COMET$(U_{\text{in}}, \textit{xAttr})$ = [*xAttr* competitive]. The logical form is *emotion*$(\neg$[*xAttr* competitive]$)$. We first construct an intermediate utterance $U_\alpha$ using the following rule: *<user> <verb> competitive*. Here, *<verb>* is a verb specific to each relation type. In our example, $U_\alpha$ could be '<user> is competitive'. Next, for each input $U_{\text{in}}$, we generate three responses. The first response $U_{\text{out}}^{-e}$ only includes pragmatic insincerity, i.e. it expresses $\neg$[*xAttr* competitive]. To construct it, we apply a rule-based algorithm to generate the negation of $U_\alpha$ in a manner similar to (Chakrabarty et al., 2020), discussed in Section 2. $U_{\text{out}}^{-e}$ could be '<user> is not competitive'. The second response $U_{\text{out}}^{-i}$ does not include pragmatic insincerity, but only markers that express an emotional attitude, i.e. it expresses *emotion*$($[*xAttr* competitive]$)$. To achieve this, in a pattern-based manner, we augment $U_\alpha$ with hyperbole and interjections, as indicated by Utsumi (2000). $U_{\text{out}}^{-i}$ could be '<user> is definitely competitive, yay!'. The third response $U_{\text{out}}$ includes both devices, i.e. it expresses *emotion*$(\neg$[*xAttr* competitive]$)$. $U_{\text{out}}$ could be '<user> is definitely not competitive, yay!'. A full list of patterns is shown in Section A in the appendix.

In the running example we focused on the *xAttr* relation type. Recall there are four relation types that we consider, *xNeed*, *xAttr*, *xReact*, and *xEffect*. As such, for each input text $U_{\text{in}}$, we generate 12 responses: three response types, $U_{\text{out}}^{-e}$, $U_{\text{out}}^{-i}$, and $U_{\text{out}}$, for each relation type. We use the pattern Ch-<relation >$^{(|-i|-e)?}$ to refer to each response of our system, *Chandler*. For instance, Ch-xAttr refers to $U_{\text{out}}$ built considering the *xAttr* relation, while Ch-xNeed$^{-e}$ refers to $U_{\text{out}}^{-e}$ built considering the *xNeed* relation.

Note that other strategies for converting the logical form of sarcasm to text are possible. For instance, using policy-based generation with external rewards (Mishra et al., 2019) might have lead to higher perceived sarcasticness of our generated responses. However, we leave this to future work. Our goal is to understand user preferences towards when sarcasm should be used, and how sarcasm should be formulated.

### 4.3 Measuring Users' Preferences

We built three surveys, labelled (a)–(c), that we published on the Prolific Academic[2] crowdsourcing platform, one for each output type, out of $U_{\text{out}}^{-e}$, $U_{\text{out}}^{-i}$, and $U_{\text{out}}$. As such, in the survey corresponding to $U_{\text{out}}$, we presented participants with the input text $U_{\text{in}}$, along with the responses produced by Chandler-xNeed, Chandler-xAttr, Chandler-xReact, and Chandler-xEffect.

In each survey, we also enclosed a response from DialoGPT (Zhang et al., 2020), a recent dialogue system that is not built to be sarcastic; a response produced by SarcasmBot, the sarcastic response generator of Joshi et al. (2015) ; and a response produced by $R^3$, the state-of-the-art sarcastic paraphrase generator of Chakrabarty et al. (2020)[3].

We make a few observations. First, DialoGPT is used as a reference system, following the reasoning of Joshi et al. (2015): responses designed to be sarcastic should have a higher perceived sarcasticness than responses from DialoGPT, which are not designed to be sarcastic. Second, note that $R^3$ is designed to produce rephrases. As such, we applied $R^3$ to the output of DialoGPT to get a sarcastic rephrase of a response to the input. Table 2 shows an example input utterance, along with responses from all systems.

All in all, each survey instance contained a specific input text, and seven responses generated as mentioned above and presented in a random order. In the survey, we asked participants to evaluate each response across four dimensions: (1) Sarcasm: How sarcastic is the response? (2) Humour: How

[2] https://prolific.co

[3] https://github.com/tuhinjubcse/SarcasmGeneration-ACL2020

| system | response |
|---|---|
| DialoGPT | I'm not sure if you're being sarcastic or not. |
| DialoGPT+$R^3$ | I'm sure if you're being sarcastic or not. No one has yet been hurt. |
| SarcasmBot | That is a very useful piece of information! LMAO |
| Ch-xNeed | Yay! Good job not knowing how to write. |
| Ch-xAttr | Yay! You're not a very unintelligent person, that's for sure. |
| Ch-xReact | You're not feeling very embarrassed right now, that's for sure. Yay! |
| Ch-xEffect | You're not really going to sigh in frustration right now, that's for sure. Brilliant! |
| Ch-xNeed$^{-i}$ | You knew how to write, that's for sure. Good job! |
| Ch-xAttr$^{-i}$ | Brilliant! You're a very unintelligent person, that's for sure. |
| Ch-xReact$^{-i}$ | You're feeling very embarrassed right now, that's for sure. Brilliant! |
| Ch-xEffect$^{-i}$ | You're really going to sigh in frustration right now, that's for sure. Brilliant! |
| Ch-xNeed$^{-e}$ | You didn't know how to write. |
| Ch-xAttr$^{-e}$ | You're not unintelligent. |
| Ch-xReact$^{-e}$ | You're not feeling embarrassed right now. |
| Ch-xEffect$^{-e}$ | You're not going to sigh in frustration right now. |

Table 2: Responses generated by all systems to the utterance "I ran out of characters :drooling_face:", as discussed in Section 4.3.

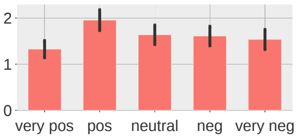

Figure 1: Mean sarcasm appropriateness score for each sentiment category, as discussed in Section 5.1. The error bars represent 95% confidence intervals.

funny is the response? (3) Coherence: How coherent is the response to the input? It is coherent if it sounds like sensible response that a person might give in a real conversation; and (4) Specificity: How specific is the response to the input? It is not specific if it can be used as a response to many other inputs. Each dimension ranged from 0 to 4, in line with previous work (Chakrabarty et al., 2020). Next, we asked participants to select their preferred response out of the seven, i.e. the one they would personally use. Finally, we asked them to judge, on a scale from 0 to 4, how appropriate it was to respond sarcastically to the shown input text.

Each survey instance was presented to three different participants. However, we did not use a voting scheme to aggregate the three survey instances into one. Rather, aggregation was conducted per-system. This is because our metrics (e.g. sarcasticness, preference towards a response, appropriateness) are inherently subjective, depending on the sociocultural background of the participants. See, for instance, the work of Oprea and Magdy (2020). As such, the concept of "correct answer" does not exist in the conventional sense. Indeed, the inter-participant agreement was low, but not surprisingly so, given that participants could have come from different sociocultural backgrounds. However, this does not entail that population statistics are not informative. As related work in this direction, consider that of Amidei et al. (2018), who make the point "an unchecked focus on reduction of disagreement among annotators runs the danger of creating generation goals that reward output that is more distant from, rather than closer to, natural human-like language." (Amidei et al., 2018) Consider also the work of Davani et al. (2021), who discuss the issue of disagreement in subjective tasks. We do, however, encourage more work in this direction.

## 5 Results

We now look at the responses that the participants provided in our survey, addressing our RQs.

### 5.1 RQ1: Should a Chatbot be Sarcastic?

#### 5.1.1 When is sarcasm appropriate?

Figure 1 shows the mean appropriateness score for each of the five sentiment categories. A one-way ANOVA test between the means yielded a $p$-value $\approx 0.001$. We therefore proceeded with Tukey's range test (Tukey, 1949), to find the means that are significantly different from one another. We noticed that sarcasm was considered significantly more appropriate by survey participants in responses to positive inputs, compared to very positive, and very negative inputs, respectively. This supports our statement from Section 2: the assumption of previous state-of-the-art generators that sarcasm should *only* be generated for negative inputs is problematic. However, even for the positive class, the mean appropriateness is less than 2. This makes it difficult to recommend responding sarcastically based on sentiment only.

To gain more insight, we proceeded with a qualitative inspection of the inputs that yielded the highest and lowest appropriateness scores, respectively. We noticed a few main themes, that we labelled *joke*, *family*, *school*, *leisure* and *death*. We then asked two humans to label all inputs across these dimensions. A third human resolved all disagreements. Finally, we computed the Pearson correlation coefficient of each theme with the sarcasm appropriateness score, across all inputs. We noticed a significant ($p < 0.05$) positive correlation

| text | approp. |
|---|---|
| I was a single mom with a sick child | 0 |
| I had a wonderful day thanks to my husband | 0 |
| I had such a great time with my family at my little prima's quince | 1 |

Table 3: Example inputs with low sarcasm appropriateness (approp.) score.

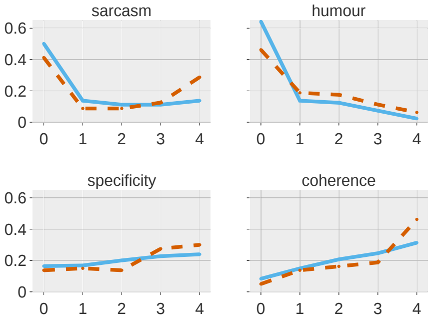

Figure 2: Distribution of the sarcasm, humour, specificity, and coherence scores of the *preferred* response; across all survey instances (continuous blue line) and across instances with a high sarcasm appropriateness (dashed red line), as discussed in Section 5.1.2.

between appropriateness and the category *joke*, and significant negative correlation with belonging to the *family* theme. We show some examples of the theme *family* with low appropriateness scores in Table 3.

Thus, according to our analysis, sarcasm seems to be most appropriate for positive inputs, and for humorous inputs, which may invite more sarcastic responses. In other situations, however, sarcasm might be interpreted as inappropriate and even offensive (Meaney et al., 2021).

### 5.1.2 When is sarcasm preferred?

We first consider the overall preference towards either sarcasm or non-sarcasm. Recall that participants also specified their preferred response for each input. The distribution of the sarcasm, humour, specificity, and coherence scores of this *preferred* response, across all survey instances, is illustrated in Figure 2 with a blue, continuous, line. The red, dashed, line illustrates the distribution across the 80 survey instances where the sarcasm appropriateness score of the input was higher than the midpoint, i.e. at least 3.

We notice considerably higher preference towards non-sarcastic and non-humorous responses. As indicated by the blue lines, over 50% of the preferred responses were those considered non-sarcastic and non-humorous by participants, the rest of the distribution being highly skewed towards the lower sarcasm and humour regions. Furthermore, note that even when sarcasm was considered highly appropriate, participants still preferred non-sarcastic responses, as indicated by the red, dashed, line in the top-left of Figure 2. Although there is a shift in the distribution towards sarcasm in this case, the skew is still towards the non-sarcastic region. Looking at the bottom row of Figure 2, on the other hand, we notice a negative skew, indicating an overall preference towards higher coherence. This is slightly the case for specificity as well.

To investigate further, we fit a logistic regression model to predict whether a response is preferred based on its sarcasm, humour, specificity, coherence scores, and two-way interactions between these variables. All coefficients are listed in Appendix B. We noticed noticed a significant ($p < 0.05$) positive relationship between coherence and preference, as well as the interaction between sarcasm and humour. The term representing the product of sarcasm and specificity had a significant negative effect on preference. In terms of the specific systems, we notice DialoGPT was preferred about 44% of the time, followed by Ch-xAttr$^{-i}$ (20%), and SarcasmBot (15%), which corresponds exactly to the coherence ranking in Table 4.

Our results indicate that responses with high coherence to the inputs are generally preferred over sarcastic responses. Sarcasm is only preferred when it is also considered humorous. On the other hand, participants seem to have actively avoided sarcastic responses that were very specific.

## 5.2 RQ2: How Should a Chatbot Formulate Sarcasm

### 5.2.1 Linguistic Devices

In Table 4 we show mean sarcasm, humour, specificity, and coherence scores provided by participants for each variant of Chandler, across all inputs. In the table, there are four groups (1–4) and three systems within each group (a–c). Rows with index (a) show scores for the complete versions of Chandler, for each if-then relation type. Rows (b) and (c) show partial versions, omitting pragmatic insincerity and emotional markers, respectively.

**Allusion** We have four strategies for alluding to the failed expectation, depending on the relation type considered. We notice the highest sarcasm score is achieved by Ch-xAttr (row 2a), followed by Ch-xNeed (row 1a), Ch-xReact (row 3a) and Ch-xEffect (row 4a). The same ranking holds for

|  |  | System | sarc. | hum. | coh. | spec. |
|---|---|---|---|---|---|---|
|  |  | DialoGPT | 0.6 | 0.3 | 2.3 | 2.0 |
|  |  | DialoGPT+$R^3$ | 0.8 | 0.3 | 0.9 | 1.3 |
|  |  | SarcasmBot | 2.5 | 0.8 | 1.4 | 0.9 |
| 1 | a. | Ch-xNeed | 1.9 | 0.6 | 1.3 | 1.6 |
|  | b. | Ch-xNeed$^{-i}$ | 1.5* | 0.5 | 1.7* | 1.9* |
|  | c. | Ch-xNeed$^{-e}$ | 1.0* | 0.4* | 1.5 | 1.7 |
| 2 | a. | Ch-xAttr | 2.1 | 0.6 | 1.3 | 1.4 |
|  | b. | Ch-xAttr$^{-i}$ | 1.6* | 0.6 | 1.8* | 1.7* |
|  | c. | Ch-xAttr$^{-e}$ | 1.1* | 0.4* | 1.3 | 1.2 |
| 3 | a. | Ch-xReact | 1.7 | 0.4 | 1.0 | 1.0 |
|  | b. | Ch-xReact$^{-i}$ | 1.4* | 0.4 | 1.3* | 1.3* |
|  | c. | Ch-xReact$^{-e}$ | 0.8* | 0.3* | 1.0 | 1.0 |
| 4 | a. | Ch-xEffect | 1.6 | 0.5 | 1.1 | 1.3 |
|  | b. | Ch-xEffect$^{-i}$ | 1.4 | 0.5 | 1.4* | 1.6* |
|  | c. | Ch-xEffect$^{-e}$ | 1.1* | 0.4 | 1.3 | 1.4 |

Table 4: Means of the sarcasm, humour, specificity, and coherence scores provided by participants, for each variant of Chandler (Ch). "*" indicates statistically significant difference from row (a) within the same numbered group (t-tests with Bonferroni correction, $p < 0.001$).

variants of Chandler that do not include pragmatic insincerity or emotional markers. Out of the allusion strategies selected, the responses perceived as most sarcastic are those that mention attributes of the user. Similarly, we notice that variants of Chandler that use the xAttr relation are also perceived and the most coherent, specific to the input, and achieve the highest humour score.

**Pragmatic Insincerity** Comparing the complete version, Ch-xAttr (row 2a), with Ch-xAttr$^{-i}$ (row 2b), the same model without pragmatic insincerity, we notice a significant drop in average sarcasm score. We observe a similar trend in group 3 for Ch-xReact$^{-i}$, indicating the importance of pragmatic insincerity. However, this did not hold for the other two relation types. Additionally, both specificity and coherence seem to significantly increase when removing pragmatic insincerity, irrespective of the relation type considered.

**Emotional Markers** Comparing complete versions of Chandler with those that omit emotional markers, we notice that the omission of such markers leads to significantly lower perceived sarcasm for all relation types. Humour is also significantly impacted by the omission of emotional markers for all relation types considered except for *xEffect* (row 4). Oh the other hand, coherence and specificity are not significantly influenced.

To sum up, the degree of perceived sarcasm is influenced by all linguistic devices considered. Out of the if-then relation types we consider, mentioning attributes of the user seems to lead to the highest perceived sarcasm, humour, specificity and coherence. Being insincere about the state of affairs leads to significantly higher perceived sarcasm, but significantly lower specificity and coherence. Emotional markers increase sarcasm and humour perception, but do not significantly impact specificity or coherence. Finally, recall that a main claim of IDT was that the degree of sarcasticness of an utterance grows with the number of implicit display conditions met. Our results support this claim.

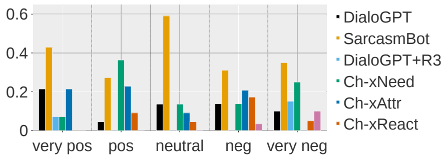

Figure 3: Normalized number of times each system was preferred for instances were the participant preferred a response that they also considered sarcastic.

### 5.2.2 Preferred Flavour

While we established that participants typically preferred non-sarcatic responses, we next set out to find what sarcasm people preferred in our experiments when they *did* prefer sarcasm. To do this, we consider the set of survey instances that showed the complete versions of Chandler, where the sarcasm score given by the participant to their preferred response was at least 3, leaving us with 107 (around 14%) of the 750 survey instances. We divide these instances into five categories, based on input sentiment. Within each category, for each generation system, we count the number of times that a response produced by that system was preferred. Figure 3 shows the normalised counts across all systems, for each sentiment category.

We observe that, for positive inputs, where sarcasm was considered significantly more appropriate than other sentiment categories, people prefer responses produced by Ch-xNeed. Interestingly, however, we observe that people prefer the fairly nonspecific, pattern-based sarcastic remarks produced by SarcasmBot for most types of input text. However, when analysing its outputs, we noticed it produced a total of only 28 unique responses (listed in Appendix C) to our 250 inputs. While in our experiments each response was only shown at most three times, in a real scenario of a user interacting with a conversational agent, the user might not appreciate repeatedly receiving the same response.

## 6 Recommendations

We recommend that future work on sarcasm generation should account for the four main findings: (1) People think sarcasm is *inappropriate* as a response to most inputs. However, if it is to be used, it is seen as most appropriate when the input is positive, but not extremely positive. People also found sarcasm to be a suitable response to jokes. (2) Even when deemed appropriate, people usually do not prefer sarcasm. Rather, coherence is the most important factor in explaining their response preferences. When people do prefer sarcasm, they like it mainly when it is also seen funny. Further, they generally dislike sarcasm that is very specific. (3) When generating sarcasm, pragmatic insincerity and emotional markers are important to include as they have a high influence of sarcasm perception. (4) Overall, people commonly prefer the simple general sarcastic responses of SarcasmBot, even compared to more sophisticated generation models, which suggests that presently, a simpler solution to sarcasm generation may actually be advantageous. Nevertheless, more investigation is required to examine if it will be desirable in long conversations, since it has limited diversity in outputs.

## 7 Conclusion

We have used a linguistically informed framework for sarcasm generation so that we could present human judges with a variety of flavors of sarcastic responses in a range of situations. Our findings suggest that sarcasm should not always be generated, but the decision to generate sarcasm itself should informed by user preferences. People find sarcasm most appropriate as a response to positive utterances and cases in which a joking environment has already been established. Further, judges preferred sarcasm most when they actually found it to be funny, and most often preferred general sarcastic responses. However, people often preferred non-sarcastic responses even more. We recommend that future work in this area carefully considers both the appropriateness and necessity of generating sarcasm at all.

## 8 Ethical Considerations

In our experiments, we noticed that some of the input tweets contained references to sensitive topics, such as religion and gender, or to tragic life events. Producing sarcasm for such inputs might be inappropriate and offensive to some (as our experiments confirmed). We clearly informed our survey participants about this possibility in the Participant Information Sheet, before accessing our survey. The sheet is enclosed in Appendix D.

## References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven R. Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of The 16th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mario J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *CoNLL*, pages 167–177. ACL.

David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *ICWSM*, pages 574–577. AAAI Press.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. ACL.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779. ACL.

Richard Bruntsch and Willibald Ruch. 2017. Studying irony detection beyond ironic criticism: Let's include ironic praise. *Frontiers in Psychology*, 8:606.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*, pages 7976–7986. ACL.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.

Herbert L Colston and Jennifer O'Brien. 2000a. Contrast and pragmatics in figurative language: Anything understatement can do, irony can do better. *Journal of Pragmatics*, 32(11):1557–1583.

Herbert L Colston and Jennifer O'Brien. 2000b. Contrast of Kind Versus Contrast of Magnitude: The Pragmatic Accomplishments of Irony and Hyperbole. *Discourse Processes*, 30(2):179–199.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*.

Shelly Dews and Ellen Winner. 1995. Muting the meaning: A social function of irony.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, Cambridge, UK.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. In *COLING*, pages 1837–1848. ACL.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of Pragmatics*, 26(5):613–634.

Aditya Joshi, Anoop Kunchukuttan, Mark James Carman, and Pushpak Bhattacharyya. 2015. Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *WISDOM at KDD*. ACM.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *EMNLP*, pages 1006–1011. ACL.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *LREC*. ELRA.

Roger J. Kreuz and Sam Glucksberg. 1989. How to Be Sarcastic: The Echoic Reminder Theory of Verbal Irony. *Journal of Experimental Psychology: General*, 118(4):374–386.

Roger J Kreuz, Debra L Long, and Mary B Church. 1991. On Being Ironic: Pragmatic and Mnemonic Implications. *Metaphor and Symbolic Activity*, 6(3):149–162.

Sachi Kumon-Nakamura, Sam Glucksberg, and Mary Brown. 1995. How about another piece of pie: The allusional pretense theory of discourse irony.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: Description and Construction of Text Structures*, pages 85–95. Springer Netherlands, Dordrecht.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7, hahackathon, detecting and rating humor and offense. In *ACL*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *EMNLP-IJCNLP*, pages 6144–6154. ACL.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2020. The effect of sociocultural variables on sarcasm communication online. *Proceedings of The 23rd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.

Silviu Oprea, Steven Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–349, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Penny M. Pexman and Meghan T. Zvaigzne. 2004. Does irony go better with friends? *Metaphor and Symbol*, 19(2):143–163.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *WSDM*, pages 97–106. ACM.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714. ACL.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. ACL.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Dan Sperber and Deirdre Wilson. 1998. Irony and relevance: A reply to seto, hamamoto and yamanashi. In R. Carston and S. Uchida, editors, *Relevance theory: Applications and implications*, pages 289–293. Benjamins, Amsterdam.

Martin Sykora, Suzanne Elayan, and Thomas W Jackson. 2020. A qualitative analysis of sarcasm, irony and related #hashtags on twitter. *Big Data & Society*, 7(2):2053951720972735.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Akira Utsumi. 1996. Implicit display theory of verbal irony: Towards a computational model of irony. In *Proceedings of the International Workshop on Computational Humor (IWCH'96)*.

Akira Utsumi. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *ACL*, pages 1035–1044. ACL.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua*, 87(1):53–76.

Steven Wilson and Rada Mihalcea. 2019. Predicting human activities from user-generated content. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2572–2582, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278. ACL.

# A Logical Form to Text Patterns

In this section we show the patterns used by Chandler to convert the logical form of sarcasm to text, as discussed in Section 4.2 of the main paper. We show patterns for each if-then relation type, *xNeed*, *xAttr*, *xReact*, and *xEffect*.

In the patterns below, <inten> is an intensifier, <suff_inten> is an intensifier added at the end of a phrase, <pos> is a positive emotion word, and <interj> an interjection. Inspired by (Utsumi, 2000) and (Joshi et al., 2015), each of these were randomly chosen from the following sets:

- <inten> : [very]
- <suff_inten> : [for sure]
- <pos> : [Good job, Well done]
- <intrj> : [Yay!, Brilliant!]

<obt> below is the object of the corresponding if-then relation object, as provided by COMET when taking in the input tweet.

## A.1 Patterns for the Complete Version of Chandler

*xNeed* patterns:

- You didn't <obt> , that's <suff_inten> . <pos> !

*xAttr* patterns:

- <interj> You're not <inten> <obt> , that's <suff_inten> .
- <interj> <pos> not being <obt> .
- <interj> You're not a very <obt> person that's <suff_inten> ."

*xReact* patterns:

- You're not feeling <inten> <obt> right now, that's <suff_inten> . <interj>

*xEffect* patterns:

- You're not <inten> going to obt_inf right now, that's <suff_inten> . <interj>

## A.2 Patterns for Chandler without Pragmatic Insincerity

*xNeed* patterns:

- You <obt> , that's <suff_inten> . <pos> !

*xAttr* patterns:

- <interj> You're <inten> <obt> , that's <suff_inten> .
- <interj> <pos> being <obt> .
- <interj> You're a very <obt> person that's <suff_inten> ."

*xReact* patterns:

- You're feeling <inten> <obt> right now, that's <suff_inten> . <interj>

*xEffect* patterns:

- You're <inten> going to obt_inf right now, that's <suff_inten> . <interj>

## A.3 Patterns for Chandler without Emotional Markers

*xNeed* patterns:

- You didn't <obt>.

*xAttr* patterns:

- You're not <obt>.
- You're not a <obt> person.

*xReact* patterns:

- You're not feeling <obt> right now.

*xEffect* patterns:

- You're not going to obt_inf right now.

## B Logistic Regression Coefficients

In Table 5 we present the full model parameters for the logistic regression experiment from section 5.1.2.

## C SarcasmBot Outputs

We noticed SarcasmBot produced a total of only 28 unique responses to our set of 250 inputs, as discussed in Section 5.2.2 of the main paper.

- Unbelievable that you just said 'sucky'! You are really very classy!
- Awesome!
- Brilliant!
- Let's party!
- Oh you poor thing!
- You owe me a drink for that awesome piece of news!
- Wow, you said 'sucks', didn't you? Your mom will be really proud of you!
- Wow, you said 'suck', didn't you? Your mom will be really proud of you!
- I'd feel terrible if I were you!
- You are such a simple person!
- Aww!! That's so adorable!
- That deserves an applause.
- I am so sorry for you!
- Yay! Yawn!
- How exciting! Yawn!
- How exciting! *rolls eyes*
- Wow! *rolls eyes*
- Yay! *rolls eyes*
- Yay! LMAO
- Wow! Yawn!
- How exciting! LMAO
- Wow! LMAO
- That is a very useful piece of information! *rolls eyes*
- That is a very useful piece of information! LMAO
- That is a very useful piece of information! Yawn!
- Unbelievable that you just said 'sobbing'! You are really very classy!
- Unbelievable that you just said 'sucks'! You are really very classy!
- Unbelievable that you just said 'bloody'! You are really very classy!

## D Participant Information Sheet

### D.1 What will I do?

Imagine someone (we'll call them PersonX), makes a statement. You will be shown a few responses to that statement. The responses were generated by chatbots (computer programs). Some sentences talk about sensitive topics, such as tragic life events. Responses to such sentences could be potentially inappropriate, or even offensive or harmful. Unfortunately, chatbots do not understand whether or not a topic is sensitive for a human. Please be fully aware of this when accepting to take part in our study.

For each response, you will be asked:

1. How sarcastic you find the response? (0 - not sarcastic, 3 - very sarcastic)

2. How funny you find the response? (0 - not funny, 3 - very funny)

3. How specific is the response to PersonX's statement? The response is specific if it mentions details that show a good understanding of PersonX's statement and its implications. Otherwise it's general. (0 - very general, 3 - very specific).

4. How coherent is the response to PersonX's statement? The response is coherent if it makes sense as a response. That is, it's a clear and sensible response that someone might actually give. It does not matter if it's specific or general. (0 - not coherent, 3 - very coherent).

Let's take a quick example. In this example, imagine that PersonX's statement is "I went to the grocery store". Here are some responses about this statement.

About being specific:

- "That's great." - Very general response. You can say this as a response to pretty much anything.

- "Nice to hear you are enjoying this sunny day." - General response. It does provides some details about the day (that it's sunny). However, those details are not uniquely related to PersonX's statement.

- "You must be tired." - More specific response. It shows an understanding that going somewhere (anywhere at all) may cause tiredness.

|  | coef | std err | z | $P > \|z\|$ | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -3.1228 | 0.140 | -22.369 | 0.000 | -3.396 | -2.849 |
| sarcasm | -0.1328 | 0.070 | -1.897 | 0.058 | -0.270 | 0.004 |
| humour | 0.0608 | 0.133 | 0.457 | 0.647 | -0.200 | 0.321 |
| specificity | 0.1338 | 0.087 | 1.542 | 0.123 | -0.036 | 0.304 |
| coherence | 0.8261 | 0.072 | 11.508 | 0.000 | 0.685 | 0.967 |
| sarcasm*humour | 0.1178 | 0.031 | 3.861 | 0.000 | 0.058 | 0.178 |
| sarcasm*specificity | -0.0620 | 0.031 | -1.990 | 0.047 | -0.123 | -0.001 |
| sarcasm*coherence | -0.0624 | 0.032 | -1.961 | 0.050 | -0.125 | -2.61e-05 |
| humour*specificity | 0.0100 | 0.044 | 0.225 | 0.822 | -0.077 | 0.097 |
| humour*coherence | -0.0487 | 0.047 | -1.038 | 0.299 | -0.141 | 0.043 |
| specificity*coherence | 0.0073 | 0.026 | 0.281 | 0.779 | -0.044 | 0.058 |

Table 5: Detailed results of logistic regression described in section 5.1.2.

- "You probably bought a lot of vegetables." - Specific response. It shows an understanding of what a grocery store is. That is, a place where you can probably buy vegetables.

- "You must have been quite hungry for carrots." - Very specific response. It shows an understanding of what a grocery store is, about what carrots are, and about the link between carrots and the store (mainly, that carrots are sold there).

About being coherent:

- "I'm cold." - Not coherent. It has nothing to do with PersonX's statement

- "I went to the grocery store". It's not a suitable response that someone would normally give.

- "I had such a wonderful dream last night, there were a lot of awesome cars painted blue." - Not coherent. It does not make sense as a response to PersonX's statement.

- "I sometimes dream about eating carrots." - More coherent response. Someone might sometimes say this as a response, although it's not a common response.

- "OK thanks." - Very coherent. One might actually say this as a response. Notice it's not specific to PersonX's statement. You can say it as a response to many other statements. Still, it's coherent to PersonX's statement. Thanks a lot for getting me those carrots, I'll pay you back next week. - Very coherent and very specific to PersonX's statement.

### D.2 Participant Information Sheet and Consent Form

- Principal investigator: ⟨our PI's name⟩

- Researcher collecting data: ⟨researcher's name⟩

- Funder (if applicable): ⟨funding bodies⟩

This study is in the process of being certified according to the ⟨details about the ethics committee of our institution ⟩. Please take time to read the following information carefully. You should keep this page for your records.

### D.3 Who are the researchers?

We are the ⟨name of our group⟩group, a research group that brings together a range of researchers from ⟨our institution⟩in order to build on our existing strengths in social media research. This research group focuses on mining structures and behaviours in social networks. The principal investigator is ⟨our PI's name⟩.

### D.4 What is the purpose of the study?

This study aims to understand what linguistic style people associate with sarcasm.

### D.5 Why have I been asked to take part?

We target everyone registered as living in ⟨country⟩on the Prolific Academic platform.

### D.6 Do I have to take part?

No—participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

### D.7 What will happen if I decide to take part?

You will be asked to fill in a survey. The flow of the survey is the following:

- You will be shown a short text (originating from a tweet) and asked whether it is, in your view, appropriate to respond sarcastically to that text.
- If you say “no”, you will be shown another text. The process will repeat until you say “yes” or 10 texts have been shown.
- If you say “yes”:
  - You will be shown 7 responses to the text that you selected;
  - For each response, you will be asked to specify, on a scale from 1 to 5: (a) How sarcastic it is; (b) How funny it is; (c) How coherent it is to the original text; It is coherent if it sounds like a reasonable response that a person might give. (d) How specific it is to the original text; It is specific if it mentions details about the original text, or its implications, that make this response not appropriate as a response to many other texts.

We estimate it will take around 3 minutes to complete the survey.

### D.8 Compensation

You will be paid £0.38 for your participation in this study.

### D.9 Are there any risks associated with taking part?

Please note: some of the texts that you will see include content that you might consider sensitive, or might trigger unwanted memories. For instance, they might mention losing a family member, losing friends, break-ups, failure in exams, or health issues.

### D.10 Are there any benefits associated with taking part?

Financial compensation of £0.38.

### D.11 What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

### D.12 Data protection and confidentiality

Your data will be processed in accordance with Data Protection Law. Throughout your entire interaction with us, the only information collected about you specifically is your Prolific Academic identification number. This data will only be viewed by the team members of the ⟨our group⟩group, listed here: ⟨our group’s website⟩. All other data, including the responses you provide, and the amount of time you took to fill in the survey, will be made public on the internet as part of Open Science, available to be indexed by search engines. The Open Science initiative is described here:
`https://en.wikipedia.org/wiki/Open_science`.

### D.13 What are my data protection rights?

⟨our institution⟩is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. However, we will have no control for the data that will be made public, as specific in the previous section. For more details, including the right to lodge a complaint with the Information Commissioner’s Office, please visit ⟨website of the datathe Information Commissioner’s office⟩. Questions, comments and requests about your personal data can also be sent to ⟨the data protection officer at our institution⟩. For general information about how we use your data, go to: ⟨website with information on research privacy at our institution⟩.

### D.14 Who can I contact?

If you have any further questions about the study, please contact the lead researcher, ⟨lead researcher’s name and email address⟩. If you wish to make a complaint about the study, please contact ⟨email address of the ethics committee at our institution⟩. When you contact us, please provide the study title and detail the nature of your complaint.

### D.15 Updated information

If the research project changes in any way, an updated Participant Information Sheet will be made available on ⟨website where updates are published⟩.

### D.16 Consent

By proceeding with the study, you agree to all of the following statements:

- I have read and understood the above information.
- I understand that my participation is voluntary, and I can withdraw at any time.
- I consent to my anonymised data being used in academic publications and presentations, as well as published publicly on the internet, as part of Open Science.
- I am aware that I will see potentially offensive, harmful, or hurtful content.
- I allow my data to be used in future ethically approved research.