Computational Sarcasm Detection Examined Through the Lens of Psycholinguistics

Doctoral thesis for obtaining the

academic degree

Doctor of Philosophy (Dr. phil.)

submitted by

Jang, Hyewon

at the

Universität Konstanz

Faculty of Humanities

Department of Linguistics

Konstanz, 2024

Computational sarcasm detection examined through the lens of psycholinguistics

Abstract

This thesis examines sarcasm detection from the combined angles of psycholinguistics and computational linguistics. It provides empirical evidence about various factors that cause sarcasm and affect the understanding of it in human communication. The thesis also uses those findings to deepen our understanding of computational sarcasm detection models.

Our first contribution is to reveal how certain important factors relate to the production and interpretation of sarcasm. Two of our experiments examine sarcasm's relation to context, emotion, and communicative intent. We find that emotion and communicative intent are highly associated with the choice to use sarcasm. Using two further experiments, we clarify the relation between speaker perspectives and observer perspectives in navigating communicative situations involving sarcasm—a topic that is an important but neglected aspect of computational sarcasm detection.

Our second contribution is to use the key factors just mentioned as a means of investigating the nature of the automatic sarcasm detection process. We report several patterns identified through three computational experiments, whose united goal is to reveal what kinds of information are encoded in sarcasm detection models, and whether they are similar to the information shown in human behavior. The first finding is that generalizable automatic sarcasm detection is difficult, due to the varied characteristics that sarcasm has. Additionally, we find that the incongruity between the speaker's affect and the content of the utterance makes the identification of sarcasm challenging both for human observers, and language models. Finally, we show that different amounts of context play different roles in computational sarcasm detection, and that this depends on the level of disagreement about sarcasm among human observers.

Our last contribution in this thesis is to introduce a new framework that smoothly integrates experimental and computational investigation of sarcasm, which can be

applied to other linguistic phenomena in future research. Our proposed framework, which is proven effective through multiple studies reported in this thesis, should bridge human communication and natural language processing, and contribute to a more multifaceted investigation of natural language in future work.

Zusammenfassung

In dieser Arbeit wird die Erkennung von Sarkasmus aus dem kombinierten Blickwinkel von Psycholinguistik und Computerlinguistik untersucht. Sie liefert empirische Erkenntnisse über verschiedene Faktoren, die Sarkasmus verursachen und das Verständnis von Sarkasmus in der menschlichen Kommunikation beeinflussen. Die Arbeit nutzt diese Erkenntnisse auch, um unser Verständnis von computergestützten Modellen zur Sarkasmuserkennung zu vertiefen.

Unser erster Beitrag besteht darin, aufzuzeigen, wie bestimmte wichtige Faktoren mit der Produktion und Interpretation von Sarkasmus zusammenhängen. Zwei unserer Experimente untersuchen die Beziehung von Sarkasmus zu Kontext, Emotion und kommunikativer Absicht. Wir stellen fest, dass Emotionen und kommunikative Absichten in hohem Maße mit der Entscheidung, Sarkasmus zu verwenden, verbunden sind. In zwei weiteren Experimenten klären wir die Beziehung zwischen Sprecher- und Beobachterperspektive bei der Navigation in kommunikativen Situationen, die Sarkasmus beinhalten - ein Thema, das ein wichtiger, aber vernachlässigter Aspekt der computergestützten Sarkasmuserkennung ist.

Unser zweiter Beitrag besteht darin, die soeben genannten Schlüsselfaktoren als Mittel zur Untersuchung der Art des automatischen Sarkasmus-Erkennungsprozesses zu nutzen. Wir berichten über mehrere Muster, die durch drei Computerexperimente identifiziert wurden, deren gemeinsames Ziel es ist, herauszufinden, welche Arten von Informationen in Sarkasmus-Erkennungsmodellen kodiert werden und ob sie den Informationen ähnlich sind, die im menschlichen Verhalten gezeigt werden. Die erste Erkenntnis ist, dass eine verallgemeinerbare automatische Sarkasmuserkennung aufgrund der vielfältigen Eigenschaften von Sarkasmus schwierig ist. Außerdem stellen wir fest, dass die Inkongruenz zwischen dem Affekt des Sprechers und dem Inhalt der Äußerung die Erkennung von Sarkasmus sowohl für menschliche Beobachter als auch für Sprachmodelle schwierig macht. Schließlich

zeigen wir, dass unterschiedliche Mengen an Kontext eine unterschiedliche Rolle bei der computergestützten Erkennung von Sarkasmus spielen, und dass dies vom Grad der Uneinigkeit über Sarkasmus unter menschlichen Beobachtern abhängt.

Unser letzter Beitrag in dieser Arbeit ist die Einführung eines neuen Rahmens, der die experimentelle und computergestützte Untersuchung von Sarkasmus nahtlos integriert und der in der zukünftigen Forschung auch auf andere sprachliche Phänomene angewendet werden kann. Der von uns vorgeschlagene Rahmen, der sich durch mehrere Studien in dieser Arbeit als effektiv erwiesen hat, sollte eine Brücke zwischen menschlicher Kommunikation und der Verarbeitung natürlicher Sprache schlagen und zu einer vielseitigeren Untersuchung natürlicher Sprache in der zukünftigen Arbeit beitragen.

Acknowledgement

I have always known that I am a person interested in languages and how the human mind processes them. Thanks to some very helpful people, in the fall of 2019, I got an opportunity to focus that interest on studying computational linguistics in the MA program Speech and Language Processing at the University of Konstanz. After five years of steep learning curves and trial and error, the usual ups and downs, and some drama here and there, I am about to submit my PhD thesis – a small piece of work about my two favorite topics 'language' and 'processing'.

As much as it feels good to have achieved something, I am well-aware of the people that I owe for guiding me through this whole process. Without a doubt, the most credit goes to my main advisor Diego Frassinelli, who gave me the opportunity to start and finish this bumpy journey of doing a PhD. I thank him for several things.

First of all, I thank him for giving me the freedom and autonomy that I needed to keep my interest in research. He has encouraged me to choose a topic that genuinely interested me and to take control of my research projects. I also thank him for providing me with helpful feedback in our meetings or whenever I needed it. Every time before our regular meetings, I would have a mound of questions that needed resolving before I could move on to the next step. And Diego helped me resolve them first by creating a space for me to freely share my thoughts that were all tangled up in my brain, and then providing open feedback that helped me turn those thoughts into more coherent and substantial action plans. I also thank him for all the detailed feedback on my numerous drafts of papers, which made me feel that I was being taken care of. Lastly, I thank him for actively trying to make my research conditions better so I could fully focus on my PhD project without any concern about practical limitations.

Diego's supervision was excellently complemented by my other advisor Bettina Braun. Even though my topic was a bit different from her strongest expertise, Bet-

tina was excellent in providing me with insightful feedback and helping me with technical details. She would often listen to my ideas and plans from scratch and come up with an insightful thought about how to go about them. Several points that she made in our meetings have steered my PhD project into the directions they unfolded in. I also thank her for always being a fair and reasonable figure that I could trust.

There are several people to thank among the non-advisors, too. First, I thank Qi Yu and Moritz Jakob for collaborating with me on research projects. I thank Massimiliano Canzi for his help with my experiments and for facilitating various social events. I thank Andrea Ferreira for being a very helpful class tutor. I thank the SLP students who came to my reading group, because I learned a lot from it. And I thank everyone in the Department of Linguistics at the University of Konstanz who influenced my work in a positive direction and made my time in Konstanz better than it could have been.

I thank Sabine Schulte im Walde and the students in her research group for their feedback on the initial part of my project.

I thank my office mates – First, I thank Zlata Kikteva for making the beginning of my PhD time fun and exciting with her powerful energy and keeping the fun alive even after moving away. And I thank Philine Link for filling my last year of PhD with inspirations and positive attitudes, and also just for our goofy giggly girl times.

I thank my parents for trying their best to understand my work and for keenly trying to learn how I am progressing through my PhD from so far away.

Lastly, I thank Colin for consistently stimulating my brain with his clever ideas and amusing actions, for giving me plenty of emotional support, and for treating me like a cat – with tenderness and affection.

Despite the common belief that pursuing a PhD is a painstaking endeavor, I dare say that it turned out to be the most fun years of my life. Doing a PhD was an excellent opportunity for a multi-focal brain training. I learned how to pick a topic worth investigating, how to make it concrete, and how to implement it. I learned how to work on creating small pieces of work first and how to glue them together to create a coherent bigger piece. I learned how to be completely immersed in a project without losing control. Learning these things made me stronger as a researcher and also as an individual. I am grateful for that.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background on sarcasm research in different academic fields

Sarcasm is a very common form of figurative language (D'Arcey et al., 2019; Dews et al., 1995). Despite its prevalence, it is quite complex, and is often closely related to other linguistic phenomena such as verbal irony, hyperbole, rhetorical questions, and humor. For this reason, it has attracted a great deal of attention in academic research (Attardo, 2000; Fox Tree et al., 2020; Hull et al., 2017; Neuhaus, 2023). Several fields such as theoretical pragmatics and psycholinguistics have asked a variety of questions about the nature of sarcasm, which we overview here.

1.1.1 Theories of sarcasm

First, there is the question of what sarcasm is. What is the unique essence of sarcasm that distinguishes it from other ways of communicating (Kreuz & Glucksberg, 1989)? How is it similar to/different from other related linguistic phenomena such as verbal irony or humor (Attardo, 2000; Rockwell, 2003)? These are the questions that research in theoretical pragmatics has mostly focused on, in the pursuit of revealing what the core essence of sarcasm is.

In the Gricean theory of sarcasm (Grice, 1975), sarcasm is identified at the presence of a blatant violation of the maxim of quality, as exemplified in (1) below:

(1) *Context*: Your friend was sure that it would not rain today, but you realize

that in fact, it is raining today.
Your response: What a great day!

The untruthful nature of the response in (1) qualifies it as sarcasm. In Echoic theories (Sperber, 1984; Sperber & Wilson, 1981), a speaker speaks sarcastically by "echoing" (as opposed to "using") an utterance to convey a negative attitude. An echoic utterance alludes to the thoughts or utterances of others, which reminds the listener of norms or failed expectations, and allows them to understand that the utterance is sarcastic. According to this theory, (1) is sarcastic because the speaker is merely "echoing" the previous anticipation of a sunny day to express their negative attitude towards that anticipation (negative because the weather turns out to be bad). Similarly, in the Pretense theory (Clark & Gerrig, 1984), a speaker (S) "pretends" to be an alternative speaker (S') speaking to an alternative listener or hearer of the utterance (H'). S poses a negative attitude towards the utterance of S', and H' is ignorant of this fact and takes it literally, while H is expected to understand everything, and therefore understand that the utterance is sarcastic. Under this theory, in (1) the speaker thinks that the weather is bad, but pretends to be someone who thinks that the weather is good, while also assuming the hypothetical presence of the pretend listener who believes that statement, and finally, the speaker also intends for the actual listener to understand all of these different layers simultaneously. In contrast, in the Implicit Display theory (Utsumi, 2000), a sarcastic utterance expresses the fact that a speaker has an unmet expectation, and conveys a negative attitude toward the failed expectation. In this theory, the speaker in (1) had an expectation that their friend's belief about the weather would be true, and expresses their negative attitude when the belief turns out to be wrong. Furthermore, in the Relevant Inappropriateness theory (Attardo, 2000), a sarcastic remark is an inappropriate utterance but nevertheless relevant to the context, which the speaker intentionally makes inappropriate, and also expects the listener to construe it that way. From the perspective of this theory, in (1) the response is relevant to the context though inappropriate as the presupposition of the context (i.e., the weather is bad) is violated in the response.

These analyses of sarcasm, which share common concepts such as the expression of an attitude or the contradiction of the real message, provide insight into understanding what sarcasm is and how it operates. As a result, a classical definition of

sarcasm that many researchers adopt is "the utterance of the opposite of the true meaning" (Gibbs, 1986; Glucksberg, 1995; Kumon-Nakamura et al., 1995). Nevertheless, it is still hard to pin down sarcasm with a single description (Gibbs, 2000). Despite the many definitions and analyses of sarcasm, many open questions remain about what exactly sarcasm entails in reality.

1.1.2 The psycholinguistics of sarcasm

One such question is about the reasons or motivations for choosing to use sarcasm in the first place. Why do people speak sarcastically at all (Dews et al., 1995; Roberts & Kreuz, 1994)? What communicative functions does sarcasm allow speakers to convey (Colston, 2023)? What do people achieve by being sarcastic (Dews et al., 1995; N. Zhu & Wang, 2020)?

Also, if a speaker makes a sarcastic utterance for whatever reason, there is often someone who hears it and reacts to it. After all, sarcasm is a communicative act. This motivates yet more questions, which relate to the emotional reaction (affect) of the listener to sarcasm. How does one feel about a sarcastic comment, as opposed to a non-sarcastic one (Leggitt & Gibbs, 2000)? Does sarcasm amuse or hurt the conversational partner (Pexman & Olineck, 2002)? Do people process sarcasm differently when it is used among friends, as opposed to strangers (Pexman & Zvaigzne, 2004)? Is sarcasm considered harsher or softer depending on the preceding context or social dynamics (Boylan & Katz, 2013)? These questions are often asked in psychology and experimental linguistics.

1.1.3 Computational sarcasm

As a result, there is a general understanding about the way sarcasm works in human communication, which has been informed by a great deal of empirical evidence. This has led to a further question, about the way sarcasm is handled by non-humans—specifically, artificial language models. The most basic and extensively addressed aspect of this question is how to make computational models detect sarcasm in text. Before the advent of large language models, sarcasm detection by artificial models often involved feature engineering (Babanejad et al., 2020; Das & Kolya, 2021; A. Ghosh & Veale, 2017; Riloff et al., 2013; Tsur, 2010), or neural networks with text representation using Transformers models (Baruah et al., 2020;

Kumar & Anand, 2020; Potamias et al., 2020; Ren et al., 2023; Zhang et al., 2023). Some work in sarcasm detection also incorporates related factors such as context (Baruah et al., 2020; Rajadesingan et al., 2015) and multimodal cues (Lu et al., 2024; Tian et al., 2023).

Despite the abundance of prior work that shows how artificial models are able to detect sarcasm, a significant gap exists that should be addressed, both to strengthen our understanding of sarcasm in human communication and to create artificial systems that can handle sarcasm more accurately. For one, there is a gap regarding the limited scope of data used for computational modeling of sarcasm. Most sarcasm detection work heavily focuses on social media as a data source (Baruah et al., 2020; Cai et al., 2019; Lu et al., 2024; Misra & Arora, 2023; Potamias et al., 2020; Rajadesingan et al., 2015; Tan et al., 2023; Yue et al., 2023). Due to convenience and availability, many studies have used data from online sources such as social media (A. Ghosh & Veale, 2016; Khodak et al., 2018; Ptáček et al., 2014; Tan et al., 2023) and product reviews (Davidov et al., 2010; Filatova, 2017) to develop systems for sarcasm detection. Though it is a valid research design to use such data to further our understanding of sarcasm, it poses potential issues for comprehensive research, as it restricts the scope of sarcasm in computational linguistics. Such online data sources are limited to communications between strangers, or occurring in specific domains (e.g., Twitter, Reddit). But human communication does not happen exclusively online, and online communication takes on different forms than face-to-face communication (Aguert et al., 2016). For example, prior work suggests that sarcasm happens more often in online communications than face-to-face ones due to the lower risk associated with using sarcasm in anonymous settings (Fox Tree et al., 2020; Hancock, 2004). As will be demonstrated in Chapter 3, one often finds templated or simplistic instances of sarcasm suited for social media in the data used in many computational experiments, which does not always reflect the kinds of sarcasm used in everyday conversations.

Another current gap in computational sarcasm research is the scarcity of work that leverages what we already understand about sarcasm through psycholinguistic work. As the focus of prior computational work has generally been to demonstrate the efficacy of a proposed system, such work has paid relatively little attention to the auxiliary elements that are affiliated with sarcasm, which we have extensive knowledge about from psycholinguistic and theoretical research. This is not to say

that there is zero work that makes use of previous knowledge for sarcasm detection. For one, there have been attempts to leverage the classical definition of sarcasm as "saying the opposite of what is intended" in sarcasm detection (Riloff et al., 2013), but as more complicated neural models become preferred, such work has become rare. There are also attempts to integrate contextual information or emotion information into sarcasm detection models (Tan et al., 2023; Vitman et al., 2023), but the connection of these elements with sarcasm is mostly loose in this type of work. The focus of such research is typically, for example, the effective integration of emotion detection models or image information (=context) into sarcasm detectors, but not to investigate the question of what kinds of emotions or contexts would affect the use of sarcasm in order to integrate these auxiliary elements to the models. In this thesis, we define *context* as a preceding situation that has a potential to motivate speakers to speak in certain manners (in this case, sarcastically). We examine the kinds of information encoded in sarcasm detection models by leveraging such contexts with causal connections to the utterance.

A more fine-grained way of leveraging the relevant factors into computational models can help us examine sarcasm detection with more clarity, such that humans can interpret the grounds for the success of sarcasm detection models. This deepens our understanding of the process of artificial sarcasm detection.

1.2 Motivation

This thesis is motivated to ameliorate the limited focus on specific types of sarcasm data, and the lack of integration of prior knowledge about sarcasm, in computational work. We leverage concepts relevant to sarcasm (e.g., context, affect, intent, perspective) from multiple angles and integrate them into computational sarcasm detection models to show what kinds of knowledge are encoded in these models.

Specifically, the starting point of this work is to create a sufficiently large dataset of psycholinguistic information about sarcasm. We conduct experiments with different focuses and hypotheses to collect sarcasm data that contains clearly motivated sarcastic utterances, and the associated contexts. Through this process, we amass additional knowledge about the production and comprehension of sarcasm, and show how this new knowledge connects to prior research. Based on this, we go on to assess sarcasm detection models in light of the newly collected data, and

integrate this psycholinguistic knowledge surrounding sarcastic utterances into the computational models to test the capabilities and underlying mechanisms of them.

1.3 Contributions of this thesis

1.3.1 Research questions

The primary research questions (RQs) in this thesis are the following:

RQ 1. What contextual factors motivate speakers to use sarcasm?

RQ 2. What commonalities, and what differences, do speakers and observers have when identifying a remark as sarcastic?

RQ 3. What are the capabilities of language models for detecting sarcasm from different sources?

RQ 4. What causes sarcasm failure in communication?

RQ 5. How do models handle sarcasm when humans disagree with one another?

1.3.2 Contributions

By answering those research questions, this thesis makes three main contributions:

1. It presents new findings about sarcasm production from psycholinguistic experiments, which have received relatively little attention in prior work compared to sarcasm comprehension.
2. It presents new findings about the information encoded in sarcasm detection models by using intermediary factors that are important in human communication – context, affect and disagreement.
3. It introduces a new framework that connects (psycholinguistic) experimental methodologies with computational research, which can be applied to other topics in future investigation.

1.3.3 Key points

We make the following arguments based on the experiments reported in this thesis. Each point is addressed in detail in the indicated sections:

1. Sarcasm often occurs because of a certain affect (emotional reaction to a situation) that a context motivates speakers to have (Sections 2.1 and 2.3).
2. Observers can mostly identify sarcasm used by speakers as well as the underlying affect of the speakers (Sections 2.2 and 2.4).
3. The strongest factor that motivates speakers to use sarcasm lies in the content of the context (*what is said*) rather than the auxiliary factors such as *how it is said* or *who says it* (Section 2.5).
4. Factors that influence the use of sarcasm in human communication can be used as keys to access computational sarcasm models and to reveal hidden facts about how they detect it (Section 3).
5. Sarcasm is broader and more complex than is claimed in previous computational work (Section 3.1).
6. Miscommunications involving sarcasm occur partially due to the broken link between the speakers' affect and their utterance, which poses a significant difficulty both for humans and language models (Section 3.2).

1.4 Organization of this thesis

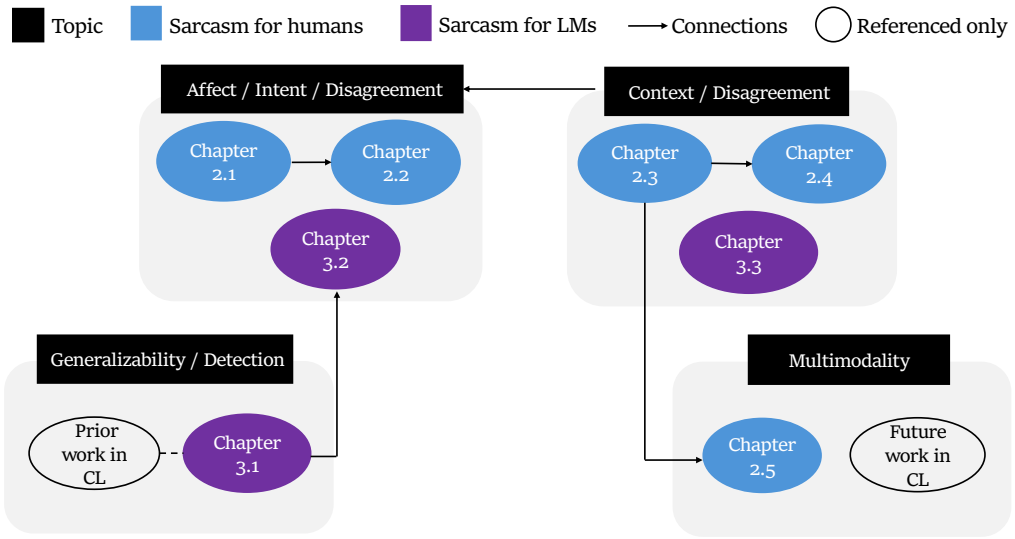

Figure 1.1: Reading threads for this thesis

Figure 1.1 provides an illustration of the structure of this thesis, which consists of four chapters. Chapter 1 (the current chapter) provides an introduction and back-

ground about sarcasm research in different academic fields. Chapter 2 focuses on sarcasm in human communication. It reports findings from five behavioral experiments about various factors that contribute to understanding sarcasm. Sections 2.1 and 2.2 focus on the connection between affect, intent, and sarcasm from the perspectives of speakers and observers, respectively. Sections 2.3 and 2.4 zoom in on the connection between context and sarcasm from the perspectives of speakers and observers, respectively. Section 2.5 addresses the connection between auxiliary factors (e.g., modality and delivery style) and sarcasm. Chapter 3 focuses on sarcasm detection by language models. It reports findings from three experiments with language models connected to the findings and data collected in Chapter 2. Section 3.1 focuses on the generalizability of sarcasm detection models motivated by the fact that sarcasm comes in different styles depending on the sources. Section 3.2 addresses the connection between affect and sarcasm encoded in language models, and its similarity to the behavior of human observers. Section 3.3 examines the relationship between the amount of context and disagreement among human observers in its effect on language models for sarcasm detection. Lastly, Chapter 4 summarizes the thesis, and discusses its limitations as well as topics for future work.

1.5 Author's publications

This thesis is based on the following publications. For all these publications, the author of this thesis contributed to methodology, stimuli creation, experiment running, data processing, formal analysis, visualization, as well as writing and editing of the papers:

1. **Hyewon Jang**, Diego Frassinelli, Generalizable Sarcasm Detection is Just Around the Corner, of Course!, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024).* (Section 3.1)
2. **Hyewon Jang**, Moritz Jakob, Diego Frassinelli, Context vs. Human Disagreement in Sarcasm Detection, *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang) @ NAACL 2024, Association for Computational Linguistics.* (Section 3.3)
3. **Hyewon Jang**, Bettina Braun, Diego Frassinelli, Intended and Perceived Sarcasm Between Close Friends: What Triggers Sarcasm and What Gets Conveyed?

Proceedings of the 45th Annual Conference of the Cognitive Science Society (CogSci 2023). (Sections 2.1 & 2.2)

4. **Hyewon Jang**, Bettina Braun, Diego Frassinelli, Contextual factors that trigger sarcasm, *under review at the journal of Metaphor and Symbol* (Sections 2.1 $\sim$ 2.3).

Chapter 2

Sarcasm for humans

This chapter reports new findings from five experimental studies with connected research questions about the factors that affect sarcasm use in human communication[1]. It discusses when sarcasm is to be expected, what motivates the use of sarcasm, which types of affect are associated with sarcasm, and how any of it is (or is not) transmitted to listeners.

We start from one of the most widely discussed functions of sarcasm, which is to express one's attitude or specific discourse goals (Colston, 2023; Glucksberg, 1995; Roberts & Kreuz, 1994). In our first study (Section 2.1), we connect two essential questions of "why do people use sarcasm?" (i.e., to achieve certain communicative functions) and "when do people use sarcasm?" (i.e., what kinds of situations would trigger sarcasm more than others?). As we expect certain contexts to prompt a speaker to develop a motivation to convey communicative intent that may be expressed through the use of sarcasm (see Figure 2.1), we provide several contextual prompts that are likely to motivate communicative functions of sarcasm which have been proposed in previous research (Colston, 2023, 1997; Dews & Winner, 1995; Gibbs, 2000), and assess the relation between context, affect (emotion in reaction to an experience), and the use of sarcasm. Then we reinforce our initial findings by shifting the focus to the finer-grained characteristics of contexts that motivate sarcasm use, and by expanding the social relationship between interlocutors (Section 2.3).

The previously mentioned studies focus on the production of sarcasm. In the next two studies, we examine the relation between sarcasm production and compre-

[1] In this thesis, we follow Fox Tree et al. (2020) and the references therein in collapsing sarcasm and verbal irony into one category and consider them as synonymous.

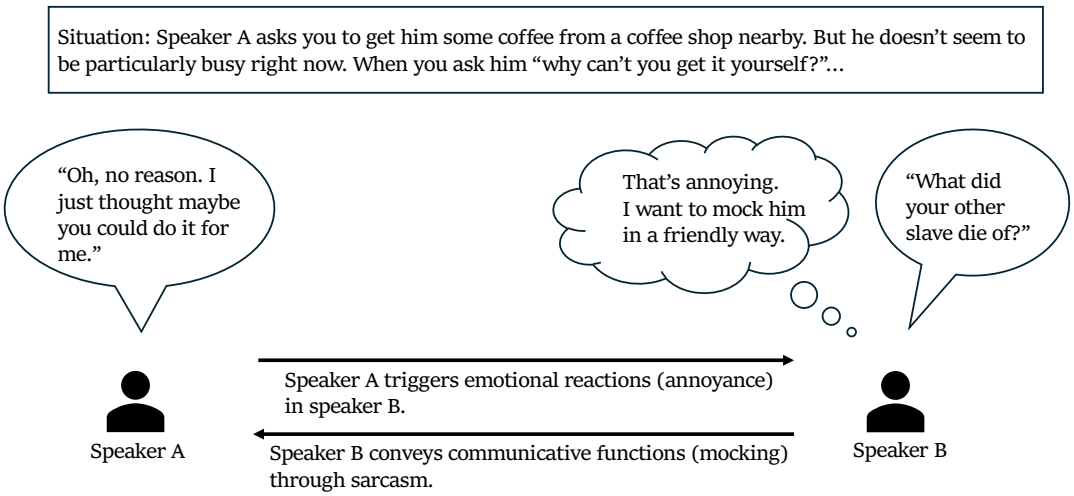

Figure 2.1: Hypothesis motivating the general structure of our experiments.

hension (Sections 2.2 and 2.4). We identify the commonalities and the differences between speakers and observers about their intended and perceived sarcasm. In our last experiment (Section 2.5), we reinforce our findings by bringing in modalities other than the text, which allows for the introduction of other elements such as the delivery style of context (i.e., how it is said) as opposed to the content itself (i.e., what is said) in examining the influence of context in triggering sarcastic intent. Some of these discussion points reemerge as important pointers for the experiments in Chapter 3.

2.1 Study 1: The role of affect and intent in sarcasm production

2.1.1 Background

The question of "why does one use sarcasm at all?" has been extensively studied in previous work, often with a focus on understanding why one would choose to say a sarcastic remark instead of a literal one (Toplak & Katz, 2000; Roberts & Kreuz, 1994), and what communicative goals one tries to accomplish by using sarcasm (Dews et al., 1995; Glucksberg, 1995). Many communicative functions of sarcasm have since been identified, such as the muting or enhancement of criticism, humor, mockery, or face-saving.

The muting of criticism. According to the Muting the Meaning Hypothesis (also called the Tinge Hypothesis), a sarcastic remark can mute or attenuate the intended meaning of a message compared to a literal remark (Dews & Winner, 1995). In Dews & Winner's experiments, participants read stories ending with either a sarcastic or literal remark and judged how critical, annoyed, or pleased they thought the speaker was. Sarcastic insults were rated as less critical, and sarcastic compliments were considered less praising.

The enhancement of criticism. In contrast to the Tinge Hypothesis (Dews & Winner, 1995), Colston (1997) found that sarcasm magnifies the negative attitude embedded in a remark. As a potential explanation for the opposing findings from Dews & Winner (1995), Colston pointed out the importance of information added by intonation. The stimuli in Colston (1997) were presented as written text, whereas auditory stimuli were used in the latter two studies. Similarly though, Toplak & Katz (2000) reported that sarcastic remarks enhanced criticism compared to their literal counterparts. Roberts & Kreuz (1994) also found that sarcasm was strongly associated with the discourse goal of "showing negative emotion".

Humor. It has also been proposed that sarcasm can provide a way to be funny, humorous, witty, or clever (Glucksberg, 1995; Colston, 2023, 2021). Dews et al. (1995) showed that sarcastic comments were perceived to be less critical and more humorous than literal comments, as well as effective for maintaining a positive interlocutor relationship. Their findings were consistent across modality (video, audio, and text). Also, a study that analyzed short recordings of natural conver-

sations between college students and their friends found that sarcasm is used to mock or tease the addressee humorously without the intention of seriously criticizing them (Gibbs, 2000). Similarly, in Matthews et al. (2006), participants in an experiment read hypothetical scenarios and indicated which response options (literal or sarcastic) they would choose. The results showed that humor was a factor that encouraged speakers to choose sarcasm.

Mockery. Sarcasm can also be used to mock the addressee. Gibbs (2000) identified the mocking function of sarcasm after analyzing recordings of conversations between friends. Pexman & Olineck (2002) had participants rate different types of contexts and responses in terms of how positive, polite, or mocking they were, and found that sarcastic insults are perceived to be more positive, polite, and also more mocking compared to literal insults.

Face-saving. Face-saving effects of sarcasm were reported in several studies (Glucksberg, 1995; Dews et al., 1995). In a study that especially focused on investigating the face-saving functions of sarcasm, Jorgensen (1996) used participants' recollection of their experiences of making sarcastic remarks. Her results showed that using sarcasm makes the speaker appear less rude and less unfair, hence achieving the face-saving effect.

The communicative functions reported in such works were in large part discovered as a result of comprehension experiments. The participants in those studies would evaluate the associated functions of a given utterance.

2.1.2 This study

The study discussed here tested whether providing contexts that are likely to motivate speakers to use the previously mentioned communicative functions in their response results in an increased use of sarcasm, compared to normal conversation. The experiment was designed to be production-based, so that participants were allowed to provide open responses to given contexts. The assumption was that some of those responses would be sarcastic due to the contexts motivating the use of the communicative functions associated with sarcasm. To restrict the number of tested variables, we limited the interlocutor relation to only one type, in which all contexts involved close friends only.

2.1.3 Method

Materials To devise situations that would likely motivate participants to convey certain attitudes, we turned to an existing dataset of sarcasm, MUStARD (Castro et al., 2019). Through a qualitative analysis, we identified a pattern in the dataset, showing that sarcastic comments are often preceded by situations in which an interlocutor is being silly or annoying, as exemplified in (1):

(1) Leonard: What, Sheldon? What, Sheldon?! What, Sheldon?!
Sheldon: Tell me what you see here.
Leonard: The blunt instrument that will be the focus of my murder trial?

Example (1) taken from MUStARD shows an instance in which the interlocutor (Sheldon) behaves in an annoying manner (hinted by the emotive utterance by Leonard), which is followed by a sarcastic remark by the other interlocutor (Leonard). Similarly, example (2) shows that the interlocutor (Joey) behaves in a silly manner, which is followed by a sarcastic remark by the conversational partner (Monica).

(2) Joey: Okay, done.
Monica: What's *pleh*?
Joey: That's *help* spelled backwards so that helicopters can read it from the air.
Monica: Huh! Whats *dufus* spelled backwards?

Motivated by such examples, we created 16 new *non-neutral* situations in which a close friend was behaving in a silly or annoying manner, and 16 *neutral* situations in which the friend was not behaving in such a way. (3) is an example of a context used in the experiment in which an interlocutor behaves in a non-neutral manner and (4) is an example of a neutral context.

(3) **Non-neutral (silly or annoying):** You are helping Steve move into a new apartment. After an hour, you realize that Steve is only carrying light stuff and you are doing all the heavy lifting. Steve says, "ugh, moving is always so stressful and chaotic..."

(4) **Neutral:** Steve and you are hanging out, and he invites his other friends. You like them because they are fun. Steve says, "I invited some other people, hope you don't mind."

The interlocutor in the contexts was a male friend[2] to increase the likelihood of sarcastic responses[3]. We checked for potential differences in lexical features across contexts (i.e., maximum/mean/minimum lexical frequency, maximum / mean / minimal lexical concreteness, count of content / function words, count of unique words) on sarcasm ratings to ensure that these features do not interfere with our main condition (contextual factors). The two levels of the main condition were matched on all these features.

Participants 60 native English speakers (30 female, $\text{Mean}_{\text{age}} = 34$) were recruited on Prolific[4]. Participants received 8 GBP per hour as compensation.

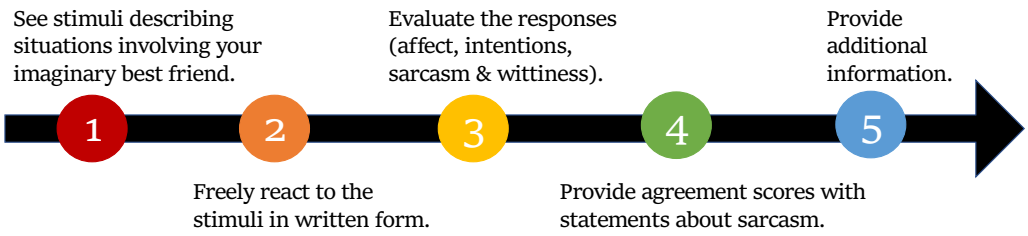

Figure 2.2: Flow of Study 1.

[2]This was based on previous findings that there is an increased likelihood of sarcasm among or directed at male speakers (N. Zhu & Wang, 2020). However, the gender of the directed speaker was not a separate manipulation in our experiment.

[3]Increasing the number of sarcastic responses was an important factor for the design of the experiment, as the responses collected here were planned to be used as data for computational experiments in Chapter 3, and sarcasm is not as common as literal language by default.

[4]`https://www.prolific.co/`. Participants in all subsequent experiments were also recruited on this platform.

Procedure Figure 2.2 shows the procedure of the experiment. The experiment was built using FindingFive[5]. Participants were asked to imagine being the best friend of the imaginary person *Steve* (Step 1 in Figure 2.2). After a practice round with one example situation, the main experiment was carried out in four blocks. In the first block, participants read each situation (N = 32) and provided open responses in any way they wished to respond to their best friend (Step 2). They provided their responses in written form. The situations were shuffled in alternating order (non-neutral and neutral) for all participants. In the second block, participants were referred back to the situations with their responses in the same order as the previous block (Step 3). They answered the following questions on a 1 - 6 Likert scale (*not at all, mostly not, not so much, somewhat, mostly, completely*), in order to report how silly or annoying they found the interlocutor *(affect)*, how sarcastic they thought their response was *(sarcasm)*, and how witty they thought their response was *(wittiness)*. *Affect* (=emotional reaction to the context) was collected because it is the subjective perception of situation that may trigger sarcasm, regardless of the originally intended *context type*. *Wittiness* was collected because prior work has found that it is closely tied to sarcasm, and we assumed that it would especially be the case given the nature of the situations. Participants also indicated all the intentions behind their responses by selecting from eight given options (*intentions*). *Intentions* were collected as they would allow the communicative goals to be tested as intermediate elements that lead speakers to speak sarcastically. The five intention options (I1 - I5 in Table 2.1) were compiled based on the communicative functions reported in the literature. The three control intentions (I6 - I8) were added to the experiment for cases in which sarcasm is not used and therefore participants would want to indicate communicative intents that are not particularly associated with sarcasm.

In the third block, participants indicated their level of agreement to several statements describing sarcasm, also on a 1-6 scale, based on their participation in the experiment (Step 5). Table 2.2 shows the statements provided to the participants. The purpose of this block was to collect participants' understanding of the concept of sarcasm without providing a predefined definition of it[6]. Participants skipped

[5] `https://eu.findingfive.com/`. All subsequent experiments involving human participants were designed using this platform.

[6] Defining sarcasm is tricky in general, demonstrated by the weak consensus of the definition among researchers (See Fox Tree et al. (2020) for details about this). We decided that a reasonable

Table 2.1: Intention options provided to participants with sources from previous work.

	Intentions	Sources
I1	to criticize interlocutor in a harsher way	Colston (1997)
I2	to criticize interlocutor in a softer way	Dews & Winner (1995)
I3	to mock interlocutor in a hilarious way	Glucksberg (1995); Gibbs (2000)
I4	to mock interlocutor in a friendly way	Dews et al. (1995); Jorgensen (1996)
I5	to speak cleverly	Colston (2023); Glucksberg (1995)
I6	to be natural	control intention for non-sarcastic responses
I7	to be direct	control intention for non-sarcastic responses
I8	to be nice	control intention for non-sarcastic responses

this block if they did not provide any sarcastic response throughout the whole experiment. In the last block (Step 6), participants indicated how often they use sarcasm in their daily life (*general sarcasm use*). The experiment lasted 45 minutes on average.

Table 2.2: Statements describing sarcasm ("Sarcasm would..") and the source of each statement from previous work.

No.	Statements	Sources
S1	be saying the opposite of what is intended	Dews & Winner (1995); Kreuz & Glucksberg (1989)
S2	convey messages in a more sophisticated way	Giora et al. (2005)
S3	strengthen the bond between the interlocutors	Dews & Winner (1995)
S4	offend the interlocutor	Toplak & Katz (2000)
S5	be perceived as humorous	Dews et al. (1995)

2.1.4 Analysis

Variable coding Each of the eight intentions was binary coded (0 / 1). The ratings collected on a Likert scale were z-transformed across items for each participant in order to control for the variability in participants' sensitivity (M = 0, SD = 1 for every participant).

Preliminary analyses As our main focus was to examine the effects of various predictors on sarcasm ratings, we first conducted two preliminary analyses to construct a model driven by hypotheses. First, to validate the experimental manipulation, we inspected the relationship between participants' *affect* (degree of silliness-annoyance) and *context type* (*neutral* vs. *non-neutral*). A linear mixed-effects model

approach is to gather participants' ideas about sarcasm post-experiment than force a prescriptive definition, as every speaker is expected to have at least a loose understanding of what sarcasm is.

(lmer; Baayen et al., 2008; Bates et al., 2015) between the two variables showed that *affect* was strongly associated with *context type* ($\beta = 1.00$, $SE = 0.71$, $t = 14.19$, $p < 0.001$), indicating a successful experimental manipulation, as the stimuli that were intended to be non-neutral were indeed perceived as such by the participants. Second, we checked for correlations between *sarcasm* and *wittiness* as we expected participants to consider the two elements similarly given the nature of our experimental settings. A Spearman correlation coefficient suggested a high positive correlation between the two variables ($r = 0.76$, $p < 0.001$), which showed that a response that was rated as highly sarcastic was also likely to be rated as highly witty, supporting prior findings (Dews & Winner, 1995; Gibbs, 2000; Matthews et al., 2006).

Main analysis The main linear mixed-effects (lmer) model had z-scored sarcasm ratings as dependent variable and z-scored *affect* ratings interacting with the eight binary-coded *intentions* as predictors. The model also had main effects for the control variables *order of stimulus presentation* and *general sarcasm use*. The *order of stimulus presentation* was included to capture potential transfer effects from using the same interlocutor across the trials. *General sarcasm use* was included because it was reasonable to assume this might have an effect on the results (e.g. Participants that are typically more sarcastic may be more sarcastic in the experiment as well. On the flip side, these same participants may also have a higher threshold for judging a response to be sarcastic). Due to the strong influence of *context type* on *affect*, *context type* was not included in the model to avoid collinearity issues, and due to high correlation with *sarcasm*, *wittiness* was also excluded from the main model. Gender, the only control variable about the participants, was also excluded from our main model because it did not reach statistical significance.

A by-item random intercept was included. Initially, a by-participant random slope was included for all the main predictors, but the slope was kept only for *affect* to ensure model convergence.

Lmer model formula

[DV] z-sarcasm score $\sim$
[Main predictors] z-affect * (intent1 + ... + intent8) +
[Control predictors] stimuli presentation order + general sarcasm use +

[By-participant random slope] (0 + z-affect|participant) +

[By-item random intercept] (1|item)

2.1.5 Results and discussion

Table 2.3: Results of the statistical modeling. Lmer coefficients predicting self-reported sarcasm ratings (z) with main predictors *affect (z)* and 8 *intentions* in interaction, and control variables.

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Main predictors	(Intercept)	0.08	0.09	0.83	0.405	
	affect	0.24	0.06	3.83	<0.001	***
	criticize harsher	0.08	0.07	1.14	0.256	
	criticize softer	0.05	0.05	0.99	0.324	
	mock hilariously	0.78	0.06	12.49	<0.001	***
	mock friendly	0.76	0.05	15.11	<0.001	***
	speak cleverly	0.22	0.06	3.86	<0.001	***
	be natural	-0.06	0.04	-1.61	0.108	
	be direct	-0.18	0.04	-4.79	<0.001	***
	be nice	-0.22	0.05	-4.22	<0.001	***
Interactions	affect:criticize harsher	-0.14	0.07	-1.95	0.051	
	affect:criticize softer	-0.13	0.06	-2.24	0.025	*
	affect:mock hilariously	-0.18	0.06	-2.82	0.005	**
	affect:mock friendly	-0.06	0.06	-1.12	0.263	
	affect:speak cleverly	-0.08	0.06	-1.38	0.168	
	affect:be natural	0.00	0.04	-0.03	0.974	
	affect:be direct	-0.12	0.04	-3.05	0.002	**
	affect:be nice	-0.07	0.06	-1.26	0.208	
	affect:stimuli order	0.01	0.00	3.03	0.002	**
Control variables	general sarcasm use	-0.03	0.01	-2.49	0.013	*
	stimuli order	0.00	0.00	-0.39	0.694	
Conditional R^2	0.39					
Significance	*: p <0.05, **: p <0.01, ***: p <0.001					

The main model explained 39% of the variance (conditional R^2). We assessed the collinearity among the predictor variables by calculating the Variance Inflation Factors (VIF; Zuur et al., 2010). The VIFs were low to moderate, as all variables were smaller than 4.7 (M = 1.53, SD = 0.84).

Table 2.3 reports the lmer coefficients and statistical significance of the predictors on the z-scored sarcasm ratings. *Affect* was a significant predictor for sarcasm

ratings, with each unit of increased non-neutrality of the context (silly or annoying) leading to 0.24 higher z-scored sarcasm ratings. The intention to *mock* the interlocutor *hilariously*, *mock* the interlocutor *in a friendly way*, and to *speak cleverly* also showed positive and statistically significant main effects on the sarcasm ratings (i.e. having the intention of *mocking* the interlocutor *hilariously* resulted in 0.78 higher z-scored sarcasm ratings than otherwise). On the other hand, intention to *be direct* or *be nice* and *general sarcasm use* had significant negative effects on the sarcasm ratings (i.e., having the intention of *being direct* led to 0.18 lower z-scored sarcasm ratings). No main effects of intention to *criticize harsher*, *criticize softer*, or *be natural* were observed. The *order of stimulus presentation* did not affect sarcasm ratings but showed a mild positive interaction with *affect*, which indicates that in cases of the speaker perceiving the contexts as silly or annoying, the sarcasm ratings were higher in the later part of the experiment.

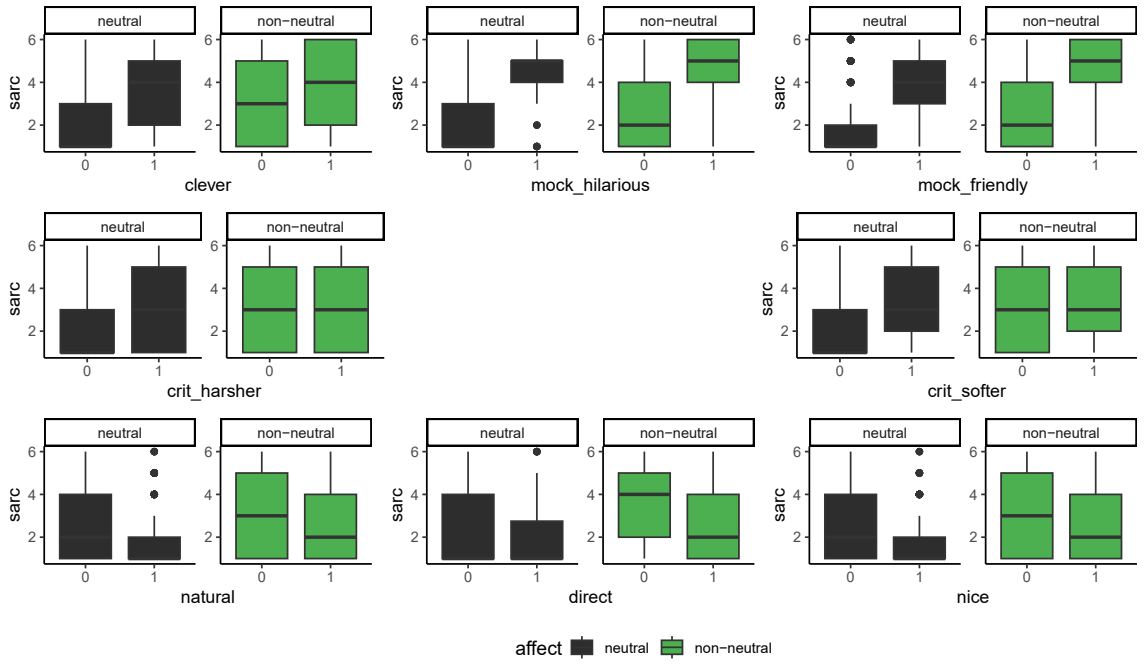

Figure 2.3: Relationship between affect and various communicative intentions in sarcasm ratings.

To summarize, when speakers found the conversational situation or conversation partner to be silly or annoying, they were more likely to speak sarcastically. If speakers had the intention to mock the conversation partner either in a hilarious or friendly way, or to say a clever remark, they were also more likely to use sarcasm. Figure 2.3 visualizes the strong relationship between sarcasm and these intentions.

As for the statements describing sarcasm, 57 out of 60 participants (95%) reported having used sarcasm in the experiment, and gave responses about what they believed describes sarcasm. Around half of these participants agreed with the classical definition of sarcasm (saying the opposite of what is intended). More than half of the participants believed that sarcasm is a sophisticated way of communication, and that it is bonding and humorous. Less than half of the participants believed that sarcasm is offensive.

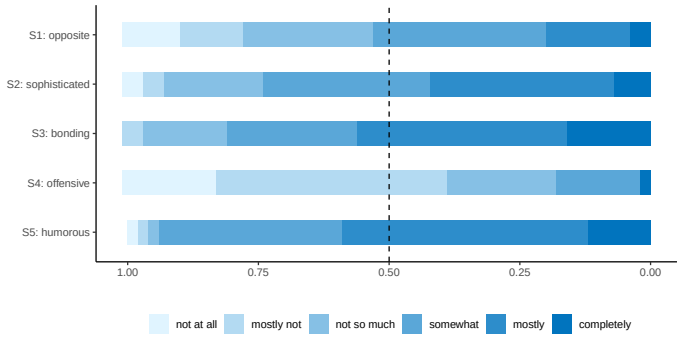

Figure 2.4: Agreement to statements describing sarcasm by participants of Study 1. See Table 2.2 for full descriptions of each statement.

2.2 Study 2: The role of affect and intent for sarcasm perception

2.2.1 Background

When a speaker speaks sarcastically, they expect a listener to understand it (Co-operative Principle; Grice, 1975). The speaker expects the listener to understand that their utterance was intended to be sarcastic, and also to understand why they spoke sarcastically. Though these things often do successfully occur as intended, sometimes certain gaps in understanding emerge from the differing perspectives of speakers and listeners. For this reason, prior work addresses the interlocutor's point-of-view (speaker vs. addressee / observer) as a factor that influences the identification or interpretation of sarcasm. The point-of-view of speakers and listeners can cause differences in how aggressive a sarcastic comment feels, the understanding of why the speaker used sarcasm, and about whether a comment is sarcastic at all.

People conceive of different reasons for sarcasm depending on whether they focus on a speaker's intent or the addressee's reaction (Toplak & Katz, 2000). One explanation for this difference frames it in terms of what each interlocutor considers important: from the speaker's perspective, the *reason* behind a sarcastic remark is more important, whereas from the addressee's perspective, the *effect* that the sarcastic remark may cause is more important (Toplak & Katz, 2000).

Additionally, people evaluate sarcastically critical arguments as more aggressive when taking the perspective of the addressee, but as more humorous when taking the perspective of the speaker (Bowes & Katz, 2011). Similarly, the criticism conveyed in a sarcastic remark is muted only when participants judge the "social impression" behind the sarcastic remark, but not when they judge the "speaker intent" (Pexman & Olineck, 2002).

Furthermore, recognizing the presence of sarcasm in an utterance can also differ between speakers and listeners. In one study, participants in a sarcasm-inducing condition reported identifying more uses of sarcasm in their own utterances compared to those in a non-sarcasm-inducing condition, but they did not report such effects about the utterances of their conversational partners (Fox Tree et al., 2020).

2.2.2 This study

This study identified the degree of alignment between sarcasm producers and evaluators. Specifically, the experiment tested whether the emotional factors triggering a speaker to use sarcasm are interpreted as such by the observer of a conversation. This experiment was an extension of Study 1 (generation experiment) in that new participants rated the responses collected in Study 1 as external observers.

2.2.3 Method

Materials The same situational prompts and the responses collected from Study 1 were used (See Figure 2.5). Obvious spelling errors from the responses were corrected and the subject *you* was modified to *John*, so that participants could read conversations between two conversation partners *Steve* and *John*.

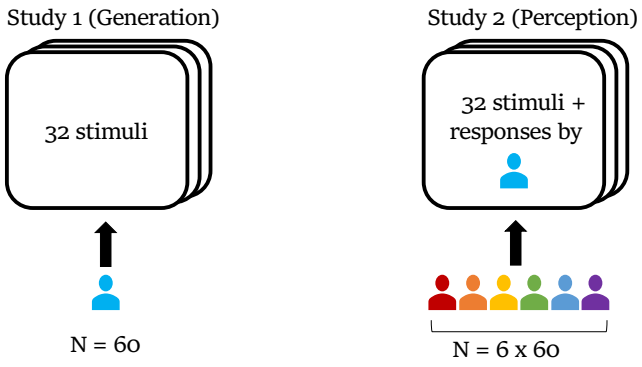

Figure 2.5: Stimuli connection between Study 1 and Study 2

Participants A new group of 360 participants (180 female, $\text{Mean}_{\text{age}} = 35$) was recruited with the same inclusion criteria as in Study 1. Participants received 8 GBP per hour as compensation. In order to account for the subjective nature of sarcasm perception in the generalization of the results (Fox Tree et al., 2020), we assigned six evaluators to each previous participant. The total number of participants in this experiment was therefore 360 (6 observers * 60 speakers). The participants assumed the role of an observer rather than the addressee, based on Dews et al. (1995), who reported that the patterns for sarcasm perception were identical between participants who took on the role of the addressee or that of a third-party.

We expected it to be more natural for the participants to be observers of the conversation than be the interlocutors in a conversation in which they never participated before.

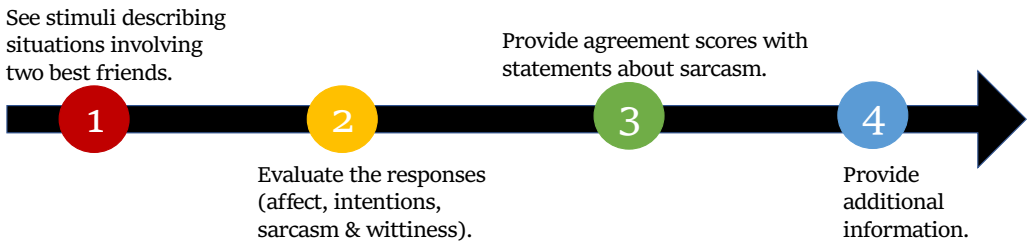

Figure 2.6: Flow of Study 2.

Procedure The experiment was conducted in three blocks. In the first block, participants saw conversations between *Steve* and *John* and answered the same 4 questions about the responses as in the Study 1 – how silly or annoying *John* found the interlocutor *(affect)*, how sarcastic *John*'s response was *(sarcasm)*, how witty *John*'s response was *(wittiness)*, and *John's* presumed communicative intent (*intentions*). In the second and third blocks, identical to Study 1, participants rated their agreement on the same statements describing sarcasm (See Table 2.2 in Section 2.1.3) and their *general sarcasm use*. The experiment lasted 30 minutes on average.

2.2.4 Analysis

Variable coding Similarly to Study 1, all the ratings (*affect*, *sarcasm*, and *wittiness*) were z-scored by participant for baseline control. The z-scored ratings provided by each group of six observers who evaluated the same conversations were averaged. Each of the eight intentions were binary coded (0 / 1) and averaged across the six observers.

Preliminary analyses As previously, we checked for the relationship between participants' *affect* (degree of silliness-annoyance) and *context type* (*neutral* vs. *non-neutral*). An lmer model showed that there was a strong effect of *context type* on *affect* for the observers as well ($\beta = 1.06$, $SE = 0.24$, $t = 4.37$, $p < 0.001$). Likewise, a Spearman correlation coefficient between *sarcasm* and *wittiness* once again suggested a high correlation between the two variables ($r = 0.69$, $p < 0.001$). The control predictor *gender* did not reach statistical significance.

Main analysis The main lmer model had the same structure as previously, except for the by-item random effect structure, such that a random intercept for each item nested within the stimuli set assigned to each group of six observers was included in the model (nested random structure).

Lmer model formula

[DV] z-sarcasm score ∼
[Main predictors] z-affect * (intent1 + ... intent8) +
[Control predictors] stimuli presentation order + general sarcasm use +
[By-participant random slope] (0 + affect| participant) +
[By-item random intercept] (1|item:group)

Supplementary analysis We also conducted a combined lmer analysis to statistically compare the different interlocutor perspectives in Study 1 (speakers) and Study 2 (observers).

[DV] z-sarcasm score ∼
[Main predictors] (z-affect + intentions) * experiment source
[By-participant random slope] (1|participant) +
[By-item random intercept] (1|item:experiment source)

2.2.5 Results and discussion

Main analysis The main model explained 47% of the variance. We observed no collinearity problems among the predictor variables as the VIFs for all variables were smaller than 3.9 (M = 1.49, SD = 0.66).

Table 2.6 reports the lmer coefficients and statistical significance of the predictors on the z-scored sarcasm ratings. For the observers also, *affect* was a significant

Table 2.4: Results of the statistical modeling. Lmer coefficients predicting other-reported sarcasm ratings (z) with main predictors *affect (z)* and *intentions* in interaction, and control variables.

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Main predictors	(Intercept)	-0.03	0.03	-0.99	0.323	
	affect	0.36	0.02	14.96	<0.001	***
	criticize harsher	0.08	0.03	2.55	0.011	*
	criticize softer	0.13	0.02	7.07	<0.001	***
	mock hilariously	0.51	0.02	20.73	<0.001	***
	mock friendly	0.50	0.02	24.95	<0.001	***
	speak cleverly	0.31	0.02	13.35	<0.001	***
	be natural	-0.15	0.02	-9.22	<0.001	***
	be direct	-0.23	0.01	-15.58	<0.001	***
	be nice	-0.16	0.02	-7.28	<0.001	***
Interactions	affect:criticize harsher	-0.09	0.03	-2.98	0.003	**
	affect:criticize softer	-0.03	0.02	-1.43	0.153	
	affect:mock hilariously	-0.15	0.03	-5.29	<0.001	***
	affect:mock friendly	-0.12	0.02	-5.18	<0.001	***
	affect:speak cleverly	-0.04	0.03	-1.44	0.150	
	affect:be natural	0.00	0.02	0.08	0.933	
	affect:be direct	-0.11	0.02	-6.58	<0.001	***
	affect:be nice	0.01	0.02	0.35	0.725	
	affect:stimuli order	0.00	0.00	-0.56	0.573	
Control variables	general sarcasm use	-0.01	0.01	-1.43	0.154	
	stimuli order	0.00	0.00	1.29	0.198	
Conditional R^2	0.47					
Significance	*: p <0.05, **: p <0.01, ***: p <0.001					

predictor for sarcasm ratings, as higher perception of non-neutrality in the contexts led to more instances of higher sarcasm ratings in the following responses. Furthermore, when the observers thought that the speakers intended to *mock* the addressee either *hilariously* or *in a friendly way*, or *speak cleverly*, they rated the speakers' responses as more sarcastic. On the other hand, when the observers thought that the speakers intended to be *direct* or *nice,* they tended to give lower sarcasm ratings to the speakers' responses (See Figure 2.7).

On the statements describing sarcasm as well, similarities were observed between Study 1 and Study 2. Out of 360 participants, 335 of them (93%) reported that they had witnessed sarcasm in the experiment and gave responses on what they believed describes sarcasm. Figure 2.8 shows the proportions of participant

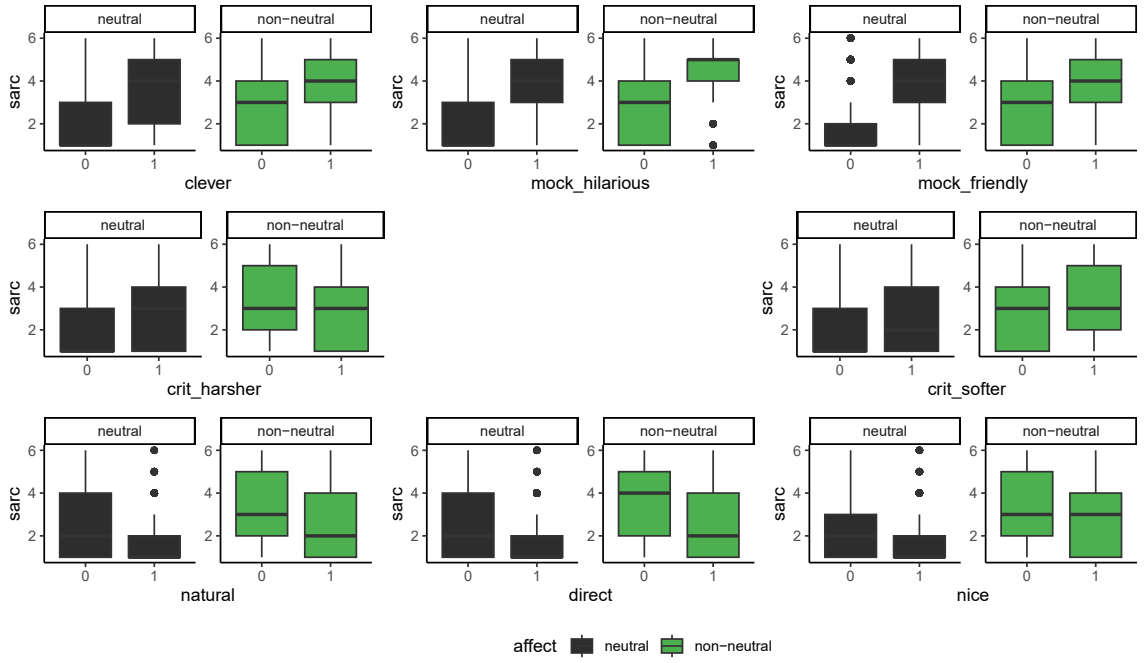

Figure 2.7: Relationship between affect and various communicative intentions in sarcasm ratings.

responses for Study 1 (left) and the current study (right), from *not at all* (1) to *completely* (6). The results show that participants evaluated sarcasm to be a positive communicative tool (See the bars for S2, S3, S5 in Figure 2.8) whether they assumed the role of a speaker or observer. These results support the humor effect argued by Dews et al. (1995), the bond-enhancement effect argued by Dews & Winner (1995), but not the "harsher criticism" effect argued by Toplak & Katz (2000) or Colston (1997).

Nevertheless, there were some differences between speakers and observers. First, only the observers gave higher sarcasm ratings to a response when they indicated that the speaker had the intention to *criticize* the interlocutor either more *harshly* or *softly*. Also, only the observers gave lower sarcasm ratings to a response when they indicated that the speakers wanted to behave *naturally*. None of these communicative intentions affected the speakers in the previous experiment. Finally, even though the majority of participants assuming the observer's role viewed sarcasm as a positive communicative tool, the proportions of such participants were significantly smaller compared to those that assumed the speaker's role. T-test results showed significant difference between the two groups for S1 (*opposite*: $t = -5.65$, $p < 0.001$) and S4 (*offensive*: $t = -4.60$, $p < 0.001$), indicating that more observers

agreed with them than speakers did, and for S3 (*bonding*: $t = 5.40$, $p < 0.001$) and S5 (*humorous*: $t = 4.61$, $p < 0.001$), meaning that more speakers agreed with them than observers did.

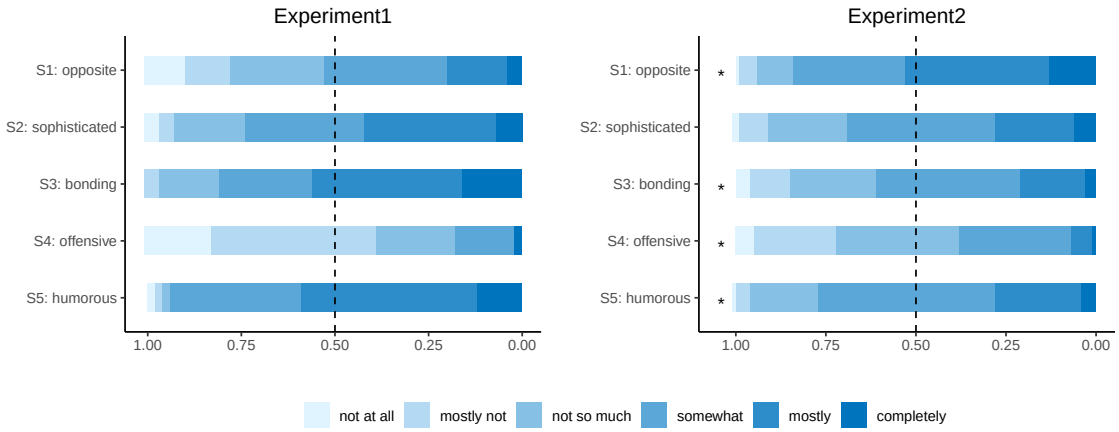

Figure 2.8: Agreement to statements describing sarcasm by speakers (Study 1) and observers (Study 2). Significant differences between the two studies are indicated with asterisks on the right side. See Table 2.2 in Section 2.1.3 for full descriptions of each statement.

Supplementary analysis Only four of the reported differences between self-reported sarcasm scores and other-reported sarcasm scores were statistically significant. We observed positive significant interactions between *experiment source* and *affect*, indicating that observers associate the perceived non-neutrality of a situation with sarcasm more strongly than speakers. Similarly, when observers believed that the speaker wanted to *cricitize* the addressee *softly* or make a *clever* remark, they rated the speaker's responses as even more sarcastic than the speakers themselves would in the same situation. An interaction in the opposite direction was observed for the intention to *be direct*, such that when observers thought that the speaker intended to be direct, they were more likely to rate the speaker's response as unsarcastic more so than the speakers would themselves. The results point to the general tendency among observers to make a stronger connection between communicative intent or emotional reactions and sarcasm perception than speakers do. We discuss the reasons for this in the general discussion (Section 2.6).

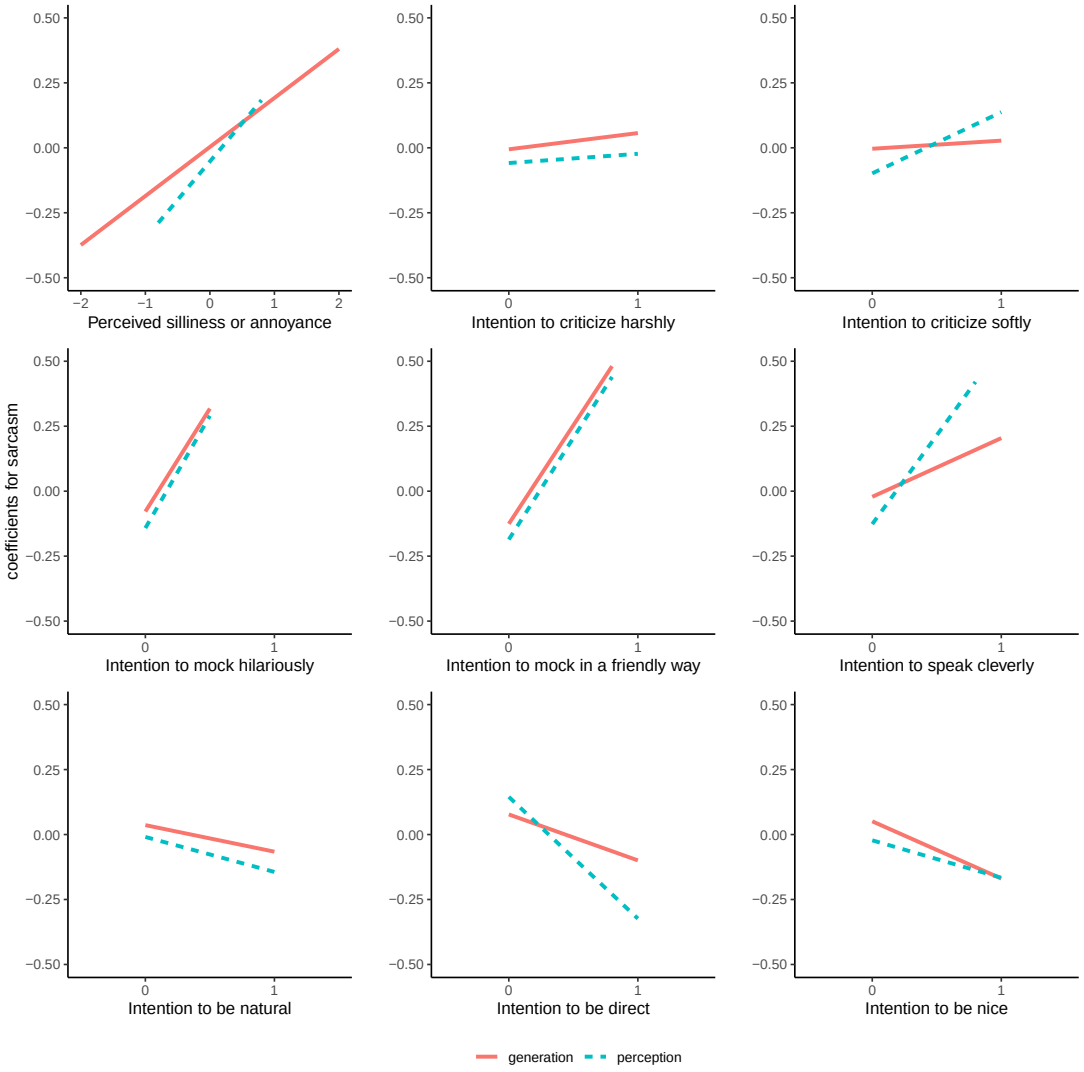

Figure 2.9: Comparison between speaker and observer perspectives. Solid red lines represent the relation between each intention and sarcasm ratings from the speakers' perspectives (Study 1). Discontinued blue lines represent the same information from the observers' perspectives (Study 2).

Table 2.5: Interaction between Study 1 and Study 2. Lmer coefficients predicting sarcasm ratings with main predictors *affect* and *intentions* interacting with *experiment source*. Only interactions are reported, with significant interactions in bold.

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Interactions	**study2:affect**	0.11	0.04	2.60	0.009	**
	study2:criticize harsher	-0.03	0.12	-0.23	0.818	
	study2:criticize softer	0.20	0.10	2.16	0.031	*
	study2:mock hilariously	0.07	0.12	0.58	0.561	
	study2:mock friendly	0.02	0.10	0.24	0.810	
	study2:speak cleverly	0.46	0.12	4.00	<0.001	***
	study2:be natural	-0.03	0.08	-0.39	0.695	
	study2:be direct	-0.29	0.07	-3.98	<0.001	***
	study2:be nice	0.08	0.09	0.88	0.380	
Conditional R^2	0.47					
Significance	*: p <0.05, **: p <0.01, ***: p <0.001					

2.3 Study 3: Types of contexts that trigger sarcasm

2.3.1 Background

Study 1 examined the connection between affect and speakers' motivation to speak sarcastically by using contexts as triggering prompts (see Section 2.1). We found multiple factors, i.e., emotional reactions and communicative intent, to be associated with increased use of sarcasm. A question that has been left unanswered relates to the connection between characteristics of sarcasm-triggering contexts, and the increased use of sarcasm. Also, Study 1 was limited to relationships with interlocutors that are close friends, which may reduce the generalizability of its findings.

2.3.2 This study

This further study therefore examined the direct connection between contextual characteristics and sarcasm use by expanding Study 1 and applying a slight shift in focus, identifying the characteristics of different contexts and their potential to motivate sarcasm. Additionally, this study also examined the influence of different interlocutor relationships (close vs. distant) on these contextual effects. This allowed for the generalization of the factors facilitating the use of sarcasm between close friends to interlocutors of a more distant relationship.

2.3.3 Method

Materials New situations were created based on the findings from Study 1 and Study 2. Of the 32 situational prompts used in them, we selected the top half (N = 16) with the highest average sarcasm ratings either by speakers or observers. We conducted a qualitative analysis to find commonalities among situations featuring silly or annoying behavior of the interlocutor. From this analysis, four categories (*context types*) emerged: Situations in which an interlocutor is behaving silly (*silly*), not seeing their own flaws (*flaw-blind*), talking about uninteresting topics (*uninteresting*), or behaving in an entitled or demanding manner (*entitled-demanding*). We also made a control category, characterized by unobtrusive and standard behavior of the interlocutor (*neutral*). We created eight new situations describing each category. The situations were constructed in a way that either a best friend or a col-

league could be involved in them[7]. Examples (5) - (9) are example stimuli belonging to each category.

(5) **Silly**: You are having a small party at your house. Steve, a little tipsy, starts mixing ketchup, mustard, potato chips, and orange juice and says "hey, look, I made something delicious!"

(6) **Flaw-blind:** Quite often, Steve's boss asks Steve to run errands that are possibly unrelated to work. A few days ago, your boss asked you to run a personal errand for him, to which you reluctantly said yes. When you complain to Steve about this, he says, "wow, that doesn't sound good. Why did you say yes?"

(7) **Uninteresting:** Steve recently inherited a fancy watch from his dead grandfather. And he keeps showing it to John and talking about it. For the fifth time, Steve says, "look, look how shiny this watch is."

(8) **Entitled-demanding:** Steve wants to take a cab to go home. He asks John to call a taxi for him. When John says "There are several cabs right there. Why not just get one of those?" Steve says "but you have an app for calling cabs. I know it gives you a discount!"

(9) **Neutral:** Steve asks John "can I borrow your USB stick?" and John gives it to him. Steve says, "thank you."

Context type was one of our primary experimental conditions. To ensure that the four context types were not overlapping, we recruited 25 English native speakers in a pilot study. They read the situations and classified them (forced choice) as one of the five context types (*silly, flaw-blind, uninteresting, entitled-demanding, neutral*). Responses were considered as correct when the participant chose the intended category. We also regarded the neutral category as acceptable as we expected indi-

[7]Though we considered testing a relationship type that is more distant than a colleague, it was extremely difficult to create situations that could involve two strangers that would exemplify the mentioned categories. Therefore, we considered colleagues as the distant relationship.

vidual variability in judging what is considered standard behavior. Situations with low correctness scores were modified to represent the intended category better. After five rounds of pilot tests (each time with different sets of 25 participants), the average accuracy for all situations was 95.75 $\pm$ 5.98 %. No effects on the sarcasm ratings from lexical features of the situation descriptions were found.

Participants 128 new native English-speaking participants (64 female, Mean_{age} = 37) were recruited with the same inclusion criteria as in the previous studies. Participants received 9 GBP per hour as compensation.

Procedure The experiment was conducted in six blocks. The relationship type (*best friend* vs. *colleague*) was manipulated within participants but between items and between blocks (first block: best friend, second block: colleague). As the order of stimulus presentation did not have an effect on sarcasm ratings in the previous experiments, to not further complicate the implementation process, the relationship type was not counterbalanced across blocks across participants[8].

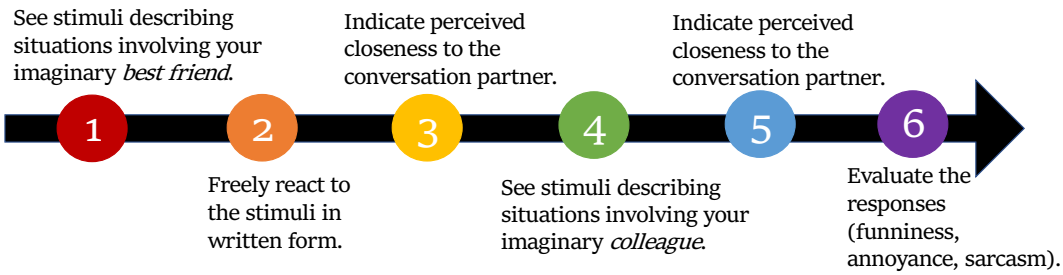

Figure 2.10: Flow of Study 3.

As shown in Figure 2.10, the general procedure was similar to Study 1. Participants responded to situation descriptions in the first two blocks (best friend and colleague) and provided evaluation ratings on a 1-6 scale for all the situations and

[8]We acknowledge a potential confound effect for relationship type.

responses in two subsequent blocks. Immediately after each response block, participants indicated how close they felt to the interlocutor (*closeness perception*). The evaluation questions asked in this experiment were: How sarcastic their response was *(sarcasm)*, how funny they found the interlocutor *(funniness perception)*, and how annoying they found the interlocutor *(annoyance perception)*. *Funniness* and *annoyance* perception were included to strengthen the findings of Study 1 (i.e., the interlocutor being silly or annoying triggers higher sarcasm ratings) and to link the characteristics of a situation with the intended communicative functions they trigger. For a random subset of the situations (20%), consisting of one situation for each condition (context type $\times$ relationship type), participants indicated how much they wanted to mock the interlocutor *(mock)* or speak cleverly *(clever)*, which were the intentions that had the strongest effects (i.e., highest regression coefficients) for higher sarcasm ratings in Study 1. The experiment lasted 34 minutes on average.

2.3.4 Analysis

Variable coding Data treatment was identical to that in Study 1.

Preliminary analysis To assess the success of the textual manipulation of *relationship type*, we ran an lmer model between *closeness perception* and *relationship type*. The model indicated a significant association between the two variables ($\beta = -0.83$, $t = -56.30$, $p < 0.001$), indicating that participants felt closer to the addressee as best friend than as colleague (manipulation successful).

Main analysis The goal of the main analysis was to test the effects of context type and relationship type on sarcasm scores. The initial lmer model had the z-scored sarcasm ratings as dependent variables and and z-scored ratings for *funniness perception* and *annoyance perception* interacting with *context type* and *relationship type*, and additionally, the same predictors interacting with *order of stimulus presentation*. Random effect structures were similar to those of Study 1: A by-item random intercept and a by-participant random slope for the continuous predictors varying within participant (*funniness perception* and *annoyance perception*) were included (see Table 2.6).

There were no 3-way interactions between *funniness perception*, *context type*, and

relationship type ($p = 0.56$) or between *annoyance perception*, *context type*, and *relationship type* ($p = 0.68$). *Order of stimulus presentation* was further removed as it had no main effect ($p > 0.25$) and did not interact with the other factors ($p > 0.10$).

Thus, the final model had *funniness perception* and *annoyance perception* as predictors interacting with *context type* and *relationship type* separately. The random effect structures remained the same.

Main lmer model

[DV] z-sarcasm score $\sim$
[Main predictors] (z-funniness + z-annoyance) *
(situation type + interlocutor relationship type) +
[By-participant random slope] (0 + z-funniness + z-annoyance|participant) +
[By-item random intercept] (1|item)

Supplementary analysis The ratings collected on the subset (*mock* and *clever*) were analyzed in a separate model. The analysis was conducted to replicate the results from Study 1 that intention to mock the addressee and to speak cleverly were very strongly associated with higher sarcasm ratings in the following utterance. We ran a linear mixed effects model between the z-scored sarcasm ratings and z-scored intentions with a by-item random intercept and a by-subject random slope for both intentions.

Supplementary lmer model

[DV] z-sarcasm score $\sim$
[Main predictors] (z-mock + z-clever)
[By-participant random slope] (0 + z-mock + z-clever|participant) +
[By-item random intercept] (1|item)

2.3.5 Results and discussion

Main analysis The main model explained 37% of the variance. The VIFs suggested a moderate collinearity among variables, with the highest VIF being 5.87 (M = 2.37, SD = 1.82).[9] Table 2.6 reports the lmer coefficients and statistical

[9]We identified a variable with the highest VIF of 5.86 (*annoyance perception*) and ran a simpler model without it. As the directions of all the effects stayed the same and we considered *annoyance perception* an important predictor, we kept the original (more complex) model.

Table 2.6: Results of the statistical modeling. Lmer coefficients predicting self-reported sarcasm ratings by participants with main predictors funniness perception (z), annoyance perception (z), context types (*neutral*, *silly*, *flaw-blind*, *uninteresting*, *entitled-demanding*) and relationship types in interaction.

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Main predictors	Intercept (neutral)	-0.26	0.11	-2.41	0.016	*
	silly	0.50	0.13	3.80	<0.001	***
	flaw-blind	0.68	0.13	5.21	<0.001	***
	uninteresting	0.30	0.13	2.34	0.019	*
	entitled-demanding	0.05	0.13	0.37	0.710	
	funniness	0.07	0.04	1.96	0.05	
	annoyance	0.38	0.05	7.29	<0.001	***
	relationship type	-0.03	0.07	-0.47	0.637	
Interactions	funniness:silly	0.10	0.04	2.17	0.030	*
	funniness:flaw-blind	0.25	0.05	4.50	<0.001	***
	funniness:uninteresting	0.20	0.05	3.74	<0.001	***
	funniness:entitled-demanding	0.17	0.06	2.79	0.005	**
	annoyance:silly	-0.01	0.06	-0.23	0.820	
	annoyance:flaw-blind	-0.06	0.07	-0.96	0.339	
	annoyance:uninteresting	-0.04	0.06	-0.65	0.517	
	annoyance:entitled-demanding	-0.15	0.06	-2.35	0.019	*
	funniness:relationship type	-0.03	0.03	-0.99	0.324	
	annoyance:relationship type	-0.05	0.04	-1.29	0.198	
Conditional R^2	0.37					
Significance	*: p <0.05, **: p <0.01, ***: p <0.001					

significance of the predictors on the z-scored sarcasm ratings. The results showed that *silly*, *flaw-blind*, and *uninteresting* contexts triggered speakers to respond with higher levels of sarcasm compared to *neutral* contexts. The *entitled-demanding* context did not affect the use of sarcasm by itself but interacted with *funniness perception* (positive) and *annoyance perception* (negative) in triggering sarcasm. *Annoyance perception* triggered higher number of sarcastic responses. *Funniness perception*, though not a significant factor on its own, interacted with *silly*, *flaw-blind*, and *uninteresting* contexts in causing higher sarcasm ratings. In contrast, *relationship type* did not affect sarcasm ratings nor did it interact with perceptions of situations (*funniness* or *annoyance*).

Supplementary analysis We confirmed that the same communicative intent found to encourage sarcasm in a close relationship has the same effects in a more distant

interlocutor relationship. Results indicated that both intentions strongly trigger a higher level of sarcasm, confirming the results from Study 1 (See Table 2.7 for comparison).

Table 2.7: Lmer coefficients predicting sarcasm ratings with main predictors of intent to *mock* and *speak cleverly* from a subset of stimuli in the current experiment (top). Lmer coefficients for *mock* and *speak cleverly* from the main model of Study 1 repeated here for comparison (bottom).

	Predictors	β	*SE*	*t*	*p*
Study 3	(Intercept)	0.17	0.04	4.73	***
	mock	0.31	0.04	8.88	***
	speak cleverly	0.31	0.03	9.09	***
Study 1	(Intercept)	0.08	0.09	0.83	
	mock hilarious	0.78	0.06	12.49	***
	mock friendly	0.76	0.05	15.11	***
	speak cleverly	0.22	0.06	3.86	***
*: p <0.05, **: p <0.01, ***: p <0.001					

Figure 2.11 suggests that in most contexts, the intent to mock or speak cleverly results in higher sarcasm ratings consistently. The context in which the interlocutor is behaving in a demanding or entitled manner shows a more complex pattern, that is, having such communicative intentions may trigger more sarcastic utterances with a higher probability in general, but there is more variability compared to other types of contexts. This aligns with the results from the main model that *entitled-demanding* situations interacted with perceptions of them (funny or annoying) in the opposite directions, suggesting that these types of situations bring about more complex effects on the use of sarcasm depending on the reactions to them, or communicative intentions.

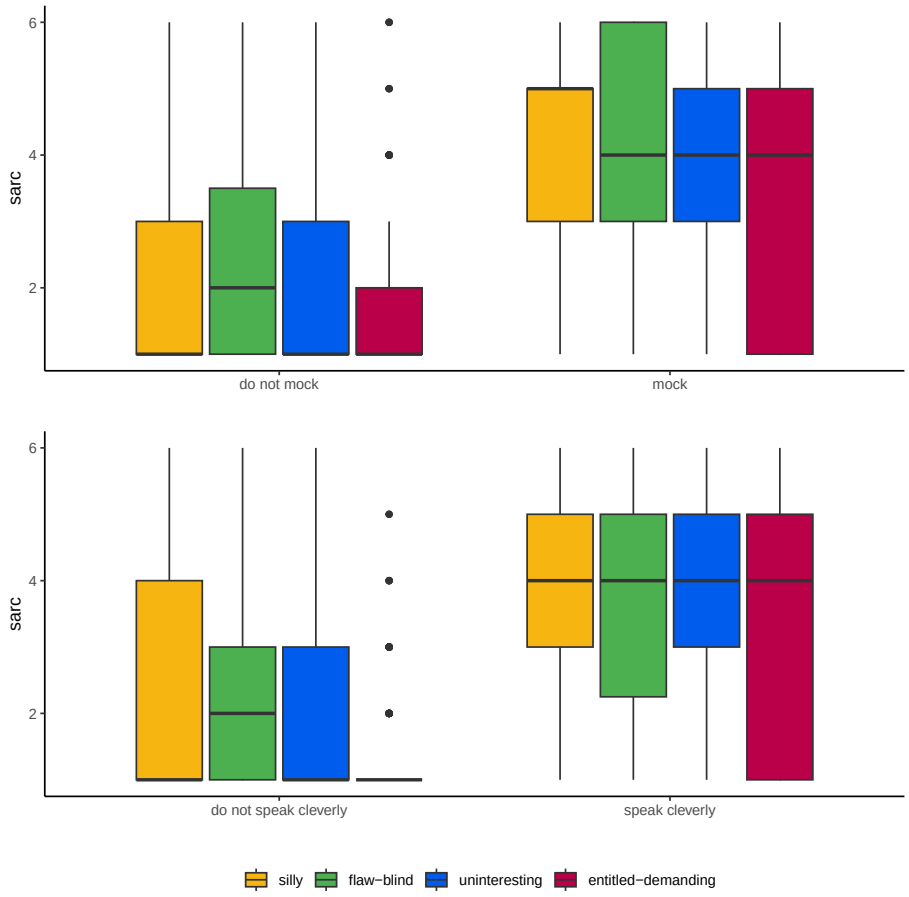

Figure 2.11: Relations between intent to mock / speak cleverly and sarcasm ratings by context type.

2.4 Study 4: Types of contexts that help interpret sarcasm

2.4.1 Background

Study 2 examined the relation between speakers' intended sarcasm and observers' perceived sarcasm by drawing comparisons with Study 1 (see Sections 2.1 and 2.2). We found that observers can generally decode the underlying intent and the emotional reactions of the speakers, but that they also show some differences from speakers.

2.4.2 This study

In parallel to Study 2, we tested the relation between speakers' intended sarcasm and observers' perceived sarcasm on the stimuli used in Study 3 (see Section 2.3). This study is a replication of Study 2 that strengthens its findings about whether the factors that motivate a sarcastic remark are also transmitted to the observers. Similarly to Study 2, a new round of participants evaluated the responses collected in Study 3 regarding sarcasm, presumed emotional reactions, and intentions.

2.4.3 Method

Materials The same materials with the responses from Study 3 were used. The same preprocessing methods were applied as in Study 2.

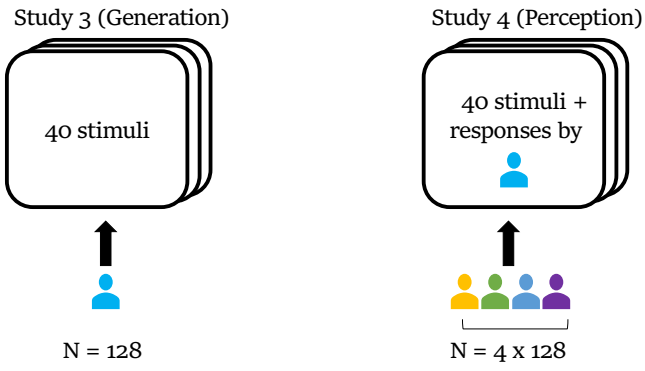

Figure 2.12: Stimuli connection between Study 3 and Study 4.

Participants A total of 5,120 participants (2,560 female, $\text{Mean}_{\text{age}} = 38$) were recruited with the same inclusion criteria. Four evaluators were assigned to the responses by each participant in Study 3 ($N_{\text{eval}} = 4 * 128$). Participants received 9 GBP per hour.

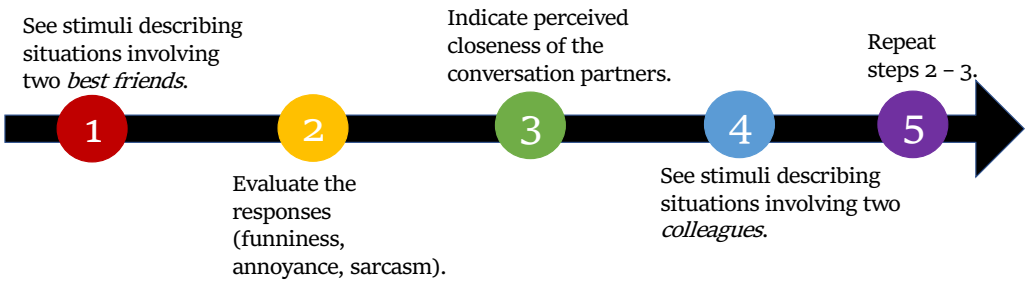

Figure 2.13: Flow of Study 4.

Procedure The experiment was conducted in five blocks (see Figure 2.13). Participants saw conversations between two close friends (*Steve* and *John*) and evaluated John' responses on the same evaluation questions as in Study 2 – how funny John found Steve *(funniness perception)*, how annoying John found Steve *(annoyance perception)*, and how sarcastic *John*'s response was *(sarcasm)*. Additionally, for random 20% of stimuli, participants rated how much John intended to mock Steve (*mock*) or speak cleverly (*clever*). At the end of the block, participants rated the presumed closeness of the two conversation partners. In the next block, participants repeated the same process with conversations between two colleagues. The experiment lasted 30 minutes on average.

2.4.4 Analysis

Variable coding Data treatment was done in the same way as Study 3, and the ratings were averaged in the same way as in Study 2.

Preliminary analysis As in Study 3, we tested if the textual manipulation of *relationship type* was successful by running an lmer model between *closeness perception* and *relationship type*. There was a significant association between the two variables ($\beta = -1.04$, $t = -139.10$, $p < 0.001$), indicating that participants perceived the two friends to be closer than the two colleagues (manipulation successful).

Main analysis The goal of the main analysis was to test the effects of context type and relationship type on sarcasm scores given by the observers. The main lmer model had the same main effect structure as in Study 3 with the same random effect structure as in Study 2.

Lmer model formula

[DV] z-sarcasm score $\sim$
[Main predictors] (z-funniness + z-annoyance) *
(situation type + interlocutor relationship type) +
[By-participant random slope] (0 + z-funniness + z-annoyance|participant) +
[By-item random intercept] (1|item:group)

Supplementary analysis The ratings collected on the subset (*mock* and *clever*) were analyzed in a separate model as in Study 3. We ran an lmer model between the z-scored sarcasm ratings and z-scored intentions with a by-item random intercept and a by-subject random slope for both intentions.

Supplementary lmer model

[DV] z-sarcasm score $\sim$
[Main predictors] (z-mock + z-clever)
[By-participant random slope] (0 + z-mock + z-clever|participant) +
[By-item random intercept] (1|item)

2.4.5 Results and discussion

The main model explained 50% of the variance. The VIFs suggested moderate collinearity among the predictors, as the scores for all variables were lower than 5.55 (M = 2.50, SD = 0.95). Table 2.8 reports the lmer coefficients and statistical significance of the predictors on the z-scored sarcasm ratings.

Table 2.8: Results of the statistical modeling. Lmer coefficients predicting other-reported sarcasm ratings with main predictors funniness perception (z), annoyance perception (z), context types (*neutral*, *silly*, *flaw-blind*, *uninteresting*, *entitled-demanding*) and relationship types in interaction.

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Main predictors	Intercept (neutral)	-0.07	0.03	-2.83	<0.01	**
	silly	0.13	0.03	3.91	<0.001	***
	flaw-blind	0.45	0.03	14.58	<0.001	***
	uninteresting	0.18	0.03	5.78	<0.001	***
	entitled-demanding	-0.11	0.03	-3.34	<0.001	***
	funniness	0.22	0.02	12.28	<0.001	***
	annoyance	0.44	0.02	20.50	<0.001	***
	relationship type	-0.01	0.02	-0.64	0.525	
Interactions	silly:funniness	0.00	0.02	0.00	0.998	
	flaw-blind:funniness	0.08	0.02	3.57	<0.001	***
	uninteresting:funniness	0.06	0.02	2.75	<0.01	**
	entitled-demanding:funniness	0.16	0.02	6.32	<0.001	***
	silly:annoyance	-0.20	0.03	-8.10	<0.001	***
	flaw-blind:annoyance	-0.13	0.03	-5.24	<0.001	***
	uninteresting:annoyance	-0.14	0.03	-5.61	<0.001	***
	entitled-demanding:annoyance	-0.25	0.02	-10.16	<0.001	***
	funniness:relationship type	0.00	0.01	-0.02	0.981	
	annoyance:relationship type	0.02	0.01	1.30	0.195	
Conditional R^2	0.50					
Significance	*: p <0.05, **: p <0.01, ***: p <0.001					

Main analysis Similarly to Study 3, the presumed perception of annoyance predicted higher sarcasm evaluation. Also, the same types of contexts that speakers reacted more sarcastically to – *silly*, *flaw-blind*, and *uninteresting* – also led observers to judge the responses as more sarcastic than neutral contexts. *Relationship type* did not affect sarcasm ratings evaluated by observers or interact with perception of situations either.

However, unlike in Study 3, *funniness* perception was positively correlated with sarcasm ratings by observers, whereas speakers in Study 3 only showed interaction effects between funniness perception and non-neutral types of contexts. *Annoyance* perception interacted with all situation types in lowering the sarcasm ratings judged by observers.

In general, the types of contexts that triggered speakers to be more sarcastic than usual were also identified by observers to have triggered higher levels of sarcasm. However, there was some difference between speakers and observers in terms of how each type of context interacts with the emotional reactions to them.

Compared to the speakers, the observers showed stronger effects on the *entitled-demanding* type of situations than the speakers did, such that the observers made an assumption that the speakers would be more sarcastic in these situations, whereas the speakers simply showed an interaction effect (positive if they found them funny and negative if they found them annoying) between this type of situation and their own emotion, to which they had direct access.

Supplementary analysis Identical to the results from all previous studies, the perceived intent to mock or speak cleverly was strongly associated with higher frequency of sarcasm use. This along with the results from the previous experiments concludes that the two intentions are consistently associated with the production and interpretation of sarcasm.

Table 2.9: Lmer coefficients predicting sarcasm ratings with the main predictors of intent to *mock* and *speak cleverly* from a subset of stimuli in Study 4.

Predictors	β	*SE*	*t*	*p*
(Intercept)	0.17	0.02	10.53	***
mock	0.29	0.02	13.83	***
speak cleverly	0.20	0.02	9.43	***
*: p <0.05, **: p <0.01, ***: p <0.001				

Figure 2.14 shows that when the observers think that the speakers had the intent to mock or speak cleverly, they were more likely to give a higher sarcasm rating to the following utterance regardless of the type of situation. This pattern is similar to that in Study 3, which reflects speakers' perspectives, but the pattern for the *entitled-demanding* situation is more clear cut for the observers than for the speakers. Taken together, speakers' emotional reactions and their communicative intent arise through an introspective process, whereas the observers rely on the external cues and their own assumptions to interpret them, thus showing some discrepancy from the speakers at times. We further discuss this topic in Section 2.6 and Section 3.2.

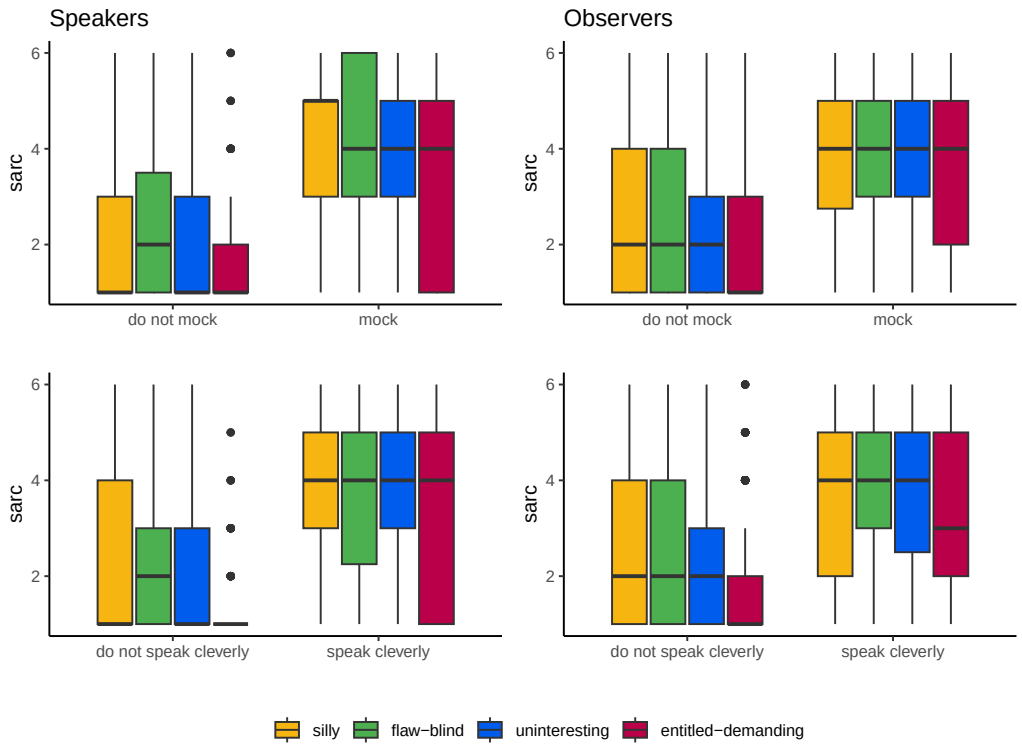

Figure 2.14: Relations between perceived intent to mock / speak cleverly and sarcasm ratings from Study 3 (speakers) and Study 4 (observers).

2.5 Study 5: The effects of other modalities in sarcasm production

2.5.1 Background

So far, all the previous experiments were carried out in text-only setups in which participants read textual prompts and responded in the textual form. However, in real-life communication, information is conveyed through auditory and visual channels as well as textual modality. There are several studies investigating the role of auditory and visual cues in the transmission and identification of sarcasm (Attardo et al., 2003; Bromberek-Dyzman et al., 2021; Caucci & Kreuz, 2012; Li et al., 2024; Woodland & Voyer, 2011). These studies focus on the comprehension of communication involving sarcasm, whereas prior work focused on the generation of sarcasm through the influence of information coming from multiple modalities is scarce (Tabacaru & Lemmens, 2014).

Investigating the effects of multimodal factors such as auditory information, facial expressions, and gestures in triggering sarcasm would allow us to pinpoint the source of the trigger – whether it is exclusively in the content of the context, or in the additional information conveyed through auditory and visual channels. This would strengthen the findings from the previous studies reported in Sections 2.1 $\sim$ 2.4 that certain emotional reactions to preceding contexts likely trigger sarcasm, as it would take into account the contributions of other factors that are present in real communication (Hancock, 2004).

2.5.2 This study

To that end, we conducted a pilot study replicating Study 3 in multimodal settings (video & audio) to assess what motivates speakers to use sarcasm other than the content of context. This pilot study introduces new elements – *delivery style* and *modality* – and reinforces the previous findings about the contextual and emotional factors associated with the production of sarcasm.

2.5.3 Method

Materials The stimuli used in Study 3 (See Section 2.3) were used as starting materials (N = 40). Going through the results from Study 3, we sorted the stimuli based on the sarcasm ratings by each speaker and observer average. We selected the top, middle, and bottom 10% of the stimuli based on the counts from all participants. We selected eight stimuli for each category of top, middle, and bottom that received the most counts (*sarcasm trigger potential*).

From the selected stimuli (N = 24), we changed the names of the interlocutor to various male names and female names to avoid transfer effect from having the same character name[10]. Two native English-speaking narrators – one female (native speaker of British English) and one male (native speaker of Canadian English)[11] – narrated the selected stimuli in front of a camera and had them video-recorded. They were instructed to narrate each stimulus in two versions, one in which they used the tones and gestures appropriate to each situation description (engaging) and one in which they narrated each situation description in a flat tone without gestures (flat), forming the variable *delivery style*.[12] The narrators narrated the situations with the names corresponding to their gender. Each narrator received 15 EUR per hour as compensation. We extracted audio from the video-recorded stimuli to create audio-only materials using the open-source software FFmpeg[13].

Participants 48 native native English-speaking participants (24 female, Mean_{age} = 36) were recruited online. Participants received 9 GBP per hour as compensation.

Procedure The participants were split into two groups (gender-balanced) and each group participated in either the audio experiment or in the video experiment. Both experiments were conducted with the identical settings except for the stimuli presentation method (video vs. audio). The participants watched the videos or listened to the audio and freely responded to each situation by typing. As this study focused on assessing what "triggers" the use of sarcasm when the contexts

[10]Though this was not a significant effect in previous studies, we speculated that the effect might be stronger in non-textual modalities.

[11]We acknowledge a potential confound in gender and English variety.

[12]We acknowledge that the variable *delivery style* needs more fine-grained control in subsequent studies.

[13]https://www.ffmpeg.org/

are provided in multimodal forms, we chose not to record responses in the multimodal forms due to the practical limitations of collecting such responses in the online experimental setup.

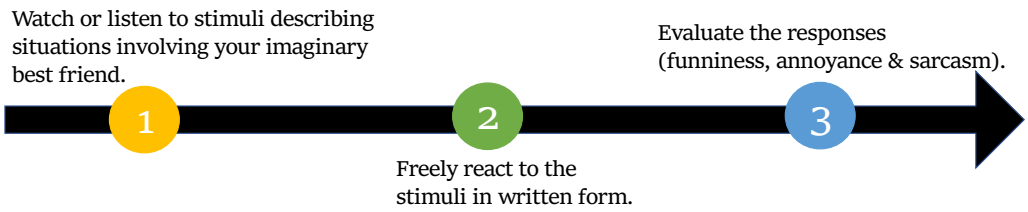

Figure 2.15: Flow of Study 5.

Following the design of Study 3, after responding to all situations, they provided evaluation ratings on a 1-6 scale to six questions about each situation and response – How sarcastic their response was *(sarcasm)*, how funny *(funniness)* or annoying *(annoyance)* they found the friend described in each situation. The experiment lasted 25 minutes on average.

2.5.4 Analysis

Variable coding All scores were z-transformed by each participant for baseline control.

Preliminary analysis We initially constructed the most complex model for assessing the effects of content, affect, modality, and delivery of contexts in triggering sarcasm, which did not converge. We therefore conducted two main analyses, each with different focuses. All p-values in these analyses were adjusted using alpha correction to avoid Type I errors.

Main analysis 1 The first main analysis assessed the connection between the speakers' affect and the level of sarcasm, along with whether the affect of the speakers is influenced by the content of it (*what is said*) or the modality or style in which the context is delivered (*how it is said*). The lmer model predicted sarcasm scores based on funniness and annoyance scores interacting with delivery style (engaging vs. flat) and modality (audio vs. video), and narrator's gender as control variable. A by-participant random slope for all continuous variables and a by-item random intercept were included.

Lmer model 1 formula

[DV] z-sarcasm score ∼
[Main predictors] (z-funniness + z-annoyance) * delivery style * modality + narrator's gender
[By-participant random slope] (0 + z-funniness + z-annoyance|participant) +
[By-item random intercept] (1|item)

Main analysis 2 The second main analysis assessed the effect of modality interacting with the sarcasm potential of the context (judged solely from the textual modality) on the level of sarcasm. To have three levels for the predictor modality (video vs. audio vs. text), we integrated the results from Study 3. We randomly selected results from 24 participants out of 128 participants in total and added the selection to the new results. The lmer model predicted the level of sarcasm (z-scored) given the modality (video vs. audio vs. text) interacting with the sarcasm triggering potential of the contexts (high vs. mid vs. low)[14]. A by-participant and a by-item random intercepts were included.

Lmer main model 2 formula

[DV] z-sarcasm score ∼
[Main predictors] modality * sarcasm-trigger-potential
[By-participant random slope] (1|participant) +
[By-item random intercept] (1|item)

[14]We conducted analyses from different angles by firstly running separate models for different modalities (video, audio, and text), secondly including only two levels (video and audio) for the variable *modality*, and lastly having *modality* as one single predictor. The results were consistent with the reported model.

2.5.5 Results and discussion

Main analysis 1 The main model explained 47% of the variance. The VIFs suggested low collinearity among all variables, with the highest VIF being 2.75 (M = 2.36, SD = 0.51). Table 2.10 reports the lmer coefficients and statistical significance of the predictors on the z-scored sarcasm ratings.

Table 2.10: Results of the statistical modeling. Lmer coefficients predicting self-reported sarcasm ratings with main predictors funniness perception (z), annoyance perception (z), modality (audio vs. video), and delivery style (engaging, flat) and control predictor narrator gender.

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Main predictors	(Intercept)	0.15	0.12	1.25	0.21	
	funniness	0.33	0.09	3.54	<0.005	**
	annoyance	0.32	0.09	3.53	<0.005	**
	delivery (engaging)	0.04	0.15	0.26	0.79	
	modality (video)	-0.07	0.14	-0.48	0.63	
Control predictor	narrator gender (male)	-0.03	0.11	-0.24	0.81	
Interactions	funniness:delivery (engaging)	0.08	0.08	1.04	0.30	
	annoyance:delivery (engaging)	-0.04	0.07	-0.63	0.53	
	funniness:modality (video)	-0.02	0.13	-0.16	0.88	
	annoyance:modality (video)	0.08	0.12	0.63	0.53	
	delivery (engaging):modality (video)	-0.16	0.19	-0.83	0.41	
	funniness:delivery (engaging):modality (video)	-0.04	0.10	-0.42	0.68	
	annoyance:delivery (engaging):modality (video)	0.01	0.09	0.16	0.87	
Conditional R^2	0.47					
Significance	*: p <0.025, **: p <0.005, ***: p <0.0005					

The affect – both funniness and annoyance – motivated speakers to speak sarcastically, which is consistent with the results from Studies 2.1 and 2.3. In contrast, the way the contexts are delivered (*how it is said*) or the modality in which they are delivered did not influence the level of sarcasm in the subsequent responses, nor did they interact with the affect. The gender of the narrator also did not show any effects on triggering sarcasm in the subsequent responses. The results suggest that the way the context is delivered to speakers is not decisive in triggering them to speak sarcastically, whereas the content of the context itself (*what is said*) goes a long way in triggering certain emotions, and thereby predicting the level of sarcasm in subsequent utterances.

Main analysis 2 The main model explained 49% of the variance. The VIFs suggested low collinearity among the predictors, with the highest VIF being 2.25 (M

= 2.00, SD = 0.42). Table 2.11 reports the lmer coefficients and statistical significance of the predictors on the z-scored sarcasm ratings.

Table 2.11: Results of the statistical modeling. Lmer coefficients predicting self-reported sarcasm ratings with main predictors modality (video, audio, text) in interaction with sarcasm-trigger-potential of contexts (high, mid, low).

	Predictors	β	*SE*	*t*	*p*	*Sig.*
Main predictors	(Intercept)	-0.47	0.16	-2.90	<0.005	**
	modality (audio)	0.13	0.27	0.47	0.64	
	modality (video)	0.11	0.27	0.43	0.67	
	sarcasm potential (mid)	0.99	0.19	5.23	<0.0005	***
	sarcasm potential (high)	2.33	0.19	12.34	<0.0005	***
Interactions	modality (audio):sarcasm potential (mid)	-0.28	0.20	-1.39	0.16	
	modality (video):sarcasm potential (mid)	-0.25	0.20	-1.29	0.20	
	modality (audio):sarcasm potential (high)	-0.76	0.20	-3.84	<0.0005	***
	modality (video):sarcasm potential (high)	-0.70	0.20	-3.52	<0.0005	***
Conditional R^2	0.49					
Significance	*: p <0.025, **: p <0.005, ***: p <0.0005					

The sarcasm triggering potential of a context, which was operationalized from the self-rated and other-rated sarcasm ratings obtained from exclusively textual setting in Study 3, affected the level of sarcasm of the responses in the current study as well, with the same pattern (low < mid < high). The added information through other modalities (audio and video) did not contribute to an increased use of sarcasm, shown from the fact that *modality* did not show any difference in the level of sarcasm of the subsequent responses. However, there was a negative interaction between higher sarcasm triggering potential and the modality, in which the level of sarcasm in the subsequent response slightly decreased if the context was delivered through video or audio, compared to pure text. We speculate that there are several reasons for this interaction. One reason could be the different cognitive load the participants may have experienced in each modality. This possibility is also backed by the supplementary analysis we conducted, with each modality in three separate models (sarcasm $\sim$ sarcasm triggering potential). The results were consistent throughout the three models, but the R2 varied for each modality (audio < video < text). When the content of a context clearly calls for an intent to speak sarcastically, it is perhaps best delivered through the simple textual modality, as the information can be processed in parallel, whereas in the video and audio modalities, the information is processed serially, preventing the triggering information from presenting itself in an impactful way. Another reason could be from the ex-

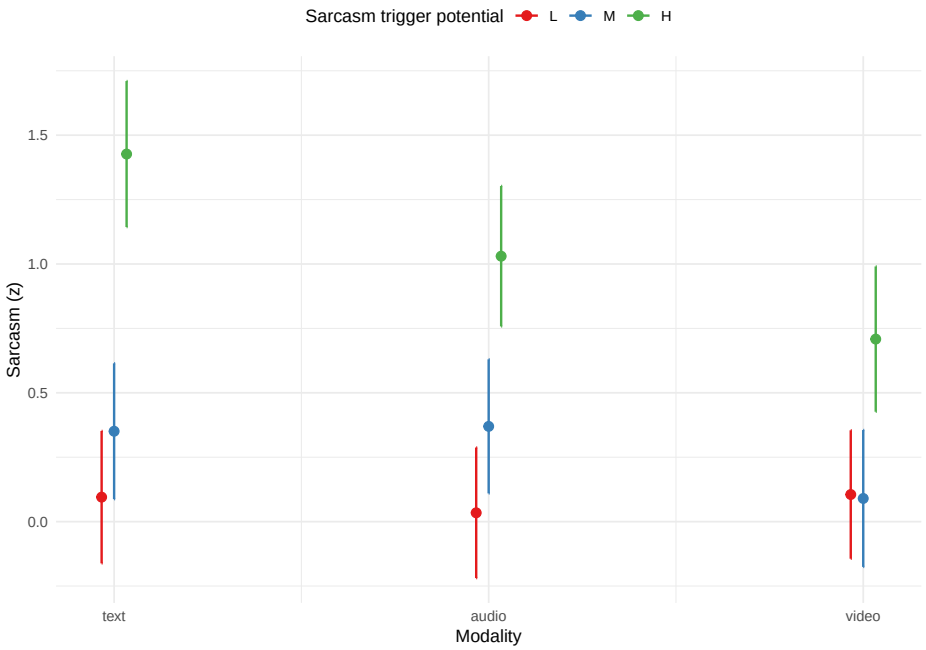

Figure 2.16: Sarcasm ratings by modality and sarcasm triggering potential of the context.

perimental design, as in the text-only settings, participants read each situation at their own pace, whereas in the video and audio settings, participants had to process each situation at the pace of the recordings.

Combining the analyses, it becomes clear that the content embedded in the context is more decisive in predicting higher or lower sarcasm in the following response than the way it is presented. The subject matter of the situation motivates speakers to experience funniness or annoyance, which is likely to be expressed through the use of sarcasm. But if the content of the context is already likely to trigger a strong intent to be sarcastic, visual and auditory modalities lower the chance of a sarcastic response to some degree.

It should be noted, though, that this study was a pilot experiment to gauge the manipulability of various multimodal factors in motivating speakers in the production of sarcasm. Some variables may not have been perfectly dissociated from one another, such as the potential confounds between the narrators, the content of the situations, the delivery styles, and so on. A follow-up study is due in future work.

2.6 General discussion

In this chapter, we investigated various factors surrounding the use of sarcasm in human communication. We started by examining which contextual factors would motivate certain communicative functions that facilitate use of sarcasm. This was based on the evidence from the literature that sarcasm is a figure of speech that facilitates the expression of one's attitude (Colston, 2023), and achieves several communicative functions and discourse goals, such as the enhancement or alleviation of criticism (Colston, 1997; Dews & Winner, 1995), appearing clever and nuanced (Colston, 2023), saving one's face (Jorgensen, 1996), or being humorous (Colston, 2021; Dews et al., 1995). We also connected these factors with the social dynamics between interlocutors, which has also been pointed out as an important aspect of sarcasm (Colston & Rasse, 2022). Several findings have emerged. We discuss the main findings in terms of context, communicative intent, interlocutor perspective, and relationship between interlocutors.

The role of context

Certain contexts that were predicted to prompt more sarcastic reactions were indeed highly associated with sarcastic responses. This occurred via the speaker's emotional reaction to the context. Perceiving an interlocutor as being silly or annoying prompted more sarcastic responses (Study 1). It is evident that the emotional reaction a situation causes has a strong effect in triggering sarcasm. Study 3 identified, in finer granularity, the specific characteristics of a context that is likely to trigger sarcasm. Situations in which the addressee was behaving silly, not seeing their own flaws, or talking about an uninteresting topic prompted more sarcastic responses from the speaker, compared to neutral situations. All these situations are likely to provoke a light negative emotional reaction in the conversation partner, which has been argued to cause sarcasm (Roberts & Kreuz, 1994). In all these situations, though, finding the interlocutor funny prompted even stronger sarcastic responses. The fact that speakers sometimes found the conversation partners to be funny in situations likely to induce negative emotions, and that it led to stronger sarcasm than not having this affect, perhaps point to the fact that sarcasm is generally associated with negative attitudes (i.e., being upset or annoyed), but that there is also an undertone of humor to it (Colston, 2021). Such an interpretation is

reinforced by the last type of contexts – the interlocutor behaving in an entitled or demanding manner. When the speaker perceived this type of context to be funny, the use of sarcastic responses by the speaker increased. On the contrary, if the speaker found the addressee annoying in the same situation, the likelihood of sarcastic responses decreased. Even though annoyance generally triggers higher levels of sarcasm (Study 1), it appears to be the case that the degree or the characteristics of the annoyance matters: Some types of annoyance trigger sarcasm, while other types of annoyance suppress it. A mild annoyance may be positively associated with higher use of sarcasm, but this effect tends to fade away when a stronger degree of negative emotion is present.

The role of communicative intent

Depending on the characteristics of a given situation, speakers develop attitudes or communicative intent, that they attempt to convey in their responses. These communicative intents or discourse goals have long been discussed as the reasons why, or for which, sarcasm is used (Colston, 2021; Dews et al., 1995; Gibbs, 2000; Roberts & Kreuz, 1994; Toplak & Katz, 2000). Our studies confirmed most of such previous findings. Two types of intentions were most strongly associated with sarcasm all throughout the four studies. The intent to mock the addressee in a hilarious or friendly manner prompted speakers to produce sarcastic responses more often, supporting previous findings (Dews & Winner, 1995; Glucksberg, 1995; Jorgensen, 1996; Gibbs, 2000). The intent to speak cleverly also triggered more sarcastic responses, in line with prior work (Glucksberg, 1995). In contrast, the intention to criticize the addressee, whether harshly or softly, had no effect on the responses. At first glance, this finding does not support Colston (1997), who found that sarcasm enhances criticism, and thus argued that sarcasm is used as a form of verbal aggression. But the reason for this discrepancy could lie in the role one has in a communicative situation - speaker, addressee, or observer (Colston, 1997) (see below). Overall, the finding that a certain communicative intent prompts the use of sarcasm is consistent with the literature. Sarcasm appears to be associated with negative attitudes (mocking and criticism) in general but mostly with ones with elements of humor as well. The intent to speak cleverly (Colston, 2023), in line with saving one's face (Jorgensen, 1996), is also a strong factor that motivates the use

and interpretation of sarcasm.

The role of interlocutor perspectives

Speaker's versus listener's perspectives in the interpretation of sarcasm have been compared in previous studies (Toplak & Katz, 2000; Bowes & Katz, 2011). We tested whether the factors that led speakers to be more sarcastic are accurately interpreted by the observers (Study 2 and Study 4). Observers of a conversation could generally identify the factors that triggered the speaker's sarcasm (finding the addressee funny, silly or annoying, having the intention to speak cleverly or mock the addressee). However, a crucial difference emerged between the two perspectives. Observers associated sarcasm with the intent to criticize, but speakers did not. These results partially agree with prior work suggesting that individuals assuming the role of the victim of a criticism felt that sarcastic criticisms were harsher than non-sarcastic criticisms (Bowes & Katz, 2011; Toplak & Katz, 2000). The discrepancy in the interpretation of sarcasm between speakers and observers could be explained by the directionality of communicative intentions and responses. Since an observer does not have access to the speaker's underlying intentions, they must infer the intentions of the sarcastic response. In contrast, the speaker has an intention first, which they then attempt to convey through an utterance. For an observer, during the process of *inferring* plausible intentions behind a sarcastic remark there is a possibility of over-interpretation, because an observer does not have direct evidence about the speaker's intent. Observers therefore likely feel that it is safer to include as many potential intentions as possible in their interpretation of why sarcasm was used, whereas the speaker does not have this issue because they are already aware of their own intentions. Thus, one sometimes finds cases in which sarcasm from the speaker's perspective is not an attempt at criticism, whereas sarcasm from the observer's perspective is likely to trigger the interpretation of criticism. The divergence between the findings of Colston (1997) and of Study 1 could also be attributed to the interlocutor perspective. The experiments in the former had participants take an observer's point-of-view, and therefore found a link between criticism and sarcasm, whereas Study 1 had participants take a speaker's perspective, and found no such link between the two elements. The results of Study 2, which tested the observer's perspective, were consistent with Colston (1997).

Given these results, one would expect that observers would judge sarcasm to be completely negative. However, we found evidence that suggests otherwise. Participants that assumed the role of speakers (Study 1) and observers (Study 2) both judged sarcasm to be bond-enhancing, humorous, and sophisticated rather than offensive. These evaluations about sarcasm do not support previous work that argued that the implied attitude behind sarcasm is that of disapproval, rejection, or aggression (Attardo, 2000; Kumon-Nakamura et al., 1995; Toplak & Katz, 2000). Instead, our results suggest that interlocutors on both sides can interpret sarcasm as something intended to be mocking, clever, bond-enhancing, and humorous as long as the communication happens as intended (Dews & Winner, 1995; Giora et al., 2005). Still, the evaluation of an instance of sarcasm as positive was significantly less likely for observers than for speakers, which is aligned with the general findings in the literature that listeners are more conservative and strict than speakers when it comes to evaluating the effect of sarcasm. When we connect these findings with the previous points of discussion on contextual influence and the role of communicative intent on sarcasm, again, we get an interpretation that sarcasm is associated with light negativity but also with humor and social bonding, though small differences exist depending on the perspectives of conversation partners.

The role of the relationship between interlocutors

Another focal point was the social dynamics between conversation partners. We expanded the interlocutor relationship types to include *close* versus *distant* (Studies 3 and 4). Factors that trigger sarcasm between close friends (i.e., the interlocutor being silly/funny, wishing to mock the interlocutor or to make a clever remark) also applied to distant interlocutors. From the observers' perspectives as well, the different relationships between conversation partners did not affect how they judged the level of sarcasm in speakers' remarks or what may have led them to be sarcastic. Indeed, the type of relationship between interlocutors did not affect sarcasm use, nor did they interact with affect (i.e., whether they found the interlocutor funny or annoying). It appears that contextual cues and the intended communicative functions they elicit are much more powerful factors in triggering sarcasm than relationship type. This partially aligns with the results of Pexman & Zvaigzne (2004), in which the relationship type had no effect on sarcasm rating itself, but affected ratings for

the pragmatic functions related to it (i.e., mockery, politeness). One relevant factor that deserves attention, though, is how much the difference in common ground between interlocutors would affect the use of sarcasm. Kreuz (1996) argued that a reason why sarcasm is used more often in close relationships is the higher chance that sarcasm would be understood well due to higher common ground shared by the speakers. Future research could test whether the varying degrees of shared knowledge in interlocutors of different relationships affects the use of sarcasm.

The role of multimodal factors

We introduced new variables in Study 5 to examine the effect of delivery style and modality on the use of sarcasm. The emotional factors that motivated speakers to speak sarcastically in text-only settings were also found to motivate speakers to use sarcasm in video and audio settings. Multimodal factors introduced in video and audio formats did not change the behavior of speakers in their choice of sarcasm. It appears that the substantial matter in a context is the strongest element that triggers speakers to use sarcasm, rather than how it is said or by whom it is said. Speakers are motivated to speak sarcastically by certain emotional reactions that the context incurs. But this process is most strongly influenced by the substance of the context rather than the the manner of delivery.

2.7 Dataset

In addition to the fact that Studies 2.1 through 2.4 investigated various hypotheses about the way sarcasm is produced and understood, this experimental process has also resulted in the collection of data that we use in Chapter 3 for sarcasm detection. We name the dataset Conversational Sarcasm Corpus (CSC) and make it publicly available at `https://github.com/CoPsyN/CSC`. The dataset consists of two parts, each from Study 1 & Study 2 (part 1), and Study 3 & Study 4 (part2).

Part 1 contains:

- 32 contexts,
- 1,920 responses given by 60 speakers,
- sarcasm ratings (1-6) reported by the speakers for each response,
- affect (*silly or annoying*) ratings (1-6) for each context,
- binary indication of 8 intentions (criticize harsher/softer, mock hilarious/friendly, be natural, be direct, be nice, speak cleverly) for each context + response pair,
- evaluations of the above by 6 observers for each context + response pair.

Part 2 contains:

- 40 contexts,
- 5,120 responses given by 128 speakers,
- sarcasm ratings (1-6) reported by the speakers for each response,
- funniness ratings (1-6) for each context,
- annoyance ratings (1-6) for each context,
- intention to mock (1-6) for 20% of context + response pairs,
- intention to speak cleverly (1-6) for 20% of context + response pairs,
- evaluations of the above by 4 observers for each context + response pair.

2.8 Chapter summary

This chapter examined the factors associated with the use of sarcasm from various angles. Study 1 focused on the factors that may motivate the use of sarcasm in conversations between close friends. It provided situations in which a close friend behaved in a neutral or non-neutral (silly/annoying) manner. Participants freely responded, hence assuming the role of the speakers, and rated the level of sarcasm of their responses and their communicative intent, and emotional reactions to the contexts (self-ratings). The self-rated sarcasm ratings were higher when the preceding context was perceived as non-neutral, and when the speakers intended to mock the addressee or make clever remarks. Study 2 investigated whether the underlying intent and emotional reaction of the speakers, as well as the level of sarcasm, could be accurately detected by external observers. The self-reported ratings in Study 1 were compared to the ratings by the observers, who evaluated the same situations and responses. Though the observers showed similar effects to the speakers, only they associated higher sarcasm ratings with the intent to criticize. Study 3 tested sarcasm-triggering contexts in finer granularity by examining the connection between different characteristics of contexts, and speakers' motivation to use sarcasm. It also expanded the social relationship between interlocutors from close to distant. Having emotional reactions to a situation generally affected the use of sarcasm, consistent with the previous findings. Regarding the characteristics of contexts, situations in which the addressee failed to recognize their own flaws, behaved silly, or talked about uninteresting topics triggered stronger sarcasm use than neutral situations. Situations in which the addressee behaved in entitled or demanding manners triggered stronger sarcasm if the speaker found them funny, but weaker sarcasm if they found them annoying. The social relationship to the addressee did not affect the frequency of sarcasm use. Study 4 replicated Study 3 from the perspective of observers. The observers coincided with the speakers in that they rated utterances occurring in similar types of situations to be more sarcastic. The effect about the social relationship between interlocutors was identical to Study 3. Study 5 introduced new elements of modality and delivery to assess their roles in the production of sarcasm. While the modality and delivery did not contribute to the production of sarcasm, the affect factors that were highly correlated with sarcasm production in all previous studies were, again, strong predictors of sarcasm

use.

Taken together, contexts in which the conversation partner behaved in a silly or annoying manner and the emotional reactions triggered by such behavior consistently prompted stronger sarcasm use in the responses that followed. Annoyance was consistently associated with an increased use of sarcasm, but past a certain level of annoyance, this effect subsided, indicating that sarcasm is mostly associated with lightly negative situations and attitudes rather than serious kinds of negativity. Beyond this, the other most prominent sarcasm-triggering factors were the intentions to mock the addressee and to respond cleverly. These factors affected sarcasm use more than the relationship between interlocutors or auxiliary factors. Such factors were conveyed to external observers successfully as well, but only external observers associated sarcasm with the intent to criticize, which speakers did not. Even so, both speakers and observers alike found sarcasm to be a positive communication tool overall.

Chapter 3

Sarcasm for language models

In Chapter 2, we made several new discoveries about the mechanisms of sarcasm in human communication. Now we shift our focus to sarcasm processing by artificial langauge models. Along with the advances in natural language processing, there has been consistent effort to transfer the ability to maneuver sarcasm into artificial systems, and such effort has clustered around the automatic detection of sarcasm. Prior work in sarcasm detection is mostly concerned with creating systems that can detect sarcasm as accurately as possible. As a result, there are studies that report accuracy scores of over 90% (D. Ghosh et al., 2015; Maynard & Greenwood, 2014). Therefore, one might consider sarcasm detection to be a mostly solved task. However, it should be noted that an absolute majority of prior work relies on datasets sourced exclusively from online communication (Babanejad et al., 2020; Baruah et al., 2020; Cai et al., 2019; Das & Kolya, 2021; A. Ghosh & Veale, 2017; Kumar & Anand, 2020; Khodak et al., 2018; Lu et al., 2024; Misra & Arora, 2023; S. Oprea & Magdy, 2020; Potamias et al., 2020; Rajadesingan et al., 2015; Ren et al., 2023; Riloff et al., 2013; Tan et al., 2023; Tsur, 2010; Yue et al., 2023; Zhang et al., 2023). As online communication has different characteristics compared to face-to-face communication (Aguert et al., 2016; Hancock, 2004), one should question whether sarcasm detection models are heavily biased towards online communication. A case in point is a study that used a different type of dataset that reported lower accuracy scores than other prior work that used social media data (Castro et al., 2019). This motivates the first question we ask in this chapter, which is whether artificial sarcasm detection models can generalize their capabilities to different types of sarcasm, occurring in different domains or types of conversations,

or for different communicative purposes (Section 3.1).

Furthermore, a question worth asking when it comes to the "best" sarcasm detection model is what constitutes the ground truth for its judgment. As sarcasm is a subjective communicative phenomenon, it is more difficult than average to objectively determine whether an utterance is sarcastic. There are previous efforts to address the different perspectives of speakers and observers, and it was discovered that there is a significant amount of disagreement between speakers and observers when it comes to evaluating sarcasm (S. Oprea & Magdy, 2020). However, the reasons for this discrepancy were not sufficiently discussed. We address this topic with two elements that we emphasized as being important for the use of sarcasm in Chapter 2: affect and speaker/observer perspective. We use these two elements to explain the gap in the detection of sarcasm (Section 3.2).

The last topic we address in this chapter is context. Chapter 2 showed that the content in a context is an important factor in motivating speakers to use sarcasm, and that observers can also grasp the entire situation to a certain degree with the context available. We take one step further to assess the importance of context for sarcasm detection models, and connect it with the level of disagreement among human observers. Leveraging context for better sarcasm detection models is not new. However, in prior work, context is often loosely defined as any additional information (e.g., images, user history online, psycholinguistic signals) rather than the preceding situation that motivates speakers to speak sarcastically. We examine the effects of context on the performance of sarcasm detection models, with the context defined as a preceding situation that may lead speakers to use sarcasm (Section 3.3).

To summarize, since sarcasm is a complex linguistic phenomenon with many factors that influence its use, this chapter capitalizes on what we already understand about sarcasm as pointers for probing the encoded information in sarcasm detection models, with the goal of making sarcasm detection more transparent and interpretable (i.e., to be able to understand the reasons why models decide that certain utterances are sarcastic or not). With the investigation of sarcasm detection models from the angles of generalizability, affect, context, and disagreement, this chapter contributes to a clearer understanding of the actual abilities of artificial sarcasm detection models. The focus of this chapter is not to improve sarcasm detection models, but to show what types of knowledge is encoded in the models

when they detect sarcasm. This scientific approach to examining sarcasm detection models provides valuable insights into which particular types of competence or limitation are to be expected when artificial models attempt to detect sarcasm.

3.1 Experiment 1: The generalizability of sarcasm detection models

3.1.1 Background

If there is any consensus around the topic of sarcasm, it is arguably the fact that sarcasm is hard to define (Fox Tree et al., 2020), but much is known about the communicative functions of sarcasm through theoretical and psycholinguistic work. As extensively addressed in Chapter 2, sarcasm can be used to mock (Gibbs, 2000; Pexman & Olineck, 2002), to criticize more harshly (Colston, 1997; Frenda et al., 2022; Keenan & Quigley, 1999; Kreuz & Glucksberg, 1989), to be humorous (Dews et al., 1995; Gibbs, 2000; Glucksberg, 1995; Matthews et al., 2006; Pexman & Olineck, 2002), or to save one's own face (Jorgensen, 1996). Sarcasm does not just have many functions, but it also comes in different forms, such as hyperbole, often with interjections and intensifiers (Joshi et al., 2017), understatement, rhetorical questions (Leggitt & Gibbs, 2000; Oraby et al., 2016), deliberate falsehood (Glucksberg, 1995; Riloff et al., 2013), or self-deprecation (Abulaish & Kamal, 2018).

Given the complexity and diversity in sarcasm, one should ask whether current models are robust enough to be effectively applied to the varied sorts of sarcasm. This question is particularly valuable in computational linguistics, as most prior work addressing sarcasm has adopted a narrow working definition, which is "saying the opposite of the true message, often with the intent to be hurtful" (Cai et al., 2019; Frenda et al., 2022; A. Ghosh & Veale, 2017; Joshi et al., 2015; Pan et al., 2020). This definition is not comprehensive (S. Oprea et al., 2021; Sperber & Wilson, 1981). However, from this basis, much prior work in automatic sarcasm detection has worked with the operationalized definitions of sarcasm, such as contextual incongruity (Joshi et al., 2015), or discrepancy between positive sentiment and negative situation (Riloff et al., 2013). Sarcasm detection has since had fluctuating success rates, from an F-score of 0.51 to 0.97 (Băroiu & Trăușan-Matu, 2022).

Prior work in sarcasm detection has used one dataset at a time, which only includes a specific style and domain of sarcasm. This can lead to frail sarcasm detection models because different datasets come from different sources. Social media is the source that most previous computational work on sarcasm has relied on, using training data such as Twitter (Abu Farha et al., 2022; Barbieri et al.,

2014; Joshi et al., 2015; Khodak et al., 2018; Ptáček et al., 2014; Van Hee et al., 2018) or Reddit (Khodak et al., 2018). Other sources of sarcasm data include online forums (Oraby et al., 2016; Walker et al., 2012), product reviews (Filatova, 2012), conversations (Chakrabarty et al., 2022), or TV series (Castro et al., 2019). Data sources matter for building a more robust sarcasm detector, since Joshi et al. (2015) shows that the same models performed quite differently (around 0.20 in F-score difference) on different datasets. The vast range of domains, styles, and topics that sarcasm can be associated with invites the question of whether sarcasm detection models fine-tuned on a specific dataset are generalizable to other datasets.

Different datasets can also have different sources of sarcasm labels. Existing datasets of sarcasm contain sarcasm labels from different sources: tags (e.g., #sarcasm or /s) (Khodak et al., 2018; Joshi et al., 2015), labels by annotators (Castro et al., 2019; Oraby et al., 2016; Riloff et al., 2013), or labels by the authors of the posts (S. Oprea & Magdy, 2020). Collecting data with labels based on tags is easy and scalable, but the data tend to be noisy and it poses a high risk of including false positives in the dataset (Khodak et al., 2018; Sykora et al., 2020). Most manually annotated data come with labels provided by crowd workers or experts (though it raises the question of "who is an expert of sarcasm?"). These labels are more reliable than tags, but they still pose a risk of wrongly reflecting the intention of the original utterance maker (S. Oprea & Magdy, 2020) as sarcasm can be missed (referred to by some as *sarchasm*) even in multimodal settings (Fox Tree et al., 2020). In text-only settings, which are what many sarcasm detection models are based on, the chance of misinterpreted sarcasm would naturally be much higher, as the available cues are more limited.

More recently, the importance of addressing the sources of sarcasm labels has also been raised, as sarcasm detection models have been found to perform significantly worse when making predictions on data with labels provided by the authors of sarcastic utterances, ranging from an F-score of 0.75 to 0.33 (Abu Farha et al., 2022; S. Oprea & Magdy, 2019, 2020). This is a factor worth examining, since most datasets were created by having human annotators rate the level of sarcasm from the observer's point-of-view (Castro et al., 2019; Riloff et al., 2013; Van Hee et al., 2018).

3.1.2 This study

We examined the robustness of sarcasm detection models by taking into account several factors. We tested the generalizability of sarcasm detection models fine-tuned on different datasets that contain sarcasm from different domains (social media/online vs. offline conversations/dialogues), styles (aggressive vs. humorous), and labels from different perspectives (author labels vs. third-party labels). We specifically addressed the following points. First, we made cross-dataset comparisons to test the generalizability of sarcasm detection models fine-tuned on various datasets. We identified the datasets that lead to the highest prediction accuracy scores across datasets (**P1**). Second, we tested the effect of label source (author vs. third-party) on model performance by using the new dataset we created (**P2**). Lastly, we analyzed different styles of sarcasm found in different datasets, supporting our argument that future work should include a more diverse shapes of sarcasm (**P3**).

3.1.3 Data

We used four datasets of sarcasm with different prominent aspects for the experiment, intended to test the generalizable capabilities of language models. Three of the datasets are publicly available. Below we describe the strengths and weaknesses of each.

1. **MUStARD** (the Multimodal Sarcasm Dataset; Castro et al., 2019) is a dataset of sarcasm that was curated out of several TV shows such as *Friends* and *The Big Bang Theory*. The dataset comes with transcripts, audio, and video files of 690 scenes. Each scene consists of contextual utterances followed by a target utterance, half of which are sarcastic (N = 345). Each scene was annotated by two annotators, who gave binary sarcasm labels to each video. The inter-rater agreement was a Kappa score of 0.23 for the majority of the videos and 0.59 for the remaining ones; a third annotator resolved disagreement.

   Strengths: MUStARD is a multimodal dataset, and the annotation was done based on both contexts and responses. Despite its multimodal nature, there is work that used the dataset in text-only settings as well (Das et al., 2023; Zhang et al., 2021).

Weaknesses: The dataset is relatively small, the annotator agreement is low, and the sentences are scripted (as opposed to naturally-occurring).

2. **Sarcasm Corpus V2** (Oraby et al., 2016) is a dataset containing online posts from 3 different online debate sites, which were extracted from the Internet Argument Corpus (Abbott et al., 2016; Walker et al., 2012). Sarcasm Corpus V2 contains a total of 9,386 written sentences, half of which are sarcastic. Each sentence was annotated by 9 annotators, who gave binary labels after being shown context posts with a following response. The inter-rater agreement was 0.80 (in percent agreement) among 9 annotators and 0.89 among the 3 best annotators[1].

   Strengths: Sarcasm Corpus V2 is large, and several annotators contributed to the annotation based on both contexts and responses.

   Weaknesses: The dataset itself does not provide contexts, and the source domain (online debate sites) can bias the data into a specific type of language use that is not commonly found in regular conversations.

3. **iSarcasmEval** (Abu Farha et al., 2022) is a dataset containing Twitter posts and sarcasm labels by the authors of the posts. iSarcasmEval contains 3,801 sentences, 1,067 of which are sarcastic[2]. Native English speakers on Prolific who were users of Twitter provided their own previous sarcastic and non-sarcastic posts with explanations of why they thought their posts were sarcastic. As the dataset has single author labels instead of third-party labels, no inter-rater agreement is applicable.

   Strengths: iSarcasmEval provides author labels, which distinguishes itself from the previous datasets.

   Weaknesses: It does not contain any contextual information.

The fourth dataset we used is the Conversational Sarcasm Corpus (CSC), described in Section 2.7, which we constructed based on the four studies described in Chapter 2. This dataset has various characteristics that distinguish it from the other three datasets. First, it is based on offline conversational contexts (as opposed to Sarcasm Corpus V2 and iSarcasmEval), is larger (than MUStARD and iSarcasmEval), and

[1]The agreement is for one of the three subsets. The authors do not report the agreement score for all data.

[2]Abu Farha et al. (2022) report slightly different numbers. The numbers reported in this paper are according to their data repository at `https://github.com/iabufarha/iSarcasmEval`.

contains naturally-occurring responses by a number of people (as opposed to MUStARD). It also contains author labels, as well as third-party labels (as opposed to MUStARD and Sarcasm Corpus V2). Since Studies 2 and 4 had multiple evaluators per context-response pair, we calculated the inter-annotator agreement in sarcasm ratings using Kendall W. The agreement was at 0.46 for Study 2.2 and 0.56 for Study 2.4, indicating moderate agreement in both.

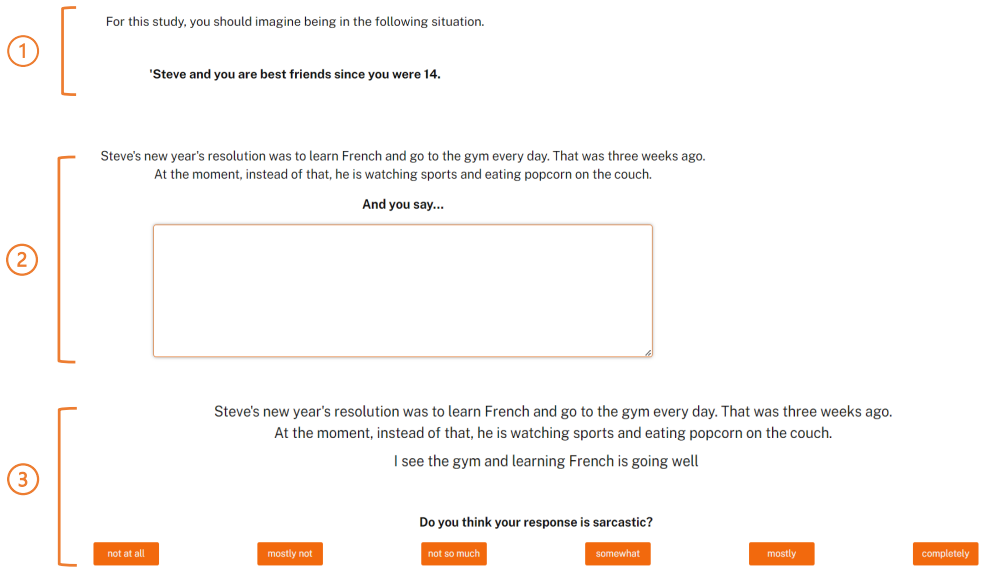

Figure 3.1: Illustration of the data collection process from Chapter 2 reiterated in a simple form.

Table 3.1: Proportions of sarcastic (S) and non-sarcastic (NS) responses by author labels (A) and third-party labels (T) from the data collection process of CSC (Part 1 and Part 2). The original 1–6 labels were binary-coded for this table.

		Part1	Part2	Total	%
A	NS	1,289	3,537	4,826	69
	S	631	1,579	2,210	31
T	NS	1,307	3,331	4,638	66
	S	613	1,785	2,398	34

The CSC dataset has the following structure: context, response, response author id, evaluator id, author label for sarcasm, third-party label for sarcasm. Out of 7,040 collected sentences, around 30% were judged as sarcastic based on binarized ratings (*completely*, *mostly*, *somewhat* to *sarcastic*), which is a higher proportion of sarcasm

than reported in prior work at around 8% $\sim$ 10% (Oraby et al., 2016; Gibbs, 2000). We run all subsequent experiments with binary-coded labels and with balanced number of labels after random downsampling (extracted using the random state of 2) to maximize comparability with the other datasets that are balanced.

Table 3.2: Dataset comparison. A: authors labels, T:third-party labels. -C: without context, +C: with context, Sim.Conv. = Simulated conversations.

	CSC	MUStARD	SC V2	iSarcasmEval
# of sarcastic sentences	2,210 (A) / 2,398 (T)	345	4,693	1,067
# of total sentences	7,040	690	9,386	3,801
# of sent. for training	4,420 (A) / 4,796 (T)	690	9,386	2,134
Average sent. length (-C)	11	12	49	18
Average sent. length (+C)	51	42	-	-
Original source domain	Sim. conversations	TV series	Online debates	Social media
Original label type	Multi (1-6)	Binary	Binary	Binary
Annotator agreement	Moderate (Kendall W 0.56)	Low (Kappa 0.23)	High (Percent agreement 0.80)	N/A
Author labels exist	Y	N	N	Y
Third-party labels exist	Y	Y	Y	N
Is multimodal	N	Y	N	N
Context exists	Y	Y	N	N

Comparing different datasets helps identify the vast spectrum of sarcasm available and further proves the need to test the generalizability of sarcasm detection models on the other datasets for a more robust sarcasm detection system. Table 3.2 shows the points of comparison among the four datasets. Apart from these quantitative and structural differences, here we report some qualitative differences (e.g., styles and topics) we find in each dataset. For CSC and MUStARD, which use contexts that are mostly friendly situations, the elements of comedy and friendly mocking are prevalent. The contexts in MUStARD often do not provide enough information to indicate that the following utterance should be sarcastic or not, because more context has to be drawn up at the episode or series level, or from the multimodal cues, which are often lost in the text-only version. Sarcasm Corpus V2 has the most aggressive, critical, and provocative type of sarcasm, which is understandable as the source of the data is online debate forums, from posts on controversial topics such as “homosexuality”, “abortion”, “gun control”, or “religion”. In contrast, iSarcasmEval has the most instances of self-deprecating sarcasm, since tweets are often used as a channel through which to express one’s own opinion about everyday topics. Even sarcastic utterances that are directed at others in iSarcasmEval exude more distant/cynical attitudes towards the target compared to Sarcasm Corpus V2 (see Example No. 1 of iSarcasmEval in Table 3.3).

Table 3.3: Two examples from each of the four datasets to illustrate the different styles of sarcasm.

Dataset	Ex.	Examples from each dataset
CSC	1	**Context**: Steve gives you a watering can on your birthday while smiling at you with a strange expression. But you don't even have a single plant. **Response**: Maybe I will use it as an outside shower.
	2	**Context**: You know that Steve does almost everything at the last minute. Today Steve tells you that he has to write a joint proposal with colleagues from a different team by the end of this week, and complains that things are moving so slowly because everybody in the other team is a procrastinator. He says, "gosh, it's so frustrating that I'm the only one trying to get the stuff moving. Everyone in this team is too laid back!" **Response**: Oh yeah, and you're not laid-back at all. You are the emperor of procrastination, Steve! If you're the most highly motivated person on that team, they really must be the most unmotivated team in human history.
MUStARD	1	**Context**: 'How do I look?', 'Could you be more specific?', "Can you tell I'm perspiring a little?" **Response**: No. The dark crescent-shaped patterns under your arms conceal it nicely.
	2	**Context**: "So. I just thought the two of us should hang out for a bit. I mean, you know, we've never really talked." **Response**: I guess you'd know that, being one of the two of us, though, right?
SC V2	1	Aaahhh, so just not accomplishing our goals and going home is not defeat, there has to be paperwork for it to be a defeat. Gotcha.
	2	Ever hear of artificial ensemination? Why is that heteros only think there is one way to produce children? I find hetero sex disturbing, and an unnatural lifestyle choice.
iSarcasmEval	1	Imagine going to university for 4 years when you could just follow Elon Musk on twitter for free.
	2	The control button just fell off my laptop. How symbolic for my life.

3.1.4 Experiment

We experimented with three encoder-only models from Transformers: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021). We limited our models to three widely-used and reliable classification models in order to focus on cross-data comparisons. We fine-tuned `bert-base-uncased`, `roberta-base`, and `microsoft/mdeberta-v3-base`. The models were fine-tuned for 2 epochs with a batch size of 64, a learning rate of 5e-5, and a weight decay of 1e-2. We used 5-fold validation and 4 different seeds (1, 11, 21, 31). The fine-tuning was implemented using the Trainer class from the Hugging Face library, and conducted on a NVIDIA A100-PCIE GPU with a total memory of 40GB. All the results reported in Section 3.1.5 are an average across all seeds and folds.

To test the generalizability of sarcasm detection spanning different datasets, we performed sarcasm detection in two ways: intra-dataset and cross-dataset detection. For the intra-dataset settings, all models fine-tuned on dataset A gave predictions on the hold-out test set of dataset A. For the cross-dataset settings, all models fine-tuned on Dataset A gave predictions on Dataset B. For all of these predictions, we did a 5-fold validation (64% training, 16% validation, 20% test).

Table 3.4: F-scores of all intra- and cross-dataset predictions. A: with author labels, T: with third-party labels, +CONT: text consisting of both context and utterance. The best finetuned LM(s) for each test set marked in bold (columnwise).

Model	Finetuned on	Predicted on								
		Intra-dataset	Cross-dataset							
			CSC-A + CONT	CSC-T + CONT	CSC-A	CSC-T	MUS + CONT	MUS	SC V2	iSarcasm
BERT	CSC-A + CONT	0.68	-	-	-	-	0.54	0.56	0.42	0.50
	CSC-T + CONT	0.73	-	-	-	-	0.55	0.57	0.51	**0.53**
	CSC-A	0.67	-	-	-	-	**0.57**	**0.58**	0.39	0.43
	CSC-T	0.70	-	-	-	-	0.56	0.56	0.46	0.47
	MUS + CONT	0.63	0.45	0.46	0.51	0.50	-	-	0.39	0.44
	MUS	0.63	0.47	0.47	**0.53**	**0.52**	-	-	0.40	0.45
	SC V2	**0.77**	0.44	0.44	0.44	0.44	0.39	0.46	-	0.45
	iSarcasm	0.59	**0.48**	**0.48**	0.52	0.51	0.44	0.50	**0.59**	-
RoBERTa	CSC-A + CONT	0.68	-	-	-	-	**0.59**	0.55	0.48	0.52
	CSC-T + CONT	0.72	-	-	-	-	0.57	**0.57**	**0.56**	**0.54**
	CSC-A	0.66	-	-	-	-	0.55	0.56	0.42	0.44
	CSC-T	0.70	-	-	-	-	0.56	**0.57**	0.51	0.51
	MUS + CONT	0.44	0.35	0.35	0.39	0.38	-	-	0.37	0.38
	MUS	0.44	0.35	0.35	0.41	0.40	-	-	0.36	0.40
	SC V2	**0.80**	**0.47**	**0.49**	**0.52**	**0.53**	0.39	0.49	-	**0.54**
	iSarcasm	0.42	0.36	0.35	0.38	0.39	0.36	0.37	0.44	-
DeBERTa	CSC-A + CONT	0.67	-	-	-	-	**0.55**	**0.57**	0.44	**0.52**
	CSC-T + CONT	0.72	-	-	-	-	**0.55**	0.56	0.53	**0.52**
	CSC-A	0.65	-	-	-	-	0.54	0.55	**0.56**	0.48
	CSC-T	0.69	-	-	-	-	0.53	0.54	0.55	0.50
	MUS + CONT	0.44	0.37	0.37	0.40	0.40	-	-	0.45	0.39
	MUS	0.43	0.35	0.35	0.43	0.41	-	-	0.36	0.40
	SC V2	**0.78**	**0.53**	**0.53**	**0.50**	**0.50**	0.37	0.47	-	0.49
	iSarcasm	0.41	0.34	0.34	0.38	0.37	0.45	0.50	0.35	-

3.1.5 Results

Table 3.4 shows the F-scores of all models in the intra-dataset and cross-dataset conditions.

Intra-dataset predictions

For all LMs, the best performance was obtained with Sarcasm Corpus V2 (SC V2), followed by the Conversation Sarcasm Corpus with third-party labels (CSC-T), and lastly iSarcasmEval. The high performance of SC V2, aside from it being the largest dataset with the most annotators, is also attributed to its source: online forums. In these forums, users can only use text for communication, potentially leading to a rich concentration of lexical cues associated with sarcasm, such as aggression and negative emotions. These textual cues are more easily identifiable by LMs, enhancing their ability to detect sarcasm. MUStARD fares worse than most other models, possibly due to the loss of multimodal cues in text-only settings, which normally would have complemented the long-dependency contexts.

The source of sarcasm labels consistently affects the model performance in the intra-dataset settings. Models fine-tuned on iSarcasmEval, which only has author

labels, show the worst performance compared to the others. For CSC also, LMs fine-tuned with third-party labels always perform better than with author labels (**P2**). This observation could suggest that language models may act more as passive observers lacking introspective abilities: Observers, either humans or language models, must rely on external cues and make inferences about an utterance to interpret it, since they have no direct access to the complex motivations of the speaker. This aligns with the existence of *sarchasm*, "missed sarcasm", which is a prevalent phenomenon in human communication (Fox Tree et al., 2020).

Cross-dataset predictions

Overall, all LMs struggle to detect sarcasm on the other datasets proportionately to their performance in intra-dataset settings. Even LMs fine-tuned on SC V2, which showed the highest performance in the intra-dataset predictions, do not generalize nearly as well to the other datasets. The datasets that cause the most struggle for fine-tuned LMs in generalization are SC V2 (BERT), MUStARD (RoBERTa, DeBERTa), and iSarcasmEval (RoBERTa), based on the average score of each row in Table 3.4. In contrast, different versions of CSC used in fine-tuning lead to the best performance in stable generalization to the other datasets for all LMs, except for one case (BERT + iSarcasmEval performed the best on SC V2 with an F-score of 0.59).

Findings from previous work employing similar methodologies suggest that, for models to generalize to other datasets effectively, they must first exhibit robust performance on the data used for their fine-tuning (Fortuna et al., 2021) or that datasets should be large (Halevy et al., 2009; Yin & Zubiaga, 2021). But, our results show that the new CSC dataset can deal with a broader range of sarcasm, despite not being the largest dataset nor yielding the highest accuracy in predictions on its own dataset. Such relative competence could have stemmed from a strong advantage of CSC over other datasets, which is the psycholinguistically-motivated collection and annotation of the data. This suggests that an additional factor in a dataset that contributes to a high generalizability of an LM includes data collection methodology.

The effects of context were tested by fine-tuning the LMs with or without context for CSC and MUStARD. For the *with-context* condition, contexts were concatenated with responses for fine-tuning. For the *without-context* condition, only responses

were used in fine-tuning. This was done because context is an important aspect in humans' processing of sarcasm, but it has been less obvious for LMs (Woodland & Voyer, 2011; Castro et al., 2019; Jaiswal, 2020; D. Ghosh et al., 2018). Also, testing the effects of context was important for CSC since it has an embedded structure in which different responses are preceded by the same context.

The results showed that LMs fine-tuned with context obtain slightly better results than LMs fine-tuned without it (in intra-dataset predictions), though the improvement was very small (between 0.01 and 0.03). In cross-dataset predictions, the presence of context did not affect the generalizability of the models either. Note that this could be the result of the way context was concatenated to the response, or of the number of fine-tuning epochs. For more comprehensive results, future research should address this aspect with targeted experimental manipulations.

The domain of the dataset does not affect the model performance in any obvious way. One might assume that the generalization ability of LMs would benefit from datasets that share a domain (i.e., CSC and MUStARD; both from conversational contexts). This was not the case, as only the LMs fine-tuned on CSC predicted sarcasm on MUStARD well (F-scores 0.53 - 0.59), but not the other way around (F-scores 0.35 - 0.53) except for two cases (BERT + MUS on CSC-A and CSC-T).

The LMs that generalized the best on CSC were in fact fine-tuned on SC V2 (for RoBERTa and DeBERTa) or iSarcasmEval or MUStARD (for BERT), most of which come from a different domain than CSC (social media or debate forums). The success of SC V2 may be attributed to the size and annotation quality of the dataset (Yin & Zubiaga, 2021).

Combining all datasets for fine-tuning did not improve model performance. When BERT was fine-tuned on all datasets combined, the average F-score was 0.71, still lower than the two highest performing datasets (SC V2 and CSC-T-CONT). This shows that a mere combination of several datasets with different sizes, styles, and label sources is insufficient to improve the generalizability of sarcasm detection.

3.1.6 Posthoc analysis

We consider the reasons for the low generalizability of LMs with a post-hoc analysis. We conducted an analysis in which we quantitatively analyze and demonstrate how LMs get accustomed to different types of cues for detecting sarcasm, when fine-

tuned with different datasets (**P3**). We only focus on BERT for this analysis.

Table 3.5: List of used LIWC categories and examples.
Original source at: https://mcrc.journalism.wisc.edu/files/2018/04/Manual_LIWC.pdf

Category	**Examples**
Negations	no, not, never
Positive emotion	love, nice, sweet
Negative emotion	hurt, ugly, nasty
Anxiety	worried, fearful
Anger	hate, kill, annoyed
Sadness	crying, grief, sad
Social processes	mate, talk, they
Family	daughter, dad, aunt
Friend	buddy, neighbor
Cognitive processes	cause, know, ought, think, know, because, should, maybe, always, never
Perceptual processes	look, heard, feeling
Drives	friend, social, win, success, superior, bully, take, prize, benefit, danger, doubt
Religion	altar, church
Swear words	fuck, damn, shit
Online register	btw, lol, thx
Agreement	agree, ok, yes
Nonfluencies	er, hm, umm

Method

We identified certain patterns of language found in each dataset that may have led to differing results. As a proxy for *styles* or *registers*, we chose 17 semantic and psycholinguistic categories provided by the linguistic analysis tool LIWC (Pennebaker et al., 2015), which quantitatively analyzes text in terms of psychological constructs such as emotion, cognition, or perception, among others (See Table 3.5). LIWC produces the ratio of words belonging to each of these categories per sentence.

Working with one dataset at a time, we took all the sarcastic instances correctly identified as sarcastic 1) *only* by BERT fine-tuned on the same dataset (*own success*) and 2) *only* by BERT fine-tuned on the other datasets (*others' success*). For each of these instances, we obtained LIWC scores for the 17 linguistic features (*Feat* in Figure 3.2) and averaged them across all the instances. We calculated the difference between the scores of *own success* and *others' success*. The score difference indicates

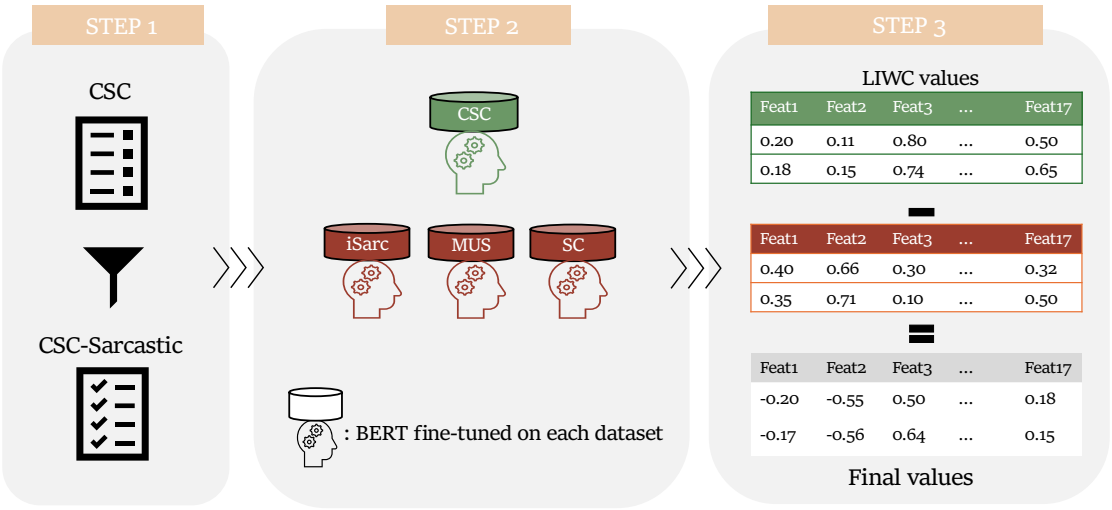

Figure 3.2: Illustration of the post-hoc analysis.

the linguistic properties that facilitate the detection of sarcasm that are uniquely present in one dataset (and not in the others).

Results

Figure 3.3 reports the unique features that help BERT detect sarcasm in each dataset. A major characteristic of Sarcasm Corpus V2 is that it contains many words related to negative emotions such as anger, and words related to social issues, swearwords, online-style words, and disfluencies. This matches our preliminary description of the dataset described earlier that Sarcasm Corpus V2 contains the most aggressive and critical types of sarcasm of out of all the datasets. MUStARD contains a lot of words related to family and drives (i.e., achievement, risk, reward, etc.). We suspect that this phenomenon may be influenced by the TV series The Big Bang Theory, where the characters often converse on topics related to achievement. CSC shows a lot of words related to agreement (i.e., Ok, yes, etc.) possibly because the speakers were instructed to talk to a close friend or colleague. The analysis also reveals a significant amount of words associated with religion, predominantly due to the use of terms like "god" or "Jesus" in expressions such as "oh my god" or "Jesus Christ". This trend might be attributed to numerous situational prompts where the interlocutor's behavior is portrayed as silly, eliciting strong emotional responses.

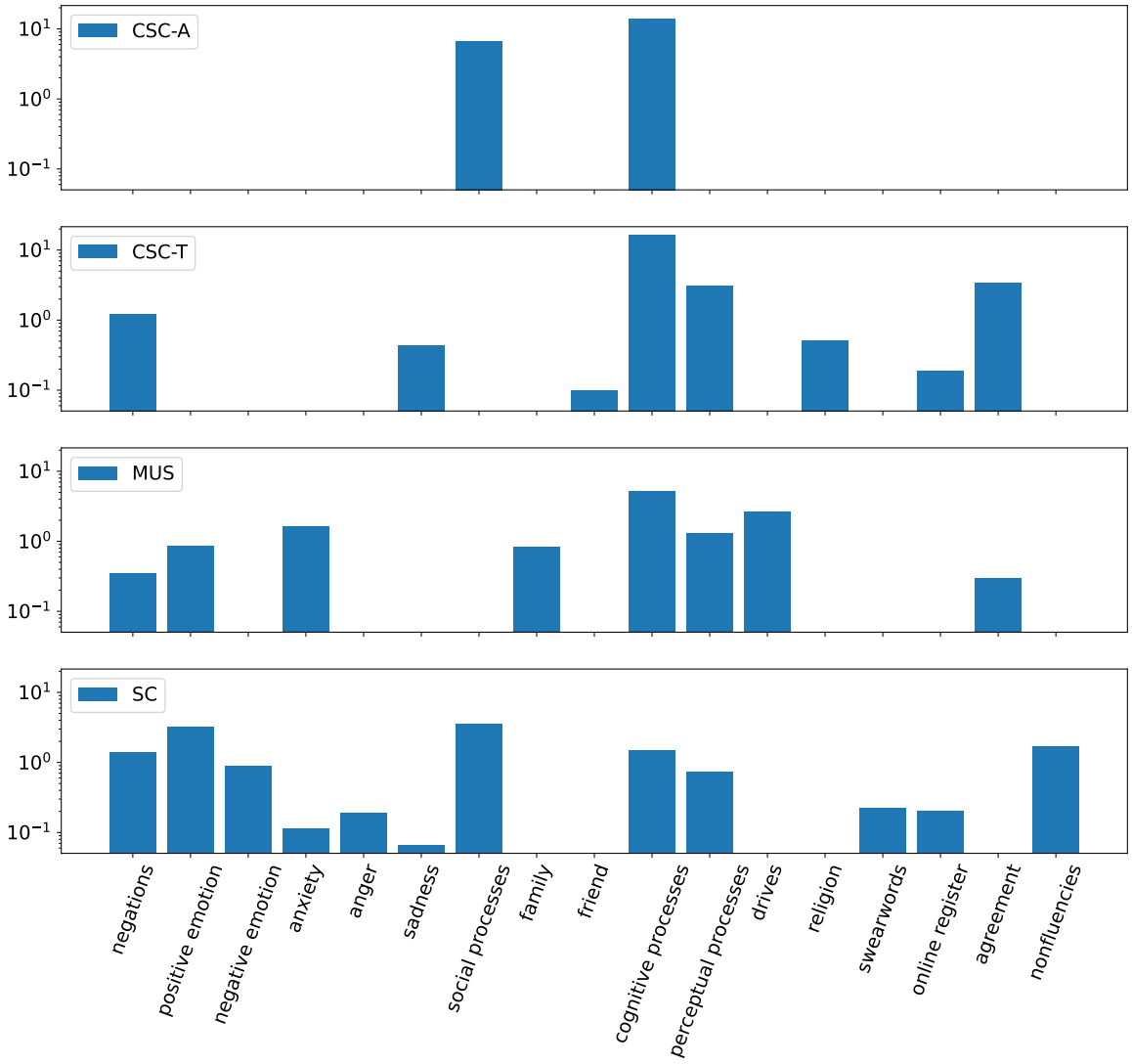

Figure 3.3: Post-hoc analysis to quantify diverging features found in different datasets of sarcasm. The height of each bar represents the log transformed difference between the score for each linguistic property found in sentences uniquely detected by LMs fine-tuned on the same data and those uniquely detected by LMs fine-tuned on the other data. The full description of the categories is provided in Table 3.5.

The results for iSarcasmEval are not reported because there were no instances for which fine-tuning on iSarcasmEval exclusively produced correct predictions (if LMs fine-tuned on iSarcasmEval made correct predictions on some instances, LMs fine-tuned on other datasets also made correct predictions).

The post-hoc analysis of different datasets and model performance shows that sarcasm does indeed come in different shapes and domains, some serious and genuinely meant to inflict verbal pain to strangers, and some humorous and friendly used among well-meaning acquaintances, which is also supported by prior psy-

cholinguistic work (Bowes & Katz, 2011; Colston, 1997; Dews et al., 1995; Frenda et al., 2022; Gibbs, 2000; Pexman & Olineck, 2002).

3.1.7 Discussion

We conducted intra- and cross-dataset sarcasm detection experiments to examine the robustness of sarcasm detection models. We used four datasets containing varying characteristics of sarcasm for comparison, and one of the datasets came from the psycholinguistic studies described in Chapter 2. Several characteristics regarding sarcasm in each dataset, such as label source (authors vs. third-party), domain (social media/online vs. offline conversations/dialogues), style (aggression vs. harmless mocking) were used as points of comparison. All LMs gave better predictions on the same dataset they were fine-tuned on, and showed much lower prediction performance on the other datasets (**P1**). Relatively still, LMs finetuned on our new dataset CSC generalized the best to the other datasets. This was the case even when the domain of sarcasm in the target dataset was different from that of CSC. LMs performed better when the ground-truth labels were from third-party annotators rather than authors themselves (**P2**). Post-hoc analysis and a closer look at each dataset supported our assumption that such low cross-dataset predictions may be attributed to sarcasm coming in various styles, intent, and shapes (**P3**). Therefore, future research should take the vast scope of sarcasm into consideration rather than only focus on a narrow definition of sarcasm as “the utterance of the opposite of the intended meaning” or as “a figure of speech intended to be hurtful, insensitive, and critical”.

3.2 Experiment 2: Affect and failure in sarcasm communication

3.2.1 Background

In Section 3.1, we showed that language models fine-tuned on observer ground-truth labels consistently perform better than those fine-tuned on speaker ground-truth labels. The two ground truths differ because the judgments of sarcasm by speakers and observers can diverge (see Chapter 2). This has repercussions for sarcasm modeling because sarcasm detection datasets mostly have labels annotated by third-party observers (Castro et al., 2019; Oraby et al., 2016), or a combination of self-labels and third-party labels (Khodak et al., 2018; Van Hee et al., 2018). This invites the question of whose sarcasm the models should aim to detect.

Discussing the divergence between intended and perceived sarcasm is not new. Prior research has also shown that the judgment of sarcasm can differ depending on whether the evaluator is the speaker or a third-party observer (S. Oprea & Magdy, 2019). However, the reasons for such discrepancy have not been discussed so far.

In order to understand why such miscommunication happens, it is useful to think about why sarcasm is used in the first place. As discussed in Chapter 2, sarcasm is used to express an attitude (Colston, 2023) or to achieve specific communicative goals, such as to be humorous (Gibbs, 2000), appear emotionally controlled (Dews et al., 1995), or mock (Pexman & Olineck, 2002). Emotion often motivates a speaker to intend such communicative goals (Jang et al., 2023). In fact, many previous studies agree that sarcasm is closely related to emotion (Filik, 2023; Leggitt & Gibbs, 2000; Veale, 2023). For instance, speakers use sarcasm when they are emotionally triggered or motivated (Jang et al., 2023), and listeners are emotionally affected by sarcastic comments in different ways (e.g., offended, criticized, amused) depending on multiple factors such as interlocutor dynamics (Pexman & Zvaigzne, 2004) or situations (Kreuz & Glucksberg, 1989). Therefore, in this section we examine the causes of sarcasm miscommunication using affect, or the "conscious emotion that occurs in reaction to a thought or experience[3]", as an intermediary factor.

[3]As defined by the Merriam-Webster Dictionary.

3.2.2 This study

We identified the factors that cause miscommunication between a speaker's intended sarcasm (or non-sarcasm) and an observer's evaluation, using affect as a focal point. We then investigated whether these factors are reflected in the classification performance of language models in ways that are similar to those found in human communication.

3.2.3 Method

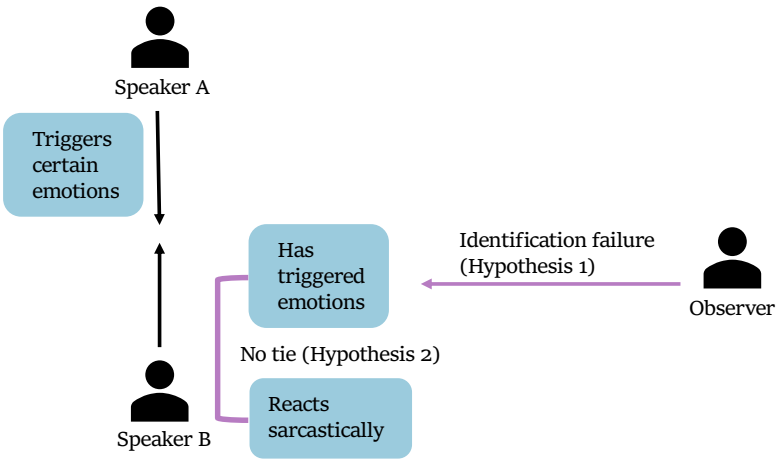

Figure 3.4: Illustration of the suggested hypotheses for explaining sarcasm failure.

Hypotheses

We identified factors that lead to mismatched sarcasm judgment in humans with two hypotheses, based on the close connection between affect and sarcasm: 1) "A failure by observers to identify a speaker's affect in a given situation causes the misaligned sarcasm judgment between speakers and observers (*observers' failure to identify affect*)" *or* 2) "An incongruity between a speaker's affect and the output utterance causes disagreement between self-rated and other-rated sarcasm (*speakers' affect-sarcasm incongruity*)". We developed these hypotheses based on the results from all five studies in Chapter 2, which showed that stronger emotional reactions

to a situation were highly correlated with higher ratings of sarcasm. In a communication setup illustrated in Figure 3.4, we identified two potential places in which a tie is broken, leading observers to reach a different judgment about sarcasm than the speakers.

Data

We used the Conversational Sarcasm Corpus (CSC) to assess differences among individuals and their underlying factors for sarcasm judgment. Since the CSC provides context-response pairs and scores for sarcasm and affect information (i.e., how silly, funny, or annoying the situations were) reported by multiple speakers and observers, it provides the perfect test to examine individual differences in sarcasm. Figure 3.6 shows the cases from the data that are consistent with either hypothesis 1 or 2.

Table 3.6: Examples of *speakers' affect-sarcasm incongruity* and *observers' failure to identify affect* leading to sarcasm failures in Conversational Sarcasm Corpus (CSC).

Hyp #	Hypothesis	Text	Sarc(SP)	Affect(SP)	Sarc(OB)	Affect(OB)
1	**Speaker's affect -sarcasm incongruity**	You got a date this evening. When you tell Steve you got a date, he asks, "oooh, what's the plan?" ***We're going for Malaysian and then a gig.***	6	2	1	1
2	**Observers' failure to identify affect**	Steve says he wants to be young and cool, so he impulsively tries to post a picture on social media, but fails because he doesn't know how. ***It's a young man's game, Steve!***	6	5	1	1

Quantifying (mis)alignment

(Mis)alignment For a systematic assessment, we quantified the misalignment between the speaker and multiple evaluators using normalized mean absolute error (MAE). However, for easier processing of the information, we adopted alignment scores instead, by subtracting the normalized MAE from 1 as follows:

$$1 - \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}|$$

Here y is the speaker score and $\hat{y}$ is an observer score. Though a conventional measure for quantifying errors is the mean squared error (MSE), the mean absolute error (MAE) aligns with the purpose of our task better because the MAE does not

penalize outliers as harshly as the MSE. A single outlier is not much of a communication failure as long as the majority of the observers make judgments similar to the speaker's original intention. A communication act is regarded as successful if most of the observers understood the speaker's intended communicative output.

Table 3.7: Agreement scores for illustration. Context + Response pairs (C + R) have a speaker score (Sp) and multiple evaluation scores by observers (Ev1 $\sim$ Ev6).

C + R	Sp	Ev1	Ev2	Ev3	Ev4	Ev5	Ev6	Avg	**Agreement**
Ex. 1	4	5	6	4	3	2	3	3.86	**0.81**
Ex. 2	4	5	4	5	4	4	1	3.86	**0.86**

The two examples in Table 3.7 have the same average score, but in Example 2, most evaluators agreed with the speaker except for one major outlier, while Example 1 shows less alignment overall between the evaluators and the speaker. Therefore, Example 1 in Table 3.7 gets a lower alignment score of 0.81 whereas Example 2 gets a higher score of 0.86.

We used the alignment formula to quantify the alignment between the speaker's communicative output and the observers' interpretation of it (e.g., between *Speaker* and all *Evs* in Table 3.7). We applied the formula to sarcasm scores (*sarcasm alignment*) and the affect scores (*affect alignment*).

Affect-sarcasm congruity For the speaker's affect-sarcasm congruity, we assigned a value of 1 if the speaker rated sarcasm and affect similarly, either as 1, 2, 3, or as 4, 5, 6. In all other cases, we assigned a value of 0.

3.2.4 Experiments

Experiment 1

The first experiment inspected the causes of sarcasm communication failure between speakers and observers. We tested our two hypotheses to identify the factors that contribute to a mismatch in sarcasm judgment between speaker and observers. We considered *annoyance* as the relevant affect as provided in the data[4]. In a separate preliminary experiment using an emotion classification model fine-tuned on the GoEmotions dataset (Demszky et al., 2020)[5], we tested the validity of *annoyance*

[4]In Part1 of CSC, the affect variable is *silly or annoying*.
[5]`https://huggingface.co/bsingh/roberta_goEmotion`

as an affect relevant to detecting sarcasm. We sampled all sarcastic instances from CSC based on the binarized speaker labels (N = 2,210) and average observer labels (N= 2,398). We predicted 28 emotions using the emotion classification model on those instances. Based on the logits, the top 20% of most important emotions were annoyance, admiration, amusement, approval, and curiosity.

Using the previously mentioned formula, we obtained alignment scores for annoyance and sarcasm, as well as the affect-sarcasm congruity scores. We then used a linear mixed-effects model to test our hypotheses. The model predicted sarcasm alignment given the affect alignment and affect-sarcasm congruity.

LMER Model:

Speaker-observer sarcasm alignment $\sim$ Speaker-observer affect alignment * Speaker's affect-sarcasm congruity + *Random Intercepts*

Experiment 2

The second experiment examined if the *affect alignment* and speaker's *affect-sarcasm congruity* influence the ability of language models to detect sarcasm.

We fine-tuned three language models, BERT, RoBERTa, and DeBERTa, on CSC for sarcasm detection based on two ground-truth label sets - speakers vs. observers. We downsampled CSC to have an equal number of sarcastic and non-sarcastic instances (N = 2,210 vs. 2,398), and used 80% for fine-tuning in 5-fold split (model initialization seeds: 1, 11, 21, 31). For each language model, we obtained predictions on the test set. Using generalized LMER models, for each LM, we assessed the relation between the correctness of LM-predicted labels, and *affect alignment* and speaker's *affect-sarcasm congruity* (6 LMER models in total).

LMER Model (for each LM):

Correctness of LM predicted labels $\sim$ Speaker-observer affect alignment * Speaker's affect-sarcasm congruity + *Random Intercepts*

3.2.5 Results

Experiment 1

Table 3.8 shows the results of the predictions for sarcasm alignment between speakers and observers, given the affect alignment between speakers and observers and

the speakers' affect-sarcasm congruity. The speaker's affect-sarcasm congruity showed a statistically significant positive effect on speaker-observer sarcasm alignment. In contrast, the alignment between the speaker and the observers on speaker's affect led to slightly negative alignment on sarcasm ratings between speaker and observers. We believe this is due to the strong interaction effect between the two predictors. In cases where the speaker's *affect-sarcasm congruity* was preserved, the alignment on affect between speakers and observers led to higher alignment on sarcasm. However, when this congruity was not maintained, the observer's correct identification of speaker's affect no longer contributed to the alignment in sarcasm judgment between speaker and observers. In other words, *affect-sarcasm congruity* appears to be a necessary condition for communication success. Without it, not much can help to reduce the communication gap. However, if this connection is established, alignment on affect can be a booster for a successful communication (See Figure 3.5).

Table 3.8: Results of the statistical modeling. Lmer coefficients predicting sarcasm alignment between speakers and observers with main predictors speaker-observer affect alignment & affect-sarcasm congruity.

Predictors	β	*SE*	*t*	*p*	*Sig.*
(Intercept)	-0.44	0.03	-13.11	<0.001	***
speaker-observer affect alignment	-0.05	0.02	-2.36	0.02	*
affect-sarcasm congruity	0.15	0.02	6.71	<0.001	***
speaker-observer affect alignment:affect-sarcasm congruity	0.42	0.03	15.94	<0.001	***
Conditional R^2	0.21				
Significance	*: p <0.05, **: p <0.01, ***: p <0.001				

Experiment 2

Table 3.9 shows the results for the predictions about the correctness of the LM-generated labels, given the affect alignment between speakers and observers and the speakers' affect-sarcasm congruity. For LMs as well, the speaker's affect-sarcasm congruity was positively correlated with LM prediction with statistical significance. The affect alignment between speakers and observers was a positive significant factor in two out of four models (RoBERTa + speaker labels & DeBERTa + speaker labels). All GLMER models showed a strong interaction term between these two predictors; instances for which speaker's affect and sarcasm were congruous were easier for LMs (higher F1-score) if the observers agreed with the speakers regarding

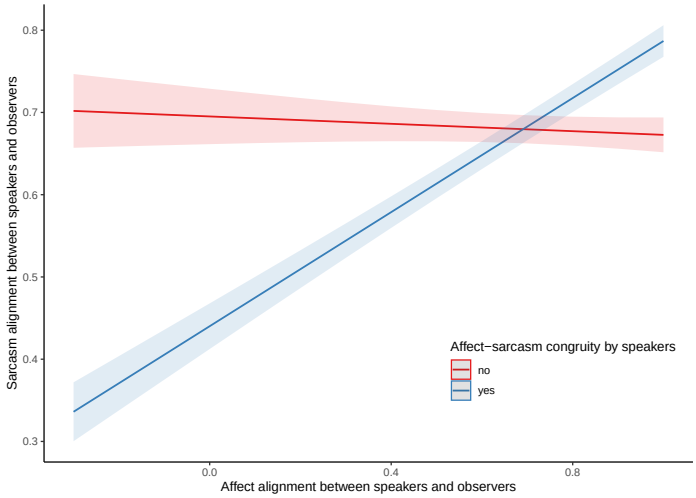

Figure 3.5: Factors that contribute to aligned sarcasm scores between speakers and observers.

underlying affect. For sarcastic instances that seemed unexpected from the observers' perspective, the benefit of affect alignment was no longer present, demonstrating comparable patterns to the human behavior from the previous experiment.

3.2.6 Posthoc analysis

Based on the results from the previous analyses, we leveraged *speaker's affect-sarcasm congruity*, the factor that was crucial for determining sarcasm failure both for humans and models, in seeking different patterns of its assistance to sarcasm detection depending on the affect-sarcasm congruity conditions. We anticipated that adding affect information to sarcasm detection models would make different contributions depending on the speaker's affect-sarcasm congruity. We divided the dataset into two segments (congruous vs. incongruous) and added affect information in the form of logits and assessed if the added information contributes to better sarcasm detection to different degrees in each segment.

We newly fine-tuned BERT and RoBERTa on CSC for sarcasm detection with an 80-20 random split (model initialization seeds: 10 and 20). DeBERTa was not included as it yielded highly similar patterns to RoBERTa in Experiment 2, and we used two seeds instead of four as the results from different seeds in Experiment 2 were almost identical.

We also fine-tuned BERT on CSC for affect detection. When adding affect information to sarcasm detection models, we extracted the logits for affect on the test

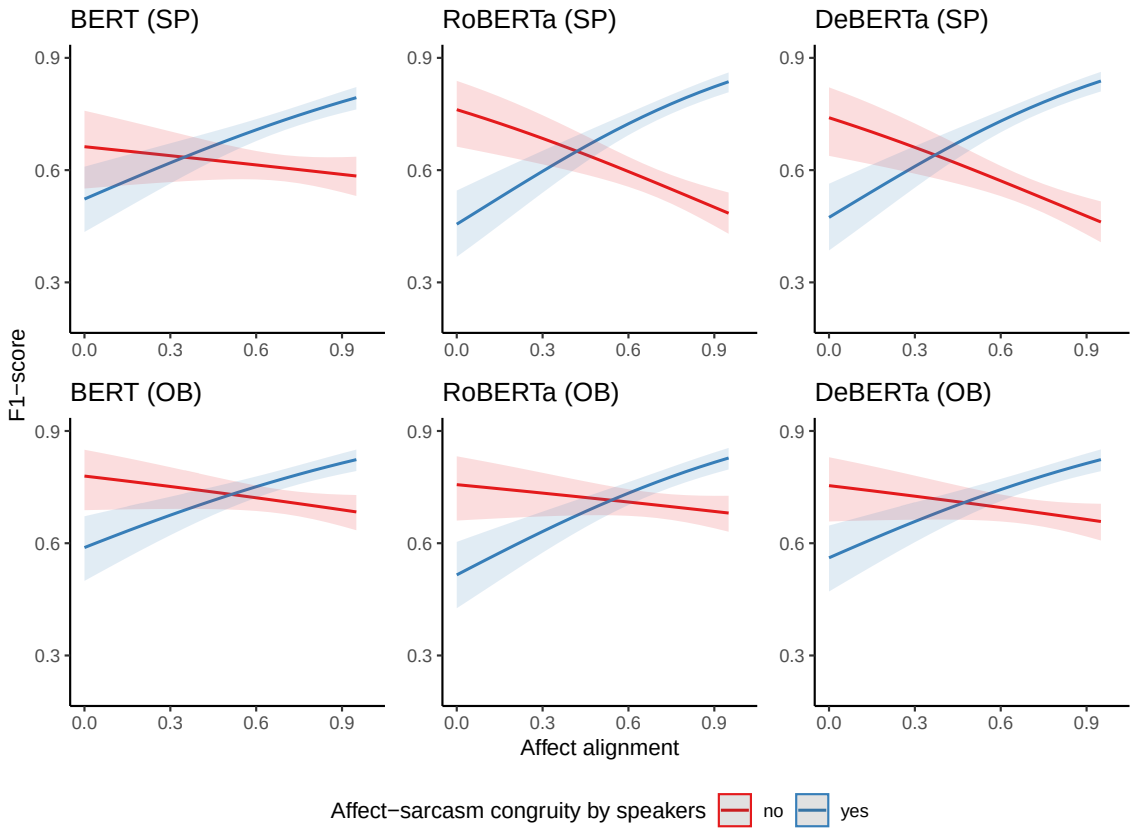

Figure 3.6: Factors that contribute to more correct decisions by BERT (left), RoBERTa (middle), DeBERTa (right) with ground-truth labels from speakers (top) and observers (bottom).

Table 3.9: Results of the statistical modeling. Glmer coefficients predicting the correctness of LM predicted labels with main predictors speaker-observer affect alignment & affect-sarcasm congruity.

	Predictors	$Exp(\beta)$	*p*	*Sig.*
BERT + SP	(Intercept)	1.96	<0.005	**
	affect alignment	0.70	0.28	
	affect-sarcasm congruity	0.56	<0.05	*
	affect alignment:affect-sarcasm congruity	5.34	<0.001	***
RoBERTa + SP	(Intercept)	3.19	<0.001	***
	affect alignment	0.28	<0.001	***
	affect-sarcasm congruity	0.26	<0.001	***
	affect alignment:affect-sarcasm congruity	24.33	<0.001	***
DeBERTa + SP	(Intercept)	2.85	<0.001	***
	affect alignment	0.28	<0.001	***
	affect-sarcasm congruity	0.32	<0.001	***
	affect alignment:affect-sarcasm congruity	22.35	<0.001	***
BERT + OB	(Intercept)	3.54	<0.001	***
	affect alignment	0.60	0.10	
	affect-sarcasm congruity	0.40	<0.01	**
	affect alignment:affect-sarcasm congruity	5.83	<0.001	***
RoBERTa + OB	(Intercept)	3.11	<0.001	***
	affect alignment	0.67	0.21	
	affect-sarcasm congruity	0.34	<0.001	***
	affect alignment:affect-sarcasm congruity	7.28	<0.001	***
DeBERTa + OB	(Intercept)	3.06	<0.001	***
	affect alignment	0.61	0.12	
	affect-sarcasm congruity	0.42	<0.01	**
	affect alignment:affect-sarcasm congruity	6.38	<0.001	***
Significance	*: p <0.05, **: p <0.01, ***: p <0.001			

set. We made sure that sarcasm fine-tuning and affect fine-tuning would be done with the same training and test split settings to avoid LMs being exposed to the same fine-tuning data for affect detection and sarcasm detection.

We used the fine-tuned models as text representation methods without classification heads. After adding affect information as logits, we used a logistic regression classifier (with a 'liblinear' solver that works better for high-dimension data, and the maximum iteration of 500) on the remaining test set with a 5-fold cross-validation.

Results

Table 3.10 shows the results on sarcasm classification and the improvement of performance with the addition of affect information in the different congruity condi-

tions.

Table 3.10: F1-scores (F-sarc) on sarcasm detection by two language models fine-tuned on ground-truth labels from speakers vs. observers (G.T.), with improvement when affect information was added as logits (+Aff), for instances where speakers' affect and sarcasm are congruous vs. incongruous (Congruity).

LM	Congruity	G.T.	F-sarc	+Aff(Ob)	+Aff(Sp)
BERT	Y	O	0.77	+0.00	+0.00
		S	0.72	+0.02	+0.01
	N	O	0.67	+0.02	+0.00
		S	0.61	+0.02	+0.05
RoBERTa	Y	O	0.79	-0.01	-0.01
		S	0.69	+0.00	+0.02
	N	O	0.68	+0.01	+0.00
		S	0.55	+0.02	+0.06

The overall performance of LMs is lower when the sarcasm and affect of the speaker are incongruous with each other compared to when they are congruous, consistent with the results from Experiments 1 and 2. Model performance is the highest when the ground-truth labels come from the observers and speaker's affect-sarcasm congruity is maintained (BERT: 0.77 & RoBERTa: 0.79) and lowest when the ground-truth labels come from the speakers and the affect and sarcasm are incongruous (BERT: 0.61 & RoBERTa: 0.55). It is the easiest for LMs to predict sarcasm in an utterance judged by the observers when the sarcasm in the utterance is congruent with the speaker's affect. Conversely, the most difficult cases for LMs are the prediction of sarcasm in an utterance based on the speaker's self-labels and when the sarcasm label is incongruent with the speaker's affect (out-of-nowhere sarcasm that was not identified by the human observers).

The improvement from the addition of affect information varies depending on the speaker congruity and the ground-truth labels. First, additional affect information does not appear to be helpful for LMs when predicting sarcasm based on observer labels, regardless of the congruity between the sarcasm and the underlying affect. On the other hand, when LMs predict sarcasm based on speaker labels, adding affect information marginally leads to better results, and this improvement is the biggest when affect and sarcasm are incongruous. In the latter case, affect information, especially by the speakers, helps LMs predict sarcasm by around 5-6%.

3.2.7 Discussion

Our second hypothesis was confirmed, that when speakers choose to speak sarcastically for no apparent reason, without evident clues from the observers' perspective, speakers and observers provide different judgments about the level of sarcasm of the utterance. Our first hypothesis, that affect identification failure leads to communication failure, did not show a statistically significant effect in itself. But when the underlying affect is congruous with the sarcastic utterance, the correct identification of speaker's affect by observers helped increase sarcasm alignment, leading to a successful communication.

The classification performance of the language models was also influenced by the same factors leading to human miscommunication. Sarcasm that seems "out-of-nowhere" for observers, because a speakers' affect seems unmatched with the output utterance, is not only extremely difficult for human observers but also for language models. However, when sarcasm is used "proportional" to speakers' affect, models handle the utterances with agreement on affect between speakers and observers better.

The overall trend in the results suggests that what speakers identify as sarcasm has more layers of underlying motivation and affect, whereas observers partially rely on the detectable affect of the speakers when they make sarcasm judgments. This naturally leads to situations in which the speakers may be emotionally affected but choose not to use sarcasm, or, conversely, they may be emotionally unaffected but choose to use sarcasm for other unspecified reasons. Language models show a behavior comparable to human observers, with their judgments of sarcasm and affect generally aligning with each other. Therefore, in this case, there is not much change in their classification performance when affect information is added. In contrast, speakers judge their affect and sarcasm introspectively and therefore independently of each other. So there is a boost in model performance when affect information is added on top of sarcasm information, and this benefit is maximized when the connection between speaker's affect and intention for sarcasm is the weakest.

3.3 Experiment 3: The role of context in relation to disagreement on sarcasm

3.3.1 Background

Previous work in cognitive science has shown the importance of context in sarcasm comprehension (Woodland & Voyer, 2011) and production (Jang et al., 2023) for humans. In computational linguistics, similar observations were made: supplying context to the target utterance boosts sarcasm detection performance of language models, though with more conflicting results: some work reported that supplying context leads to a performance boost in sarcasm detection by neural models (D. Ghosh et al., 2018; Jaiswal, 2020), whereas other work reported no such benefit in using context for the same task (Castro et al., 2019).

However, there has not been much effort in exploring the benefit of varying amounts of contextual information, or in addressing what counts as context. A rigorous definition of "context" is often absent in computational work, as any auxiliary material available for a task at hand is considered a context. Any number of preceding strings can be considered "context", such as previous posts on social media (Jaiswal, 2020; Joshi et al., 2016), previous utterances in a dialogue (Castro et al., 2019), or any additional information such as eye-tracking (Mishra et al., 2016) or images (Schifanella et al., 2016).

3.3.2 This study

This section examined the role of the presence and amount of contextual information in detecting sarcasm. Here, we defined context as the preceding textual utterances that can trigger sarcasm in people, and examined what is a good amount of contextual information to facilitate sarcasm identification for humans and language models. We also showed how context interacts with the level of disagreement among human evaluators.

3.3.3 Method

Initial data

We created a new dataset based on the Multimodal Sarcasm Detection Dataset (MUStARD; Castro et al., 2019). The MUStARD dataset contains written transcriptions of "contexts" (preceding utterances) and the following "response" (or "utterances" in the original dataset) from multiple TV series, and binary labels of sarcasm for the responses (*sarcastic* or *not sarcastic*). We selected 24 contexts that are sufficiently generalizable or ordinary. Situations that did not contain specific keywords or setups heavily associated with a particular TV show were considered generalizable. We collected new responses to these selected situations in an online setting. For a higher level of control, all the selected contexts were limited to ones from the TV series "Friends" and situations happening between two conversation partners. The names of all conversation partners were modified to detach the stimuli from the TV show as much as possible.

The condition in our experiment was the amount of context (long context, short context, no context). We capitalized on the fact that the original dataset contained contextual information in multimodal form which was often not reflected in the transcripts. We described the whole context in a narrative form, by manually referring to the scenes and episodes of the TV show, which counted as *long contexts*. The contexts in raw form in the original dataset were *short contexts*. Therefore, each context was represented twice both as *short context* (SC) in its original utterance form and as *long context* (LC) in a descriptive form. The average number of words was 26 for SC and 66 for LC.

Data collection

Response collection For each LC and SC, we collected new responses to make the stimuli comparable, given that the original dataset had responses only to short contexts. This also allowed us to collect spontaneous responses from multiple lay people, as opposed to responses generated by professional screenwriters.

We recruited 32 native English-speaking participants[6] based in the UK, USA,

[6]We only recruited male participants because some of the selected situations were much more suitable for male speakers than female speakers and the already small number of generalizable contexts could not be further reduced. A follow-up study should include gender as a variable for a more comprehensive evaluation of the use of sarcasm by humans.

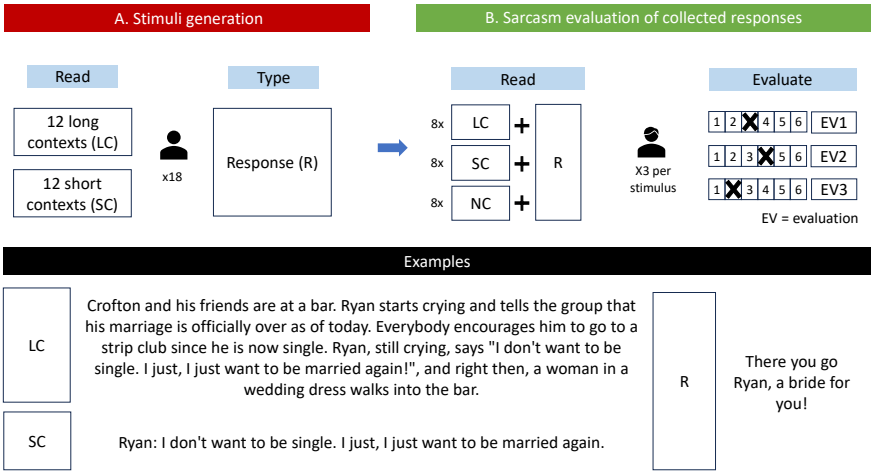

Figure 3.7: Data collection (A), data evaluation (B), and example stimuli for long (LC) and short (SC) contexts.

Canada, Australia, New Zealand or Ireland. They read 24 contexts and freely responded to each, following the general framework from Chapter 2. Half of the contexts (N = 12) were presented as SC and the other half as LC (See A in Figure 3.7). At the end of the collection, participants reported their familiarity to the TV show Friends and the rate by which they recognized any of the situations as being from the show.

To exclude bias coming from the familiarity to the TV show in using sarcasm, we discarded data from the participants who were *quite familiar*, *very familiar*, or *extremely familiar* to the show or who recognized at least 3 scenes from the show. After removing data from 14 such participants, data by 18 respondents remained, creating 432 (18×24) context and response pairs.

Evaluation collection In an online experiment, new participants evaluated the level of sarcasm of the collected responses in isolation (NC) or placed after long context (LC) or short context (SC). In conditions SC and LC, each context is paired with the generated responses and condition NC consists of the responses only (See Table 3.11).

Each stimulus was evaluated by three participants recruited with the same criteria as in the response collection. Each participant was presented with 24 stimuli, distributed evenly across the 3 conditions (See B in Figure 3.7). Participants rated the sarcasm level of the responses on a six-point Likert scale (*not at all*, *mostly not*,

Table 3.11: Number of items for different combinations of context (**C**) and response (**R**).

	Condition	N
i	SC (24) + R (18)	432
ii	LC (24) + R (18)	432
iii	NC (R-only)	432
Total		1,296

not so much, *somewhat*, *mostly*, and *completely*). Participants who failed attention check questions or were familiar with the TV show were replaced with new ones.

Table 3.12 shows the proportions of sarcasm (binary-coded from the six-point scale; *completely*, *mostly*, *somewhat* into *sarcastic*) in each contextual condition by three evaluators and by their average per stimulus[7]. The probability of judging a response as sarcastic increases when contextual information is present. Around 38% of instances that were judged as "not sarcastic" in the NC condition were judged as "sarcastic" when more context became available (LC or SC condition). However, adding context also increases disagreement among evaluators (lower Kappa).

Table 3.12: Proportions of sarcastic responses (binary-coded) by context amount according to three distinct evaluations per stimulus (EVs) and inter-rater agreement (Fleiss' Kappa) by context amount.

	AVG	EV1	EV2	EV3	Kappa
LC	0.46	0.36	0.44	0.49	0.10
SC	0.42	0.36	0.43	0.41	0.13
NC	0.23	0.25	0.25	0.28	0.18

Table 3.13 reports the Spearman's correlation coefficients (r) calculated between the original ratings (1-6 Likert scale) that each evaluation group (EV) assigned to responses alone (NC) and responses following long contexts (LC) or short contexts (SC).

[7]The data consisting of contexts, responses, and evaluation ratings are available at `https://github.com/CoPsyN/sarcasm-friends`.

Table 3.13: Inter-rater agreement of the original ratings (1-6) measured by Spearman's correlations between each pair of evaluation (EV), $p < 0.005$.

	EV1-EV2	EV1-EV3	EV2-EV3
LC	0.26	0.17	0.15
SC	0.26	0.17	0.19
NC	0.18	0.24	0.20

3.3.4 Experiments

Experiment 1

The first analysis tested whether the presence and amount of contextual information are important factors for humans to identify sarcasm in the following response. To increase the comparability of the behavior of humans and LMs, we binarized the sarcasm ratings. The overall inter-rater agreement across all stimuli measured by Fleiss' Kappa was 0.17 (See Tables 3.12 and 3.13 for agreement by condition).[8]

To analyze if there are differences in the sarcasm label (yes/no) distribution given the contextual manipulation, we fit a generalized linear mixed-effects model for each evaluation with by-participant and by-item random intercepts.

Glmer model:

Binary sarcasm labels $\sim$ Context amount + (1|item) + (1|participant)

We used the *emmeans*-package for post-hoc pairwise comparisons (Lenth, 2023). The *emmeans* library conducts a pairwise comparison of the three context conditions (LC vs. SC, LC vs. NC, and SC vs. NC) by performing automatic alpha correction.

Experiment 2

The second experiment tested if manipulating the amount of context directly affects the performance of three language models in the detection of sarcasm on the following response. As gold standard we use the human-evaluated scores described in the Evaluation collection (see Subsection 3.3.3).

We performed sarcasm detection using `bert-base-uncased` (Devlin et al., 2019), `roberta-base` (Liu et al., 2019), and `distilbert-base-uncased` (Sanh et al., 2019).

[8]For comparison, the Kappa score reported in the original MUStARD paper is 0.23 (Castro et al., 2019).

We fine-tuned these pretrained models on the contexts and responses from the MUStARD dataset excluding the 24 contexts we used in our data collection process. We used the collected data as test data, for which each fine-tuned LM classified the responses paired with long contexts, short contexts, or no contexts as either *sarcastic* or *not sarcastic*.

Given the high subjectivity in identifying sarcasm indicated by the low inter-rater agreement (Kappa 0.17), instead of simply combining the multiple human ratings per data point, we predicted the (binary-coded) ratings by the three evaluations (EVs) independently as well as combined. We conducted an error analysis comparing the results from the three EVs.

Each language model was fine-tuned for 2, 5, and 10 epochs with a batch size of 64, a learning rate of 5e-5, and a weight decay of 1e-2. The fine-tuning was implemented using the Trainer class from the Hugging Face library, and conducted on an NVIDIA A100 GPU with a total memory of 40GB. Four different seeds and five folds for validation were used. All the reported results in this section are an average of all the models (4 seeds $\times$ 5 folds) trained for 10 epochs, which yielded the best prediction results.

3.3.5 Results

Experiment 1

For all evaluations (EVs), the presence of context, either long or short, triggered significantly higher probability of perceiving sarcasm in the following response ($p < 0.001$). Long contexts caused more frequent sarcasm judgment compared to short contexts only in EV3 ($p < 0.005$), but not in EV1 ($p = 0.98$), EV2 ($p = 0.97$), or AVG ($p = 0.27$). The results indicate that the presence of context is important for human evaluators to identify sarcasm, but a greater amount of context does not necessarily lead to any added benefit.

Experiment 2

Overall, the three LMs achieve comparable classification results. Supplying context, either short or long, always improves the performance of all LMs. The performance results in Table 3.14 suggest that there are no strong differences between supplying

Table 3.14: Macro F-scores of sarcasm detection on the collected dataset described in Subsection 3.3.3 by three LMs trained on MUStARD for 10 epochs. Labels provided by each evaluation (EV) or combined (averaged and binarized; C) across three EVs.

		EV1	**EV2**	**EV3**	**C**
BERT	LC	0.49	0.54	0.55	0.55
	SC	0.53	0.52	0.54	0.54
	NC	0.47	0.39	0.41	0.34
RoBERTa	LC	0.46	0.52	0.55	0.53
	SC	0.54	0.51	0.53	0.52
	NC	0.36	0.34	0.38	0.29
DistilBERT	LC	0.53	0.55	0.54	0.53
	SC	0.53	0.52	0.53	0.54
	NC	0.44	0.38	0.40	0.32

long context and short context. A noteworthy aspect of these results is that despite low agreement among three evaluations, the prediction results by context amount show similar patterns for all EVs (i.e., LC and SC cause a higher number of correct predictions than NC).

3.3.6 Posthoc analyses

We conducted two error analyses to identify reasons for similar sarcasm classification patterns pertaining to the amount of context, despite low agreement among human evaluators to begin with. The first analysis looked deeply into the classification results to find patterns in classification performance with different evaluations considered as ground-truths. The second analysis examined the relation between disagreement level and the amount of context in classification results by models, inspired by the fact that the amount of context caused different levels of agreement scores in human evaluators (See Subsection 3.3.3).

Disagreement among human evaluators

To identify the reasons behind the similar patterns in model performance despite low agreement, we divided the data into *agreed-upon* (all evaluators agreed on a label) and *disagreed-upon* (evaluators disagreed on the label: 2 vs. 1) instances of sarcasm based on the binarized labels. From the *disagreed-upon* category, we extracted the number of instances for which LMs chose the majority label (better

choice) or the minority label (worse choice), neither of which is completely correct or incorrect. Table 3.15 shows that LMs choose the labels given by each evaluation at a similar rate. The pattern in which there is little variation in the proportions of instances with matches to different ground-truth labels (e.g., EV1 and EV2 vs. EV1 and EV3) suggests that LMs misclassify some sentences when tested with labels from one evaluation, but misclassify other sentences when tested with labels from another evaluation, thus holding the general classification patterns stable[9].

Table 3.15: Proportions of predictions by all LMs. Correct & incorrect predictions apply to *agreed-upon* instances. Majority (better choice) & minority (worse choice) predictions apply to *disagreed-upon* instances.

Type	**Prediction**	**Evaluations that predictions match**			**Proportions**		
					BERT	**RoBERTa**	**DistilBERT**
Agreed	Correct	All			0.58	0.56	0.56
	Incorrect	None			0.42	0.44	0.44
		Match_EV1	**Match_EV2**	**Match_EV3**			
Disagreed	Majority	0	1	1	0.19	0.18	0.19
		1	0	1	0.17	0.18	0.19
		1	1	0	0.20	0.18	0.19
	Minority	0	0	1	0.14	0.16	0.15
		0	1	0	0.15	0.14	0.13
		1	0	0	0.14	0.16	0.15

The interaction between context amount and degree of disagreement

To analyze the interaction between the amount of context (LC, SC, NC) and disagreement levels (agreed vs. disagreed), we categorized the predicted labels according to these factors. Table 3.16 shows that for *agreed-upon* instances, providing context helps LMs predict (more) correct labels than when no contexts are available (LC/SC >NC for correct & majority). For *disagreed-upon* instances, providing longer context shows some benefit in improving the detection of sarcasm compared to providing shorter or no context (LC >SC/NC).

In summary, for sentences with a high agreement, the presence of context is important for LMs to significantly improve their performance of sarcasm detection, but adding more context does not present clear benefit compared to a lower amount

[9]There is potential limitation to this statement because different ground-truth labels are pseudo grouped in the current setup. Randomizing the source of human labels (Ev1, Ev2, Ev3) to different instances of context-response pairs and seeing the results replicate will need to be tested in future work.

Table 3.16: Proportions of classification choice of LMs (average across all seeds and folds) by context length × disagreement level.

		Agreed-upon			Disagreed-upon		
		Correct	Incorrect	Std.	Majority	Minority	Std.
BERT	LC	0.60	0.40	0.07	0.54	0.46	0.05
	SC	0.60	0.40	0.08	0.51	0.49	0.05
	NC	0.55	0.45	0.16	0.50	0.50	0.05
RoBERTa	LC	0.61	0.39	0.09	0.53	0.47	0.05
	SC	0.60	0.40	0.11	0.50	0.50	0.05
	NC	0.54	0.46	0.19	0.50	0.50	0.08
DistilBERT	LC	0.59	0.41	0.08	0.53	0.47	0.05
	SC	0.60	0.40	0.10	0.51	0.49	0.05
	NC	0.55	0.45	0.18	0.51	0.49	0.09

of context. For sentences with disagreement, on the other hand, the contribution of longer contextual information presents more benefit compared to shorter contextual information.

3.3.7 Discussion

In this section, we systematically tested the amount of contextual information required for humans and language models to evaluate the following utterance in terms of sarcasm. We showed that in general, the presence of context leads to better detection of sarcasm both by humans and by three LMs. But, providing a higher amount of information in the context did not present clear additional benefit for humans, which was also true for LMs for sentences for which human evaluators agreed on a label. However, when humans disagreed, the performance of language models improved when a longer context was provided. Finally, we showed that low inter-rater agreement did not affect the overall classification patterns, due to a high variability in the sentences that the models misclassify each time they are tested against labels from different human evaluators. This is a relevant finding for many NLP tasks prone to disagreement and susceptible to subjectivity, which must continue to be investigated in future research.

3.4 General discussion

An expression of sarcasm may, at times, have simple linguistic properties, such that it is easily handled by artificial models (Chakrabarty et al., 2022; Riloff et al., 2013). However, there are several additional layers of information (e.g., motivation, affect, context) compressed into a sarcastic utterance that, once teased apart, give us an indication about its deeper meaning. However, this information is difficult to access for artificial language models (as well as humans, in some cases). In this chapter, we examined computational sarcasm detection by leveraging these additional layers to better understand how sarcasm is processed by language models.

Generalizability The first topic we focused on was the fact that sarcasm is not manifested in limited forms or templates. This can pose difficulty for artificial models. Therefore we tested the robustness of sarcasm detection models by testing whether they can successfully identify sarcasm with varied characteristics, shapes, and contexts (Experiment 1). Experiment 1 examined the robustness of sarcasm detection models by testing whether they can successfully detect sarcasm with different characteristics, shapes, and contexts. We compared the intra- and cross-dataset sarcasm detection performance of several language models using four sarcasm datasets that demonstrate different characteristics of sarcasm. We used a new dataset (CSC) for this process, which was created from repeated psycholinguistic experiments described in Chapter 2. Several characteristics regarding sarcasm in each dataset, such as label source (authors vs. third-party), domain (social media/online vs. offline conversations/dialogues), style (aggression vs. harmless mocking) were used as points of comparison. We consistently observed that language models perform better when the ground-truth labels are from third-party annotators rather than the authors themselves. This suggests that language models are a kind of observer, which needs cues to interpret linguistic expressions, just as human observers do. Additionally, all language models showed much worse predictions on the datasets that they were not fine-tuned on, suggesting fragility in interpreting sarcasm that comes in different styles. Nevertheless, language models fine-tuned on CSC, which was created as a result of several psycholinguistic experiments, generalized the best to the other datasets. This was the case even when the domain of sarcasm in the target dataset was different from that of CSC. A post-hoc analysis and

a closer look at each dataset supported our proposal that such low performance in cross-dataset predictions may be attributed to sarcasm coming in various styles and shapes with hidden communicative intent. The results of the experiments suggested that sarcasm comes in various shapes and styles, rather than simply being the opposite utterance of the intended meaning. This has repercussions for prior work that views sarcasm from the narrow angle of a figure of speech intended to be hurtful, insensitive, and offensive, in that devising a framework that can accommodate the vast scope of sarcasm without such bias could be a way to improve computational sarcasm systems.

Speaker affect and disagreement The complex nature of sarcasm revealed in Experiment 1 motivated more investigation into the subjectivity and variability in the judgment of sarcasm. In Experiment 2, we specifically investigated the causes for communication failure in sarcasm - discrepancy between evaluations provided by speakers and observers - and examined if knowledge about the cause can be effectively used to see how sarcasm is processed by language models.

As a starting point, we used the findings from Chapter 2, where we reported that speakers are motivated to speak sarcastically often when they are emotionally affected by the preceding context. We investigated the underlying factors that lead to instances of sarcasm communication failure, focusing on the misalignment between speakers and observers. We hypothesized that the affect of speakers caused by a situation is often a link to the level of sarcasm of the following utterance, and that the cases for which this link is broken likely result in sarcasm miscommunication between the speakers and observers. We observed through statistical tests that when speakers' affect and sarcasm levels are unmatched, observers often misjudge the sarcasm load in the speakers' utterances, leading to sarcasm failure. This pattern was also manifested in the classification performance by language models. The models detected utterances' sarcasm much more poorly if the speaker's affect and sarcasm were unmatched. When affect information was additionally supplied to the sarcasm detection models in these cases, a higher improvement in detection was observed.

The demonstration of observer-like behavior by language models was a recurrent finding. In Experiment 1 as well, language models exhibited higher performance to observer ground-truth than speaker ground-truth. In Experiment 2, language

models showed sarcasm failure in similar conditions to human observers. This also aligns with what we proposed in Chapter 2. We argued that observers' interpretation of sarcasm takes a backward and connected path, from external cues to presumed emotions of speakers, to the final judgment about sarcasm commensurate to the interpretation of those cues. This is as opposed to a forward and disparate path, with emotional reaction and motivation for sarcasm on separate circuits. These factors provide partial reasons why there are occasional sarcasm failures in both natural and artificial communication settings.

Disagreement and context It is evident that sarcasm carries the risk of not being understood, but it may also be the case that it is understood by some, but not all. The latter situation would result in disagreement among multiple external observers. A higher possibility of disagreement by human evaluators would naturally pose more difficulty for language models as well. Experiment 3 addressed this topic by examining the degree of disagreement among multiple observers, in connection with the amount of context, which is an important factor in the identification of sarcasm for external evaluators. We systematically tested the amount of contextual information required for humans and language models to evaluate the following utterance in terms of sarcasm. In general, the presence of context led to a better detection of sarcasm both by humans and by language models. However, providing a higher amount of information in the context did not present clear additional benefit for humans. This was also true for language models for sentences about which human evaluators agreed on a label. Even so, when humans disagreed, the performance of language models improved if a longer context was provided.

3.5 Chapter summary

This chapter provided results from multiple computational experiments to show how language models detect sarcasm, and what information is encoded in them. We did this by examining several additional factors and varieties of sarcasm. Experiment 1 addressed the generalizability of language models in detecting sarcasm of different styles, and in different data. Language models can detect the kinds of sarcasm on which they are fine-tuned, but cannot detect other kinds reliably. This points to the complexity and variety of sarcasm, which can be a challenge for language models. Experiment 2 investigated why sarcasm communication sometimes fails, that is, why speakers and observers sometimes give different evaluations of sarcasm to the same utterance. The affect of speakers was an important cue for observers. When the speakers' emotional reaction level to a context was unmatched with the level of sarcasm in the following utterance, observers often provided a different sarcasm judgment from that of the speakers. This tendency was also found in language models: the classification of sarcasm was much harder for the affect-sarcasm mismatch cases than the match cases, and the degree of performance improvement when the affect information was additionally supplied was higher for the mismatch cases as well. Experiment 3 also addressed disagreement, but its focus was shifted to disagreement among multiple observers. The amount of context was another factor that was addressed. Observers considered a higher proportion of utterances as sarcastic when context was supplied, but a longer context with more information did not change this evaluation. Language models showed similar behavior, but when human observers disagreed more strongly among themselves, having longer contexts helped language models resolve ambiguity and detect sarcasm more accurately.

Chapter 4

Conclusion and future work

4.1 Summary

In this thesis, we showed how the knowledge obtained from psycholinguistic experiments can be effectively used to clarify our understanding of the underlying mechanisms of sarcasm detection models. We have presented, solely through contextual manipulation, several new findings about the motivation for sarcasm in human communication, and about the ways in which that information is encoded in language models, which has similar patterns to human observers.

In human communication, the decision to speak sarcastically is often motivated by certain emotions, primarily amusement or annoyance toward the addressee. Such affect triggers speakers to convey communicative functions such as mocking or trying to appear clever, which in turn leads to a higher chance of producing a sarcastic remark. Observers of a conversation can typically decode the affect and intent that motivated speakers to speak sarcastically, though not perfectly. Given the close association between affect and sarcasm, when speakers choose to use sarcasm for no apparent reasons from the perspectives of an observer, the chance of miscommunication increases. The dependence on affect is also encoded in language models fine-tuned on sarcasm detection similarly to human observers, and thus they detect sarcasm (claimed by the speakers) less accurately when the congruity between affect and sarcasm is not established, compared to when affect and sarcasm are congruous with each other.

Human observers naturally show disagreement among one another when judging how sarcastic a given remark is. When the disagreement is higher, language

models classify sarcasm (according to the average human judgment) with more error. Specifically in such situations, providing more information through context helps language models detect sarcasm with greater accuracy.

This thesis has identified several reasons why sarcasm may occur, such as when speakers are emotionally motivated to convey a certain communicative intent, but such reasons do not constitute a sufficient condition for the occurrence of sarcasm. There can be other unidentified and unspecified motivations behind a sarcastic remark. Furthermore, sarcasm also comes in a variety of shapes and styles. This is the reason why it can lead to biases when language models are only exposed to certain types of datasets stemming from limited sources, such as online communication. Despite what one might expect given the commonly accepted characteristics of sarcasm, (as a hurtful and critical way of communicating), it in fact serves a wider range of functions. This is one factor that may be responsible for the low generalizability of sarcasm detection models.

4.2 Limitations and future work

There are several relevant topics that were beyond the scope of this thesis. First, the contextual prompts we created and used to elicit sarcasm from participants mostly involved amicable and humorous situations in close relationships. We conducted a pilot study using a simulation of social media context with heavily controversial topics such as religion, politics, and gender issues, but the preliminary results suggested a low level of sarcasm in the responses, often with reports of unwillingness to engage in such discussions at all. This perhaps aligns with what we discovered in our studies, that mild annoyance motivates the use of sarcasm, but that not all kinds of annoyance tend to. Though sarcasm is found on online forums involving serious discussions, perhaps it only happens among already willing participants of such conversations, whereas amicable situations are more universally capable of prompting sarcasm. As pointed out in Chapter 3.1, sarcasm as used in different domains and contexts must researched more thoroughly in future work.

Another issue connected to the previous point is that, due to the experimental and logistical constraints, we only covered conversations happening between two parties with one conversational turn each. The conversations were thus somewhat artificial in their structure. This fact likely excludes potentially relevant conver-

sational phenomena such as banter, which can build up in longer conversations. The investigation of sarcasm occurring in conversations with multiple turns among multiple parties, as well as in more natural settings, must be left to future work.

Additionally, the multimodal sarcasm experiment reported in Chapter 2.5 was not conducted fully fledged, as it was meant as a starting point for leveraging audiovisual information into sarcasm production research, and linking it to computational models. Much more work on this topic is needed.

Finally, our attempt to connect psycholinguistic angles with computational modeling of sarcasm was limited to sarcasm detection. With recent work focused on sarcasm generation (Chakrabarty et al., 2020; S. V. Oprea et al., 2022; Zhao et al., 2023), or sarcasm understanding in a more general sense (Chakrabarty et al., 2022), the framework of the methodology proposed in this thesis should be applied to a wider scope of computational sarcasm modeling.

4.3 Moving forward

We combined psycholinguistic methodology with computational sarcasm detection because we acknowledged how complex sarcasm is. However, sarcasm is not the only topic that this method can be employed for. The combination of psycholinguistic experiments and computational studies can be applied to numerous other linguistic topics, especially to linguistic phenomena that involve multiple layers of twist, motivation, or emotion that are not clearly reflected in the lexical or syntactic characteristics of text (e.g., argumentation, euphemism, metaphor, politeness, etc.). Addressing such complex linguistic topics would benefit from employing a new and interdiscriplinary approach, such as the one proposed in this thesis, because the topics that language models still struggle with the most, despite their general success, tend to involve the creative (Tian et al., 2024), logical (Lal et al., 2021), or subtle (W. Zhu & Bhat, 2021) aspects of human language.

4.4 Final thoughts

This extensive investigation of sarcasm has ironically left us with even more fundamental questions about it, which we briefly mention here. First, what exactly is

sarcasm, really? Theoretically-minded people might like to strictly define sarcasm, and determine how exactly to distinguish it from irony, humor, and so on. However, lay people's ideas about sarcasm seem to include a wider set of characteristics – something indirect, something twisted, intense, and loaded, something layered and therefore clever. Furthermore, who exactly is the authority of determining whether something is sarcastic? If a speaker says that they were being sarcastic in their remark, is it still in fact sarcastic even if nobody else understands that intention? Finally, if artificial agents were able to comprehend and/or produce sarcasm fluently, would this be a good thing?

References

Abbott, R., Ecker, B., Anand, P., & Walker, M. (2016). Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 4445–4452). Portorož, Slovenia: European Language Resources Association (ELRA).

Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 802–814). Seattle, United States: Association for Computational Linguistics.

Abulaish, M., & Kamal, A. (2018). Self-Deprecating Sarcasm Detection: An Amalgamation of Rule-Based and Machine Learning Approach. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 574–579). IEEE.

Aguert, M., Laval, V., Gauducheau, N., Atifi, H., & Marcoccia, M. (2016). Producing irony in adolescence: A comparison between face-to-face and computer-mediated communication. *Psychology of Language and Communication*, *20*(3), 199–218.

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, *32*(6), 793–826.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, *16*(2), 243-–260.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Babanejad, N., Davoudi, H., An, A., & Papagelis, M. (2020). Affective and Contextual Embedding for Sarcasm Detection. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 225–243). International Committee on Computational Linguistics.

Barbieri, F., Saggion, H., & Ronzano, F. (2014). Modelling sarcasm in twitter, a novel approach. In *proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 50–58).

Băroiu, A.-C., & Trăușan-Matu, (2022). Automatic sarcasm detection: Systematic literature review. *Information*, *13*(8), 399.

Baruah, A., Das, K., Barbhuiya, F., & Dey, K. (2020). Context-aware sarcasm detection using BERT. In B. B. Klebanov et al. (Eds.), *Proceedings of the second workshop on figurative language processing* (pp. 83–87). Online: Association for Computational Linguistics.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bowes, A., & Katz, A. (2011). When Sarcasm Stings. *Discourse Processes*, *48*(4), 215–236.

Boylan, J., & Katz, A. N. (2013). Ironic Expression Can Simultaneously Enhance and Dilute Perception of Criticism. *Discourse Processes*, *50*(3), 187–209.

Bromberek-Dyzman, K., Jankowiak, K., & Chełminiak, P. (2021). Modality matters: Testing bilingual irony comprehension in the textual, auditory, and audio-visual modality. *Journal of Pragmatics*, *180*, 219–231.

Cai, Y., Cai, H., & Wan, X. (2019). Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2506–2515). Florence, Italy: Association for Computational Linguistics.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual*

meeting of the association for computational linguistics (pp. 4619–4629). Florence, Italy: Association for Computational Linguistics.

Caucci, G. M., & Kreuz, R. J. (2012). Social and paralinguistic cues to sarcasm. *Humor*, *25*(1), 1–22.

Chakrabarty, T., Ghosh, D., Muresan, S., & Peng, N. (2020). Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7976–7986). Online: Association for Computational Linguistics.

Chakrabarty, T., Saakyan, A., Ghosh, D., & Muresan, S. (2022). FLUTE: Figurative language understanding through textual explanations. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 7139–7159). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

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

Colston, H. L. (1997). Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, *23*(1), 25–45.

Colston, H. L. (2021). Humor and figurative language: Good for a laugh, and more. In M. Strick & T. E. Ford (Eds.), *Current issues in social psychology: The social psychology of humor* (pp. 92–108). Oxon, UK: Routledge.

Colston, H. L. (2023). Irony as social work: Opposition, expectation violation, and contrast. In R. W. Gibbs & H. L. Colston (Eds.), *The cambridge handbook of irony and thought* (pp. 81–95). New York: Cambridge University Press.

Colston, H. L., & Rasse, C. (2022). Figurativity: Cognitive because it's social. In H. L. Colston, T. Matlock, G. J. Steen, & C. F. Burgers (Eds.), *Dynamism in metaphor and beyond* (pp. 243–264). Amsterdam: John Benjamins.

Das, S., Ghosh, S., Kolya, A. K., & Ekbal, A. (2023). Un paralleled sarcasm: a framework of parallel deep lstms with cross activation functions towards detec-

tion and generation of sarcastic statements. *Language Resources and Evaluation*, *57*(2), 765–802.

Das, S., & Kolya, A. K. (2021). Parallel Deep Learning-Driven Sarcasm Detection from Pop Culture Text and English Humor Literature. In I. Pan, A. Mukherjee, & V. Piuri (Eds.), *Proceedings of Research and Applications in Artificial Intelligence* (Vol. 1355, pp. 63–73). Singapore: Springer Singapore.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Association for Computational Linguistics* (pp. 107–116).

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020, July). GoEmotions: A dataset of fine-grained emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4040–4054). Online: Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social functions of irony. *Discourse Processes*, *19*(3), 347–367.

Dews, S., & Winner, E. (1995). Muting the Meaning A Social Function of Irony. *Metaphor and Symbolic Activity*, *10*(1), 3–19.

D'Arcey, J. T., Oraby, S., & Tree, J. E. F. (2019). Wait signals predict sarcasm in online debates. *Dialogue & Discourse*, *10*(2), 56–78.

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 392–398). Istanbul, Turkey: European Language Resources Association (ELRA).

Filatova, E. (2017). Sarcasm detection using sentiment flow shifts. In *Proceedings of the thirtieth international florida artificial intelligence research society conference.*

Filik, R. (2023). Emotional responses to sarcasm. In R. W. Gibbs & H. L. Colston (Eds.), *The cambridge handbook of irony and thought* (pp. 255–271). New York: Cambridge University Press.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, *58*(3), 102524.

Fox Tree, J. E., D'Arcey, J. T., Hammond, A. A., & Larson, A. S. (2020). The sarchasm: Sarcasm production and identification in spontaneous conversation. *Discourse Processes*, *57*(5-6), 507–533.

Frenda, S., Cignarella, A. T., Basile, V., Bosco, C., Patti, V., & Rosso, P. (2022). The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, *193*, 116398.

Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. In A. Balahur, E. van der Goot, P. Vossen, & A. Montoyo (Eds.), *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 161–169). San Diego, California: Association for Computational Linguistics.

Ghosh, A., & Veale, T. (2017). Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 482–491). Association for Computational Linguistics.

Ghosh, D., Fabbri, A. R., & Muresan, S. (2018). Sarcasm analysis using conversation context. *Computational Linguistics*, *44*(4), 755–792.

Ghosh, D., Guo, W., & Muresan, S. (2015). Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1003–1012). Lisbon, Portugal: Association for Computational Linguistics.

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, *115*(1), 3.

Gibbs, R. W. (2000). Irony in Talk Among Friends. *Metaphor and Symbol*, *15*(1-2), 5–27.

Giora, R., Federman, S., Kehat, A., Fein, O., & Sabah, H. (2005). Irony aptness. *Humor*, *18*(1), 23–39.

Glucksberg, S. (1995). Commentary on nonliteral language: Processing and use. *Metaphor and Symbol*, *10*(1), 47–57.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Studies in syntax and semantics iii: Speech acts* (pp. 183–198). New York: Academic Press.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE intelligent systems*, *24*(2), 8–12.

Hancock, J. T. (2004). Verbal Irony Use in Face-To-Face and Computer-Mediated Conversations. *Journal of Language and Social Psychology*, *23*(4), 447–463.

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, *abs/2111.09543*.

Hull, R., Tosun, S., & Vaid, J. (2017). What's so funny? modelling incongruity in humour production. *Cognition and Emotion*, *31*(3), 484–499.

Jaiswal, N. (2020). Neural sarcasm detection using conversation context. In *Proceedings of the second workshop on figurative language processing* (pp. 77–82).

Jang, H., Braun, B., & Frassinelli, D. (2023). Intended and perceived sarcasm between close friends: What triggers sarcasm and what gets conveyed? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

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

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, *50*(5), 1–22.

Joshi, A., Sharma, V., & Bhattacharyya, P. (2015). Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association*

for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 757–762). Association for Computational Linguistics.

Joshi, A., Tripathi, V., Bhattacharyya, P., & Carman, M. J. (2016). Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In S. Riezler & Y. Goldberg (Eds.), *Proceedings of the 20th SIGNLL conference on computational natural language learning* (pp. 146–155). Berlin, Germany: Association for Computational Linguistics.

Keenan, T. R., & Quigley, K. (1999). Do young children use echoic information in their comprehension of sarcastic speech? a test of echoic mention theory. *British Journal of Developmental Psychology*, *17*(1), 83–96.

Khodak, M., Saunshi, N., & Vodrahalli, K. (2018). A large self-annotated corpus for sarcasm. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Kreuz, R. J. (1996). The use of verbal irony: Cues and constraints. In *Metaphor: Implications and Applications* (pp. 23–38). Psychology Press.

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, *118*(4), 374.

Kumar, A., & Anand, V. (2020). Transformers on sarcasm detection with context. In B. B. Klebanov et al. (Eds.), *Proceedings of the second workshop on figurative language processing* (pp. 88–92). Online: Association for Computational Linguistics.

Kumon-Nakamura, S., Glucksberg, S., & Brown, M. (1995). How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*, *124*(1), 3.

Lal, Y. K., Chambers, N., Mooney, R., & Balasubramanian, N. (2021). TellMeWhy: A dataset for answering why-questions in narratives. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 596–610). Online: Association for Computational Linguistics.

Leggitt, J. S., & Gibbs, R. W. (2000). Emotional Reactions to Verbal Irony. *Discourse Processes*, *29*(1), 1–24.

Lenth, R. V. (2023). emmeans: Estimated marginal means, aka least-squares means [Computer software manual]. (R package version 1.8.6)

Li, Z., Gao, X., Zhang, Y., Nayak, S., & Coler, M. (2024). A functional trade-off between prosodic and semantic cues in conveying sarcasm. *arXiv preprint arXiv:2408.14892*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, Q., Long, Y., Sun, X., Feng, J., & Zhang, H. (2024). Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion*, *104*, 102203.

Matthews, J. K., Hancock, J. T., & Dunham, P. J. (2006). The Roles of Politeness and Humor in the Asymmetry of Affect in Verbal Irony. *Discourse Processes*, *41*(1), 3–24.

Maynard, D., & Greenwood, M. (2014). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 4238–4243). Reykjavik, Iceland: European Language Resources Association (ELRA).

Mishra, A., Kanojia, D., Nagar, S., Dey, K., & Bhattacharyya, P. (2016). Harnessing cognitive features for sarcasm detection. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1095–1104). Berlin, Germany: Association for Computational Linguistics.

Misra, R., & Arora, P. (2023). Sarcasm detection using news headlines dataset. *AI Open*, *4*, 13-18.

Neuhaus, L. (2023). Irony and its overlap with hyperbole and understatement. In R. W. Gibbs & H. L. Colston (Eds.), *The cambridge handbook of irony and thought* (pp. 310–324). New York: Cambridge University Press.

Oprea, S., & Magdy, W. (2019). Exploring Author Context for Detecting Intended vs Perceived Sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2854–2859). Florence, Italy: Association for Computational Linguistics.

Oprea, S., & Magdy, W. (2020). iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics.

Oprea, S., Wilson, S., & Magdy, W. (2021). Chandler: An explainable sarcastic response generator. In H. Adel & S. Shi (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations* (pp. 339–349). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Oprea, S. V., Wilson, S., & Magdy, W. (2022). Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7686–7700). Dublin, Ireland: Association for Computational Linguistics.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., & Walker, M. (2016). Creating and characterizing a diverse corpus of sarcasm in dialogue. In R. Fernandez, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, & A. Gainer (Eds.), *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue* (pp. 31–41). Los Angeles: Association for Computational Linguistics.

Pan, H., Lin, Z., Fu, P., Qi, Y., & Wang, W. (2020). Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1383–1392). Online: Association for Computational Linguistics.

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015* (Tech. Rep.). University of Texas at Austin.

Pexman, P. M., & Olineck, K. M. (2002). Does Sarcasm Always Sting? Investigating the Impact of Ironic Insults and Ironic Compliments. *Discourse Processes*, *33*(3), 199–217.

Pexman, P. M., & Zvaigzne, M. T. (2004). Does Irony Go Better With Friends? *Metaphor and Symbol*, *19*(2), 143–163.

Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309–17320.

Ptáček, T., Habernal, I., & Hong, J. (2014). Sarcasm detection on Czech and English Twitter. In J. Tsujii & J. Hajic (Eds.), *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 213–223). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm Detection on Twitter: A Behavioral Modeling Approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 97–106). Shanghai China: ACM.

Ren, Y., Wang, Z., Peng, Q., & Ji, D. (2023). A knowledge-augmented neural network model for sarcasm detection. *Information Processing Management*, *60*(6), 103521.

Riloff, E., Qadir, A., Surve, P., Silva, L. D., Gilbert, N., & Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 704–714). Association for Computational Linguistics.

Roberts, R. M., & Kreuz, R. J. (1994). Why Do People Use Figurative Language? *Psychological Science*, *5*(3), 159–163.

Rockwell, P. (2003). Empathy and the expression and recognition of sarcasm by close relations or strangers. *Perceptual and motor skills*, *97*(1), 251–256.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, *abs/1910.01108*.

Schifanella, R., De Juan, P., Tetreault, J., & Cao, L. (2016). Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th acm international conference on multimedia* (pp. 1136–1145).

Sperber, D. (1984). Verbal Irony: Pretense or Echoic Mention? *Journal of Experimental Psychology*, *113*(1), 130–136.

Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. *Radical Pragmatics*, 295–318.

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

Tabacaru, S., & Lemmens, M. (2014). Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding. *The European Journal of Humour Research*, *2*(2), 11–31.

Tan, Y. Y., Chow, C.-O., Kanesan, J., Chuah, J. H., & Lim, Y. (2023). Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless personal communications*, *129*(3), 2213–2237.

Tian, Y., Ravichander, A., Qin, L., Le Bras, R., Marjieh, R., Peng, N., … Brahman, F. (2024). MacGyver: Are large language models creative problem solvers? In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 5303–5324). Mexico City, Mexico: Association for Computational Linguistics.

Tian, Y., Xu, N., Zhang, R., & Mao, W. (2023). Dynamic routing transformer network for multimodal sarcasm detection. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2468–2480). Toronto, Canada: Association for Computational Linguistics.

Toplak, M., & Katz, A. N. (2000). On the uses of sarcastic irony. *Journal of Pragmatics*, *32*(10), 1467–1488.

Tsur, O. (2010). ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (pp. 162–169). Association for the Advancement of Artificial Intelligence.

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

Van Hee, C., Lefever, E., & Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. In M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, & M. Carpuat (Eds.), *Proceedings of the 12th international workshop on semantic evaluation* (pp. 39–50). New Orleans, Louisiana: Association for Computational Linguistics.

Veale, T. (2023). Great expectations and epic fails: A computational perspective on irony and sarcasm. In R. W. Gibbs & H. L. Colston (Eds.), *The cambridge handbook of irony and thought* (pp. 225–234). New York: Cambridge University Press.

Vitman, O., Kostiuk, Y., Sidorov, G., & Gelbukh, A. (2023). Sarcasm detection framework using context, emotion and sentiment features. *Expert Systems with Applications, 234*, 121068.

Walker, M., Tree, J. F., Anand, P., Abbott, R., & King, J. (2012). A corpus for research on deliberation and debate. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 812–817). Istanbul, Turkey: European Language Resources Association (ELRA).

Woodland, J., & Voyer, D. (2011). Context and Intonation in the Perception of Sarcasm. *Metaphor and Symbol, 26*(3), 227–239.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science, 7*, e598.

Yue, T., Mao, R., Wang, H., Hu, Z., & Cambria, E. (2023). Knowlenet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion, 100*, 101921.

Zhang, Y., Liu, Y., Li, Q., Tiwari, P., Wang, B., Li, Y., ... Song, D. (2021). Cfn: A complex-valued fuzzy network for sarcasm detection in conversations. *IEEE Transactions on Fuzzy Systems*, *29*(12), 3696-3710.

Zhang, Y., Ma, D., Tiwari, P., Zhang, C., Masud, M., Shorfuzzaman, M., & Song, D. (2023). Stance-level sarcasm detection with bert and stance-centered graph attention networks. *ACM Transactions on Internet Technology*, *23*(2), 1–21.

Zhao, W., Huang, Q., Xu, D., & Zhao, P. (2023). Multi-modal sarcasm generation: Dataset and solution. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 5601–5613). Toronto, Canada: Association for Computational Linguistics.

Zhu, N., & Wang, Z. (2020). The paradox of sarcasm: Theory of mind and sarcasm use in adults. *Personality and Individual Differences*, *163*, 110035.

Zhu, W., & Bhat, S. (2021). Euphemistic phrase detection by masked language model. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Findings of the association for computational linguistics: Emnlp 2021* (pp. 163–168). Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, *1*(1), 3–14.