THE UNIVERSITY
of EDINBURGH

Computational Sarcasm Detection and Understanding in Online Communication

Silviu Vlad Oprea

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2023

Abstract

The presence of sarcasm in online communication has motivated an increasing number of computational investigations of sarcasm across the scientific community. In this thesis, we build upon these investigations. Pointing out their limitations, we bring four contributions that span two research directions: sarcasm detection and sarcasm understanding.

Sarcasm detection is the task of building computational models optimised for recognising sarcasm in a given text. These models are often built in a supervised learning paradigm, relying on datasets of texts labelled for sarcasm. We bring two contributions in this direction. First, we question the effectiveness of previous methods used to label texts for sarcasm. We argue that the labels they produce might not coincide with the sarcastic intention of the authors of the texts that they are labelling. In response, we suggest a new method, and we use it to build iSarcasm, a novel dataset of sarcastic and non-sarcastic tweets. We show that previous models achieve considerably lower performance on iSarcasm than on previous datasets, while human annotators achieve a considerably higher performance, compared to models, pointing out the need for more effective models. Therefore, as a second contribution, we organise a competition that invites the community to create such models.

Sarcasm understanding is the task of explicating the phenomena that are subsumed under the umbrella of sarcasm through computational investigation. We bring two contributions in this direction. First, we conduct an alaysis into the socio-demographic ecology of sarcastic exchanges between human interlocutors. We find that the effectiveness of such exchanges is influenced by the socio-demographic similarity between the interlocutors, with factors such as English language nativeness, age, and gender, being particualry influential. We suggest that future social analysis tools should account for these factors. Second, we challenge the motivation of a recent endeavour of the community; mainly, that of augmenting dialogue systems with the ability to generate sarcastic responses. Through a series of social experiments, we provide guidelines for dialogue systems concerning the appropriateness of generating sarcastic responses, and the formulation of such responses.

Through our work, we aim to encourage the community to consider computational investigations of sarcasm interdisciplinarily, at the intersection of natural language processing and computational social science.

Lay Summary

Sarcasm is present in online communication. For instance, it can be observed in exchanges between users on the Twitter social media platform.

The opinions and sentiments that Twitter users express in their exchanges are mined and monitored by computational systems. The insight provided by such systems can drive crucial marketing, administration, and investment decisions. For instance, the sentiment expressed by Twitter users about a certain product recently launched by a company could drive that company's future decisions. The company might decide to invest more capital in the development of that product if the sentiment is positive. However, the presence of sarcasm can conceal the intended meaning of a tweet, making it challenging for computational systems to discern the real sentiment expressed by that tweet. For accurate discernment, it is crucial that these systems detect the presence of sarcasm.

This has motivated an increasing number of computational investigations of sarcasm across the scientific community. In this thesis, we build upon these investigations. Pointing out their limitations, we bring four contributions.

First, we construct a novel dataset of tweets, each tweet being accompanied by a label indicating whether it is sarcastic, or non-sarcastic. This dataset can be used by future research in the design of computational systems that detect sarcasm in tweets.

Second, we organise a competition where we ask participating teams to design and submit such systems. We analyse the performance that their submissions achieve at detecting sarcasm in our dataset.

Third, we conduct an analysis into the socio-demographic ecology of sarcastic exchanges. For instance, we show that people of the same age understand each other's sarcasm better in online communication, compared to people of different ages. This suggests that, when discerning the sarcastic nature of a tweet, computational systems should account for the age of the Twitter user who posted that tweet.

Finally, we reflect upon a recent endeavour of the community, mainly, that of augmenting dialogue systems, i.e. chatbots, with the ability to generate sarcastic responses. Such endeavours are motivated by the potential to create more approachable, human-like chatbots, considering that sarcasm is a natural part of human exchanges. Through a series of social experiments, we provide guidelines for chatbots concerning the appropriateness of generating such responses, and the formulation of such responses.

Acknowledgements

First and foremost, I thank God. “Let the words of my mouth and the meditation of my heart be acceptable to you, O Lord, my rock and my redeemer” (Psalm 19:14). May everything that I do and that I am be Yours, for Your glory, and not mine. After God, my gratitude goes to two heroes that I look up to, both of whom are wonderful women. First, I thank Mary, mother of Jesus, and my dearest mother. Second, I thank St. Thérèse of Lisieux.

I thank my supervisor, Walid Magdy. He has been constantly generous with his time, during the day and night, supporting, encouraging, and motivating me, both profesionally and personally, during easy times, and during difficult times. In the process, we became friends. I thank God for having sent him in my way. Thank you, Walid.

I thank my second supervisor, Bonnie Webber. Her feedback has decisively shaped the course of my PhD journey. She helped me see my work in a wider context, and dramatically influenced my choices when my view was too narrow. Thank you, Bonnie.

I thank Steven Wilson, for his constant availability and support in the last two years of my PhD. In the process, we also became friends. His support continued even after leaving the University of Edinburgh to start his role as an assistant professor at Oakland University. Thank you, Steve.

I was blessed to be accompanied by many other wonderful people on my journey towards the completion of my PhD. I am grateful to my third supervisor Maria Wolters, to all members of the SMASH group who pursued their PhD at the same time as I have, including Abeer and Dilara. I extend special gratitude to my friends Ibrahim Abu Farha and Youssef Al Hariri, also members of the SMASH group. I also thank my fellow PhD student from the Centre for Doctoral Training, Michael Camilleri.

Nothing would have been possible without my Anca, my wife and my best friend. You are a gift to me. I love you and I appreciate all that you do to support me.

Words are not enough to thank my dear mother, Gheorghina, and my dear father, Simion; my dear sisters, Maria and Tereza, who have brought so much undeserved, unconditional love and joy into my life; and the other wonderful women in my life who were nothing less than mothers, my aunt and my two grandmothers, all named Maria. I also thank Maria and Marius, my wife’s parents, and Mara Hosu, my wife’s sister, for having adopted me as son and brother. Finally, I thank Alex Todoran, who is nothing less than my brother.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Silviu Vlad Oprea*)

To Jesus, my Friend (John 15:15),
my Lord and my God (John 8:28).
And to Mary, the Theotokos (Luke 1:43),
and my dearest Mother (John 19:27).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

This thesis is concerned with the computational investigation of sarcasm, as manifested in a particular mean of expression, mainly online textual communication.

Sarcasm is a category of linguistic phenomena that exhibit characteristics challenging to encompass into a single explicative theory. Multiple theories have been suggested (Grice, 1975; Amante, 1981; Sperber and Wilson, 1981; Clark and Gerrig, 1984; Kreuz and Glucksberg, 1989; Utsumi, 1996; Wilson, 2006). They are discussed in Section 2.1. Among these, two characteristics are commonly pointed out. First, sarcasm occurs when there is a discrepancy between the literal and the intended meanings of an utterance. Second, through this discrepancy, the speaker of sarcasm expresses dissociation towards a particular state of affairs, often in the form of contempt or derogation.

Sarcasm is present in online communication. This is indicated by an increase in the number of computational investigations of online sarcasm across the scientific community in recent years. The survey of Băroiu and Trăușan-Matu (2022) shows this trend. In this thesis, we reflect upon these investigations. Pointing out their limitations, we bring contributions that span two research directions: computational sarcasm detection, and computational sarcasm understanding. For brevity, we refer to these directions henceforth as *sarcasm detection* and *sarcasm understanding*, respectively.

Investigations in the former direction, sarcasm detection, focus on building computational models optimised for the task of classifying texts as either sarcastic, or non-sarcastic. We refer to such models as *sarcasm detection* models. Building sarcasm detection models is motivated by the potential of sarcasm to conceal the intended meaning of a text. Concealing the intended meaning makes it challenging to extract useful signals from that text. For instance, the presence of sarcasm can be an obstacle to the accurate extraction of sentiment (Veale and Hao, 2010; Filatova, 2012; Reyes and Rosso, 2012; Maynard and Greenwood, 2014). Signals such as sentiment can drive crucial marketing, administration, and investment decisions (Medhat et al., 2014).

Investigations in the latter direction, sarcasm understanding, focus on explicating the phenomena that are subsumed under the category of sarcasm. This might include pointing out how sarcasm arises, the functions it serves, and the socio-demographic ecology of sarcastic exchanges. This is a direction that is less explored than the former. Nevertheless, advancing sarcasm understanding could implicitly advance other directions, including detection.

1.2 Research Questions

For each of the two research directions, we now point out limitations of previous investigations. In response to these limitations, we formulate five research questions. The contributions of this thesis arise in the process of addressing these research questions. There are two subsections below, each corresponding to one of the research directions.

Sarcasm Detection

Many sarcasm detection models introduced so far focus on lexical and pragmatic cues in the text being classified. Consider, for instance, Campbell and Katz (2012), Riloff et al. (2013), Joshi et al. (2016b), and Tay et al. (2018). However, sarcasm is a contextual phenomenon and detecting it in a text might require external information, including information about the author of that text. A limited amount of recent sarcasm detection models consider such information. See Wallace et al. (2015), Rajadesingan et al. (2015), Bamman and Smith (2015), Amir et al. (2016), and Hazarika et al. (2018). We refer to them as *contextual* models. These contextual models are built in a supervised learning paradigm, relying on datasets of texts labelled for sarcasm; that is, datasets of texts, where each text is labelled as either sarcastic, or non-sarcastic. Two methods of labelling texts for sarcasm have been suggested so far: *manual labelling* by human annotators; and *distant supervision*, where texts are considered sarcastic if they meet predefined criteria, such as including the token *#sarcasm*. However, contextual models have only been evaluated on distant supervision datasets. It is unclear if they generalise on datasets labelled manually. We suggest investigating this quantitatively by addressing the following research question.

RQ 1 Do contextual sarcasm detection models perform similarly on datasets labelled manually, and on datasets labelled using distant supervision?

We address this question in Chapter 3. We find the performance to differ significantly, suggesting that the two labelling methods might capture different phenomena. As such, it is unclear which method should be used for producing accurate sarcasm labels for a given text. That is, labels that coincide with the sarcastic intention of the author of that text. This motivates our next research question.

RQ 2 How can we create a dataset of texts labelled for sarcasm, where the label of a text captures the sarcastic intention of the author of that text?

We address this question in Chapter 4. We argue that previous labelling methods are suboptimal. We suggest a new method and use it to create iSarcasm, a dataset of tweets labelled for sarcasm. Here, *tweets* are posts from Twitter[1], a microblogging service where users interact through short posts. At the time of writing this thesis, the maximum length of a post is 280 characters.

We evaluate state-of-the-art sarcasm detection models on iSarcasm, and also collect third-party labels from human annotators. Models achieve a considerably lower performance on iSarcasm, compared to the performance they report on previous datasets. They also achieve a considerably lower performance, compared to human annotators. These performance discrepancies lead to our next research question.

RQ 3 How can we build more effective sarcasm detection models?
We address this question in Chapter 5.

Sarcasm Understanding

We formulate two research questions in the direction of sarcasm understanding.

We motivate the first question as follows. As mentioned above, sarcasm detection models achieve a considerably lower performance on iSarcasm, compared to human annotators. However, despite being higher than model performance, in our experiments, human performance is still less than 62%, quantified using the f-score. RQ 3 focuses on model performance. We now focus on human performance. The low f-score could indicate that sarcasm detection in text is challenging even for humans. Motivated by this observation, we switch our focus to studying sarcastic exchanges between human interlocutors. In the context of this thesis, by *sarcastic* exchange we mean an exchange that consists of one sarcastic utterance. We aim to determine the factors that could influence the ability of interlocutors to detect each other's sarcasm in such exchanges. We focus on four socio-demographic factors that characterise the interlocutors, mainly their age, gender, country, and English language nativeness. More specifically, we study the socio-demographic ecology of sarcastic exchanges between human interlocutors, in terms of these factors. That is, we aim to determine whether socio-demographic similarity between interlocutors, in terms of these factors, can influence their ability to detect each other's sarcasm. For instance, whether interlocutors of the same age are more able to detect each other's sarcasm, compared to interlocutors of different ages. Analysing previous studies of sarcasm in linguistics and sociolinguis-

[1] https://twitter.com

tics, we find that they support an affirmative answer. However, most of these studies draw their conclusions from qualitative analyses. There is a shortage of quantitative evidence in this direction, which leads to our next research question.

RQ4 Are interlocutors with similar socio-demographic backgrounds more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds?
We address this question in Chapter 6.

We formulate a second question in the direction of sarcasm understanding. It arises after reflecting upon a task recently introduced in the community, sarcasm generation (Joshi et al., 2015a; Mishra et al., 2019; Chakrabarty et al., 2020). That is, the task of creating dialogue systems, i.e. chatbots, able to generate sarcastic utterances. Approaches to sarcasm generation introduced so far are mainly motivated by the potential to create more approachable, human-like chatbots, considering that sarcasm is a natural part of human discourse. We suggest reconsidering this motivation, as a community, for two reasons. First, sarcasm is not a communicative goal in itself. Rather it is a device that can be used to achieve a wide variety of goals. Some of these goals, such as criticising and mocking, could cause offence. As such, they are likely undesirable in exchanges between humans and chatbots. Second, even if a machine seeks potentially desirable goals, it is unclear whether sarcastic utterances have the same effect on humans when coming from chatbots. In response, we suggest it is imperative, not least from an ethical perspective, to consider the following research question.

RQ5 In what conversational context is it appropriate for a chatbot respond sarcastically, and how should it formulate sarcasm such that it is understood by humans?
We address this question in Chapter 7.

***

In summary, in this thesis we address the following five research questions:

RQ1 Do contextual sarcasm detection models perform similarly on datasets labelled manually, and on datasets labelled using distant supervision?

RQ2 How can we create a dataset of texts labelled for sarcasm, where the label of a text captures the sarcastic intention of the author of that text?

RQ3 How can we build more effective sarcasm detection models?

RQ4 Are interlocutors with similar socio-demographic backgrounds more able to detect each other's sarcasm, compared to interlocutors of dissimilar backgrounds?

RQ5 In what conversational context is it appropriate for a chatbot respond sarcastically, and how should it formulate sarcasm such that it is understood by humans?

1.3 Contributions

This thesis brings four main contributions to the community. They span the two research directions introduced above, sarcasm detection and sarcasm understanding. There are two subsections below, each listing contributions corresponding to a research direction.

Sarcasm Detection

Contribution 1: A new method of labelling texts for sarcasm, along with iSarcasm, a dataset created using this method. This contribution arises from the work described in Chapter 3 and Chapter 4, as we address RQ 1 and RQ 2.

- We analyse previous labelling methods, manual labelling and distant supervision. We argue that they might capture different phenomena, but neither phenomena necessarily coincides with sarcastic intention of the authors of the texts being labelled.
- We suggest a method that actively involves the authors in the labelling process. Using this method, we create iSarcasm, a dataset of tweets labelled for sarcasm. Apart from sarcasm labels, iSarcasm also contains socio-demographic information about the author of each tweet, mainly their age, gender, and country.
- We evaluate state-of-the-art sarcasm detection models on iSarcasm, and also collect third-party labels for the tweets in iSarcasm from human annotators. Models achieve a considerably lower performance, compared to both the performance they report on previous datasets, and to annotators. This suggests the need for more effective models.

- However, human performance is still low at less than 62% F-score, which could indicate that the sarcasm detection task is challenging even for humans.

Contribution 2: The co-organisation of a competition that invites the community to create more effective sarcasm detection models. This contribution arises from the work described in Chapter 5, as we address RQ 3.

- We crowdsource the task of building more effective sarcasm detection models at the 16th International Workshop on Semantic Evaluation.
- In this purpose, we collect further examples of texts labelled for sarcasm using the labelling method introduced above.

Sarcasm Understanding

Contribution 3: An analysis of the socio-demographic ecology of sarcastic exchanges between human interlocutors. This contribution arises from the work described in Chapter 6, as we address RQ 4.

- We bring quantitative evidence supporting the claim that interlocutors with similar socio-demographic backgrounds are more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds.
- We find the most influential factors to be age, English language nativeness, and gender.
- We suggest that future social analysis tools, when analysing a sarcastic text, should account for the socio-demographic factors that characterise the author of that text. We also indicate how such factors could be inferred when detecting sarcasm in Twitter data, if missing.

Contribution 4: Guidelines for dialogue systems concerning the appropriateness of generating sarcastic responses, and the formulation of such responses. This contribution arises from the work described in Chapter 7, as we address RQ 5.

- We introduce Max, a novel sarcastic response generator grounded in a formal linguistic theory of sarcasm.
- We use Max to generate sarcastic responses to a given set of input utterances, and ask human annotators to label each response on several dimensions, including appropriateness and sarcasticness.

- We find that people consider sarcastic responses from chatbots inappropriate for most inputs; they might be appropriate for inputs that have a positive sentiment, or elements of humour. But, even when considered appropriate, people still prefer non-sarcastic responses.
- We also find that pragmatic insincerity and emotional markers are linguistic devices that, when included in generated sarcastic responses, increase the chance of sarcasm being recognisable.

1.4 Outline

Figure 1.1 illustrates the structure of this thesis, along with the research questions addressed.

The rest of the thesis is organised as follows:

- In Chapter 2 we conduct a literature review of previous investigations of sarcasm. Our review spans linguistic and computing research.
- In Chapter 3 we address RQ 1. We suggest a contextual sarcasm detection model that we evaluate on datasets representative of both manual labelling and distant supervision, achieving state-of-the-art performance. However, we find the performance to be consistently higher on distant supervision datasets, suggesting that the two labelling methods might capture different phenomena. This work counts towards Contribution 1 of this thesis. It was published in Oprea and Magdy (2019).
- In Chapter 4 we address RQ 2. We begin by arguing that both previous labelling methods are suboptimal. We then suggest a new method and create iSarcasm. We evaluate state-of-the-art sarcasm detection models on iSarcasm, and also collect third-party sarcasm labels from human annotators. This work also counts towards Contribution 1 of this thesis. It was published in Oprea and Magdy (2020b).
- In Chapter 5 we address RQ 3. We describe a task that we published at the 16th International Workshop on Semantic Evaluation for crowdsourcing more effective sarcasm detection models, and review the submissions. This work counts towards the Contribution 2 of this thesis. It was published in Abu Farha et al. (2022a).
- In Chapter 6 we address RQ 4. We emulate sarcastic exchanges via the process of third-party annotation. Specifically, we collect sarcasm labels for tweets in

iSarcasm from third-party human annotators in different rounds. In each round, we vary the socio-demographic similarity between the annotators and the authors of those tweets. Analysing the annotations allows us to make a statement about the research question. We conclude by discussing the potential implications of our work for future social analysis tools. This work counts towards Contribution 3 of this thesis. It was published in Oprea and Magdy (2020a).

- In Chapter 7 we address the RQ 5. We introduce Max and observe the reaction of human annotators to sarcastic responses generated by Max. We conclude with a set of guidelines for future work in sarcasm generation. This work counts towards the Contribution 4 of this thesis. It was published partly in Oprea et al. (2021), and fully in Oprea et al. (2022).
- In Chapter 8 we summarise our answers to the five research question, we discuss the limitations of the approaches that we employed to address the questions, and propose future directions of research.

Note that we opted to aggregate the limitations and future work suggestions of individual chapters and present them all at the end of the thesis, in Chapter 8. We believe this would allow the reader to better appreciate each limitation and future work direction in the context of all others.

1.5 Publications

The following articles have been published while conducting the work described in this thesis:

- Silviu Vlad Oprea and Walid Magdy. 2019. Exploring Author Context for Detecting Intended vs Perceived Sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859. Association for Computational Linguistics. Detailed in Chapter 3 and amounting to Contribution 1 of this thesis.
- Silviu Vlad Oprea and Walid Magdy. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289. Association for Computational Linguistics. Detailed in Chapter 4 and amounting to Contribution 1 of this thesis.
- Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, Walid Magdy. 2022. iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of The 16th International Workshop on Semantic Evaluation*. Association for

Computational Linguistics. Detailed in Chapter 5 and amounting to Contribution 2 of this thesis.

- Silviu Vlad Oprea and Walid Magdy. 2020. The Effect of Sociocultural Variables on Sarcasm Communication Online. In *Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1)*, article 029 (May 2020), 22 pages. Association for Computing Machinery. Detailed in Chapter 6 and amounting to Contribution 3 of this thesis.
- Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2021. Chandler: An Explainable Sarcastic Response Generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–349. Association for Computational Linguistics. Detailed in Chapter 7 and amounting to Contribution 4 of this thesis.[2]
- Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Detailed in Chapter 7 and amounting to Contribution 4 of this thesis.

[2]Since publishing this paper, we renamed our sarcastic response generator from *Chandler* to *Max*.

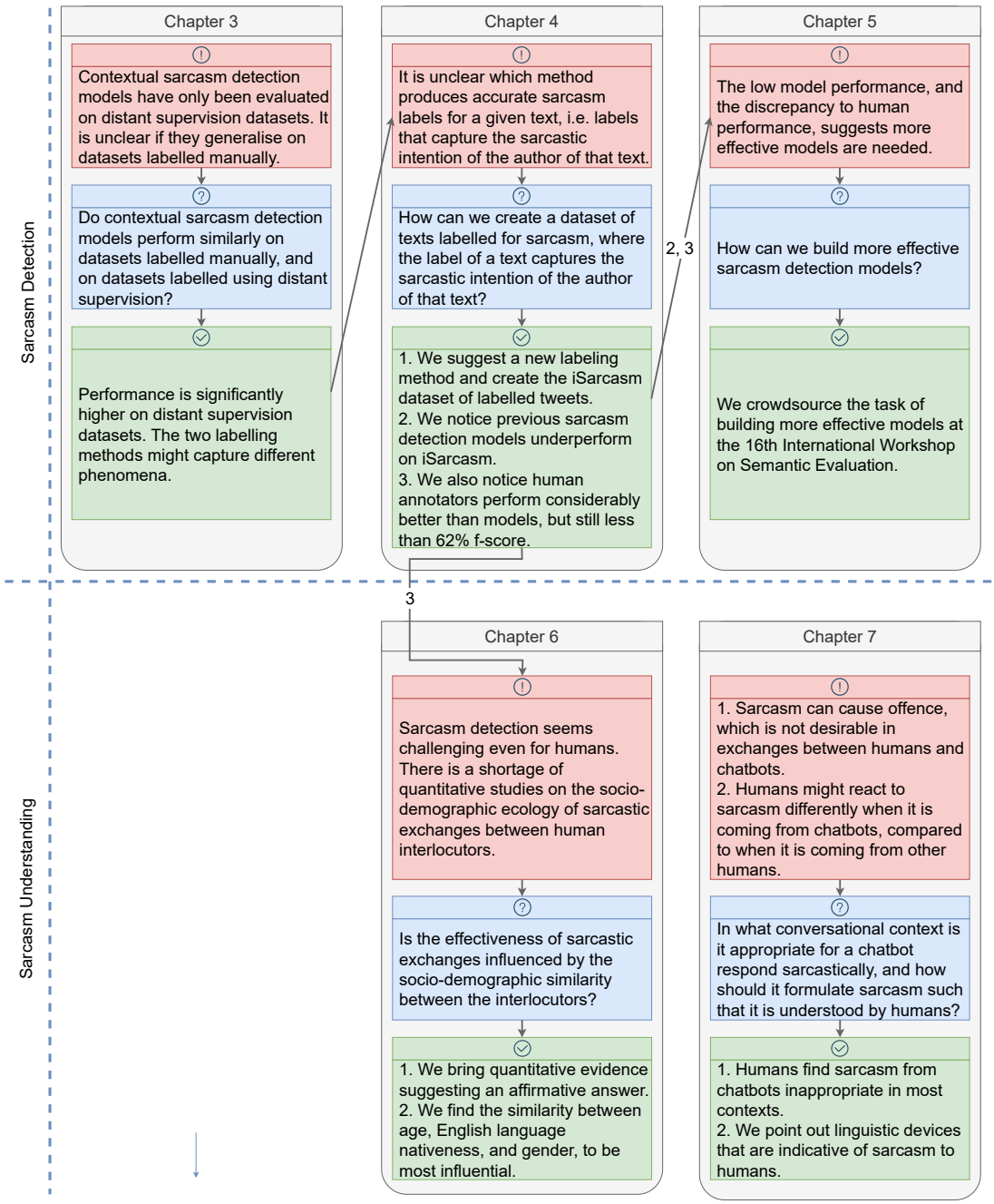

Figure 1.1: Thesis outline, as discussed in Section 1.4, including: summaries of limitations of previous work (red background, marked with ⓘ); research questions that arise in response to those limitations, along with chapters of this thesis where they are addressed (blue background, marked with ⓠ); and summaries of our answers to each question (green background, marked with ⓒ). The limitations, research questions, and answers, are grouped according to the research direction they belong to, out of sarcasm detection (upper half of the figure) and sarcasm understanding (lower half). The directions are described in Section 1.1.

Chapter 2

Background

In this chapter we introduce background information that will help towards making the thesis self-contained. Section 2.1 introduces theories of sarcasm suggested in linguistic literature. These are formal explications of the linguistic phenomena that are subsumed under the category of sarcasm. For instance, they might say how sarcasm arises in an utterance, or the purpose that sarcasm serves in human exchanges. Section 2.2 overviews previous attempts at building datasets of texts labelled for sarcasm, while Section 2.3 overviews previous sarcasm detection models.

2.1 Linguistic Theories of Sarcasm

In this section we overview linguistic theories of sarcasm. Note that in the context of this thesis, we use the term *sarcasm* to refer to the same category that most such theories refer to as *verbal irony*. We adopt this convention be consistent with computing literature, which indeed preponderantly uses the term *sarcasm*.

2.1.1 Gricean Theory

One of the first theories of sarcasm is provided by Grice, who views it as a flouting of the first maxim of Quality (Grice, 1975). Here, a flouting is a blatant violation that gives rise to a conversational implicature. In Grice's view, the speaker of a sarcastic utterance conversationally implicates the opposite of what they say. For instance, if the speaker of sarcasm says "What a great movie", they conversationally implicate that the movie was bad.

A main limitation of the Gricean view is that the flouting that gives rise to incongruity between what is said and what is implicated might not be necessary, and it is definitely not sufficient, for sarcasm to occur. To see that it might not be necessary, consider sarcastic understatements, such as saying "This was not the best movie ever" to mean the movie was bad. Granted, there is still a level of incongruity in that the literal meaning does not express the extent to which the speaker disapproves of the movie. But there is no opposition between the two meanings, as the Gricean theory assumes. To see that incongruity is not sufficient for sarcasm to occur, consider the fact that flouting of maxims could be employed to give rise to other phenomena as well. As an example, consider the metaphor "Time is money".

Despite this disadvantage, discussed in more detail by Sperber and Wilson (1981), there are instances of sarcasm that this view does explain, such as the one mentioned above, saying "What a great movie" to implicate that the movie was bad. We invite the interested reader to consult Grice (1975) for more examples, and more details on this view of sarcasm.

2.1.2 Indirect Speech Act Theory

Sarcasm can also be viewed through the lens of the speech act theory (Austin, 1962). Speech acts are acts performed by speaking, i.e. acts performed by the propositions that utterances express. Such acts include requesting, asking, promising, and blam-

ing. If the literal and non-literal meanings of an utterance are not performing the same acts, then the non-literal meaning is referred to as an indirect speech act (Searle, 1975). Searle formulates a set of *felicity* conditions that effective speech acts should meet (Searle and Searle, 1969). That is, a set of rules that must be met for a speech act to achieve its purpose. These include the *propositional content* condition, asking interlocutors to understand language, not merely act like they do; the *preparatory* condition, asking the speaker to have both evidence for the truth of the proposition P that they are about to make, and knowledge that the listener is not aware of P; the *sincerity* condition, asking that the speaker believes P to be true; and the *essential* condition, presupposing that the speaker intends the listener to act upon their utterance. Based on Searle's work, Amante (1981) sees sarcasm as blatant failure to satisfy one or more of these conditions. That is, sarcastic language is deceptive, but superficially so, the speaker intending to expose their infelicitous act to the listener.

As an example, consider the utterance "Brilliant job!" addressed by Alice to Bob, after looking in the oven to discover an overbaked cake that was forgotten there by Bob. Alice violates the preparatory felicity condition. Indeed, Alice has no reason to believe her proposition P, i.e. that it was brilliant that the cake was overbaked. In fact, she has evidence for the negation of P. She also knows that Bob considers the negation of P to be the case, rendering the literal sense of her utterance redundant. Alice also violates the sincerity condition. Her statement is, thus, infelicitous. She offers cues to her intentions, for instance, by using "brilliant", a word often used to construct hyperbole. The result is the construction of a latent meaning that stands in antithesis to the literal one, i.e. a critique about Bob's forgetfulness.

This view of sarcasm suffers from limitations similar to those of the Gricean view. First, violation of felicity conditions is not necessary for sarcasm to occur, as argued by Colston (2000) and Utsumi (2000). In this direction consider again sarcastic understatements. Violation is also not sufficient, as it does not provide grounds to discriminate between sarcasm and other indirect speech acts, such as metaphors.

2.1.3 Echoic Theories

Consider again the sarcastic utterance "What a great movie" spoken after a movie the speaker thought was bad. Sperber and Wilson (1981) offer a different account of sarcasm. They argue that the purpose of the sarcastic utterance cannot be to convey the belief that the movie was bad, since the belief can only be understood by the listener

from the utterance if the listener knows the utterance is sarcastic. However, the listener can only know it is sarcastic if they know the speaker's belief in advance. This makes the utterance completely uninformative if its purpose is to convey the speaker's belief about the movie. Instead, in the view of Sperber and Wilson (1981), the speaker is trying to convey a belief not about the movie, but about the utterance "What a great movie" itself. The utterance is an echoic mention of the speaker's initial expectation to see a good movie. This *echoic mention theory* of sarcasm explains why sarcastic utterances are made and why the meaning they implicate can be incongruous to the literal meaning. However, it does not differentiate between sarcastic and non-sarcastic echoic mentions. Kreuz and Glucksberg (1989) address this limitation by introducing the *echoic reminder theory* of sarcasm which adds the constraint that the echoic mention should always remind the listener of a violated social norm or a failed expectation.

2.1.4 Pretense Theories

Clark and Gerrig (1984) introduce the *pretense theory* of sarcasm. It claims that sarcasm arises in an exchange when the speaker pretends to be an injudicious person speaking to an imaginary uninitiated audience who would interpret their utterance literally. This way, the speaker expresses a negative attitude towards the pretended injudicious person, the imaginary audience, and the situation portrayed through their acting. The actual listener is expected to discover the pretense and this way understand the sarcasm. A variant of the pretense theory is considering sarcasm a pretense that the interlocutors jointly engage in. That is, they both pretend to perform a serious communication act in an imaginary situation. Their joint pretense that this situation is taking place is what generates sarcasm. An implication, that constitutes a main limitation of this approach, as pointed out by Utsumi (2000), is that the listener needs to share the sarcastic intention with the speaker beforehand, so that they (the listener) can engage in the joint pretense. Another limitation, shared with the original pretense theory, is the failure to distinguish between sarcastic and non-sarcastic pretense. An example of the latter is parody.

2.1.5 Implicit Display Theory

Utsumi (2000) argues that none of the theories discussed so far provides a complete account of sarcasm, in that the conditions they presuppose are neither necessary, nor sufficient, for sarcasm to occur.

As an alternative, they suggests the Implicit Display Theory (IDT) of sarcasm, which focuses on explaining how sarcasm is distinguished from non-sarcasm in an exchange.

The IDT first defines the concept of an ironic environment. We say a situation in which an utterance occurs is surrounded by an ironic environment if the discourse context includes the following components:

1. The speaker has expectation Q at time t_0;
2. Q fails at time $t_1 > t_0$; and
3. The speaker has a negative attitude towards the failure of Q.

In Utsumi (1996)'s view, such a situation within the discourse context facilitates the use of sarcasm. Note that the negative attitude could have several intensities, could be serious, or joking. Note also that the idea of linking sarcasm to an expectation is not new to Utsumi (1996), rather it is supported by previous work (Kreuz and Glucksberg, 1989; Kumon-Nakamura et al., 1995).

Next, according to the IDT, an utterance is sarcastic if and only if it is given in a situation surrounded by an ironic environment and it implicitly displays all three components of the ironic environment. Implicit display is realised if the following linguistic devices are present in the utterance:

1. allusion to the speaker's failed expectation Q;
2. pragmatic insincerity, realised by intentionally violating one of the pragmatic principles, e.g. Grice's maxims (Grice, 1975), or the felicity conditions for well-formed speech acts (Searle and Searle, 1969); and
3. implication (indirect expression) of the speaker's negative attitude towards the failure of Q.

Under the implicit display theory, the listener should assign the utterance a degree of sarcasm that is proportional to the degree to which the utterance achieves implicit display of the ironic environment. That is, sarcasm is a prototype-based category. The prototype is that instance of sarcasm which satisfies all three conditions of implicit display mentioned above. An utterance is more or less sarcastic based on how little or much it deviates from prototypical sarcasm. For instance, it might deviate by meeting only two out of the three conditions of implicit display.

2.2 Datasets of Sarcastic Texts

In this section we overview previous attempts at building datasets of texts labelled for sarcasm. That is, datasets of texts, where each text is labelled as either sarcastic, or non-sarcastic.

Two methods have been suggested so far to label texts for sarcasm: distant supervision and manual labelling. We describe each method in a subsection below, also mentioning corresponding datasets.

2.2.1 Distant supervision

This is the most common method. Texts are considered positive examples of sarcasm, i.e. are considered sarcastic, if they meet predefined criteria. This criteria includes: containing specific tags, such as *#sarcasm* for Twitter data, and */s* for Reddit data; and being posted by specific social media accounts. Texts that do not match the criteria are considered negative examples of sarcasm, i.e. are considered non-sarcastic. Table 2.1 gives an overview of datasets constructed this way, along with tags or accounts they associate with sarcasm.

The main advantage of distant supervision is that it allows building large labelled datasets with minimal effort. However, as we discuss in Section 4.2.1, the labels produced can be noisy.

2.2.2 Manual labelling

An alternative to distant supervision is collecting texts and presenting them to human annotators for labelling. Filatova (2012) asks annotators to find pairs of Amazon reviews where one is sarcastic and the other one is not, collecting 486 positive and 844 negative examples. Abercrombie and Hovy (2016) annotate 2,240 Twitter conversations, ending up with 448 positive and 1,732 negative labels, respectively. Riloff et al. (2013) use a hybrid approach, where they collect a set of 1,600 tweets that contain #sarcasm or #sarcastic, and another 1,600 without these tags. They remove such tags from all tweets and present the tweets to a group of human annotators for final labelling. We call this the *Riloff* dataset. A similar approach is employed by Van Hee et al. (2018). They build a balanced dataset of 4,792 tweets. We call it the *SemEval-2018* dataset.

2.3 Sarcasm Detection Models

In this section we overview previous attempts at building sarcasm detection models. For a more exhaustive overview, and details beyond what we provide here, we invite the interested reader to consult the reviews of Băroiu and Trăușan-Matu (2022) and Moores and Mago (2022).

Based on the information considered when classifying a given text as either sarcastic, or non-sarcastic, we identify two categories of models across literature, that we refer to as *local models* and *contextual models*. We overview each category in a subsection below.

2.3.1 Local Models

Local models only consider information available within the text being classified.

One set of approaches first extract a set of predetermined features from the given text, then provide those features to downstream classification algorithms. For instance, Tsur et al. (2010) introduce SASI, a Semi-supervised Algorithm for Sarcasm Identification. It first extracts two types of features: syntactic features, which are punctuation-based; and pattern-based features. It then uses the k-nearest neighbours algorithm for classification. Davidov et al. (2010) use a similar approach. Veale and Hao (2010) analyse web-harvested similes to identify characteristics that separate sarcastic similes from non-sarcastic ones. From this analysis they suggest a rule-based algorithm for classification. González-Ibáñez et al. (2011) aim to detect sarcasm in tweets. Given a tweet, they extract features such as: unigrams; the presence of positive emojis, such as smileys; the presence of negative emojis, such as frowning faces; and whether the tweet is a reply to another tweet. For classification, they experiment with support vector machines and logistic regression.

Another set of approaches consider linguistic incongruity (Campbell and Katz, 2012) to be a marker of sarcasm. To determine linguistic incongruity, Riloff et al. (2013) look for the presence of a positive verb used in a negative sentiment context. Joshi et al. (2016b) first extract vector representations of the words in the text, also known as word embeddings. Then, they identify linguistic incongruity between pairs of words in the text with the cosine similarity between the corresponding vector representations. Tay et al. (2018) use a neural network with a self-attention mechanism (Vaswani et al., 2017). They identify incongruity between pairs of words with

the attention scores between the corresponding vector representations of those words.

2.3.2 Contextual Models

Contextual models utilize both information that originates in the text being classified, and contextual information that originates outside the text. There is a limited amount of work in this direction. Wallace et al. (2015) aim to detect sarcasm in Reddit data. When classifying a post, they include information about the subreddit where that post was found. Rajadesingan et al. (2015) and Bamman and Smith (2015) aim to detect sarcasm in Twitter data. When classifying a tweet, they include contextual information about the user who posted that tweet. This consists of manually-curated features extracted from the historical tweets of that user. Amir et al. (2016) merge all historical tweets of that user into one historical document and use the Paragraph Vector model (Le and Mikolov, 2014) to build a representation of that document. Building on their work, Hazarika et al. (2018) extract, in addition, personality features from the historical document using a model pre-trained on a personality detection benchmark corpus. The corpus is that published by Matthews and Gilliland (1999), containing essays labelled with the Big-Five personality traits, i.e. Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

article	data source	labelling criteria
Davidov et al. (2010)	Twitter	tags: #sarcasm, #sarcastic, #not
Barbieri et al. (2014b)	Twitter	tags: #sarcasm, #education, #humor, #irony, #politics
Ptáček et al. (2014)	Twitter	tags: #sarcasm, #sarcastic, #irony, #satire
Bamman and Smith (2015) Joshi et al. (2015b)	Twitter	tags: #sarcasm, #sarcastic
González-Ibáñez et al. (2011) Reyes and Rosso (2012) Liebrecht et al. (2013) Bouazizi and Ohtsuki (2015) Bharti et al. (2015)	Twitter	tags: #sarcasm
Barbieri et al. (2014a)	Twitter	accounts: @spinozait,@LiveSpinoza
Khodak et al. (2018)	Reddit	tags: /s

(a) Datasets labelled via distant supervision, as discussed in Section 2.2.1. In the table, when "labelling criteria" is a list of tags, it refers to the fact that, when building the datasets described by the articles under "article", texts were considered sarcastic if they included those tags. When it is a list of accounts, it refers to the fact that texts posted by the respective social media accounts were considered sarcastic.

article	data source	data format
Riloff et al. (2013) Benamara et al. (2017) Cignarella et al. (2018) Van Hee et al. (2018) Bueno et al. (2019)	Twitter	labelled tweets
Abercrombie and Hovy (2016)	Twitter	labelled tweet-reply pairs
Filatova (2012)	Amazon	labelled product reviews

(b) Datasets labelled manually, or using a combination of distant supervision and manual labelling, as discussed in Section 2.2.2.

Table 2.1: Datasets for sarcasm detection suggested in previous work, labelled using either: (a) distant supervision (Table 2.1a); or (b) manual labelling, or a combination of the two methods(Table 2.1b); as discussed in Section 2.2.

Chapter 3

Exploring Author Context for Detecting Intended vs Perceived Sarcasm

In this chapter we address RQ 1 of this thesis. The work presented herein, together with that presented in Chapter 4, count towards Contribution 1 of this thesis. This chapter is based on the paper "Exploring Author Context for Detecting Intended vs Perceived Sarcasm" that we published in the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Oprea and Magdy, 2019).

3.1 Introduction

Many sarcasm detection models introduced so far consider lexical and pragmatic cues in the text that is being classified. We mentioned some of these models in Section 2.3.1. However, as dramatically pointed out by (Wilson and Sperber, 1992) in an article that introduces a particular theory of sarcasm that we overview in Section 2.1.3, "the linguistic form of an utterance grossly underdetermines its interpretation". Indeed, lexical and pragmatic cues that originate in the text being classified might not be sufficient for disambiguating the sarcastic intention of the author of that text. Rather, their intention might only be made apparent when considering contextual information that originates outside the text, including information about the author.

In response, a limited amount of recent sarcasm detection models aim to consider such information, reporting state-of-the-art results. We discussed these models in Section 2.3.2. Such contextual models are built in a supervised learning paradigm. Specifically, they are parametric functions that input the numerical encoding of a given text. The output is interpreted as the probability that the input text is sarcastic. A maximum likelihood approach is usually used for training, i.e. parameter estimation, given a dataset of texts, each labelled as either sarcastic, or non-sarcastic. As such, the effectiveness of these models depends on the availability and quality of training datasets.

Two methods of labelling texts for sarcasm have been suggested so far: manual labelling by human annotators; and distant supervision. When distant supervision is used, texts are considered sarcastic if they meet predefined criteria. One example of such a criterion is to consider sarcastic those texts that include a specific token. When labelling posts from the Twitter platform, this string is often *#sarcasm* or *#sarcastic*; when labelling posts from the Reddit platform, it is usually */s*. For more details, please consult Section 2.2, where we overviewed sarcasm detection datasets introduced by previous work.

However, despite reporting encouraging results, to our knowledge, contextual sarcasm detection models introduced so far have been evaluated on distant supervision

datasets. It is unclear if they generalise when trained on datasets labelled manually. We suggest investigating this quantitatively by addressing the first research question of this thesis.

RQ 1 Do contextual sarcasm detection models perform similarly on datasets labelled manually, and on datasets labelled using distant supervision?

To address this question, we consider datasets representative of both labelling methods. We then suggest several novel contextual sarcasm detection models. In line with previous contextual models mentioned above, the models we suggest, when classifying a tweet, consider both the tweet itself, and information about the author of that tweet. In particular, we identify such author information with a single vector representation of the historical tweets of that author. Our models differ in how that representation is constructed. We use the term *user* to refer to the author of a tweet and the phrase *user embedding* to refer to such a representation.

In this scenario, we suggest answering RQ 1 by addressing the following, more granular, research questions:

RQ 1.1 Given a tweet t posted by user u^t with user embedding e^t, is e^t predictive of the sarcastic nature of t?

RQ 1.2 Is the predictive power of e^t on the sarcastic nature of t the same if t is labelled manually, compared to when t is labelled via distant supervision?

While the models we suggest outperform previous models on all datasets, we show that user embeddings are considerably more predictive of distant supervision labels, than of manual labels.

The rest of this chapter is organised as follows. Section 3.2 introduces the datasets that we use for testing our contextual models. Section 3.3 introduces the models. Section 3.4 reports the experiments we conducted to answer our research questions. Section 3.5 summarises the chapter, states the answers to the granular research questions of this chapter, and motivates the work conducted in the next chapter.

3.2 Sarcasm Datasets

We test the contextual models that we suggest on two popular datasets of tweets labelled for sarcasm, the Riloff dataset (Riloff et al., 2013) dataset and the Ptacek (Ptáček et al., 2014) dataset.

dataset	size	sarcastic	non-sarcastic	training	validation	test
Riloff	701	192	509	551	88	62
Ptacek	27,177	15,164	12,013	21,670	2,711	2,797

Table 3.1: Label distribution across our datasets, and distribution into training, validation and test sets, as discussed in Section 3.2.

3.2.1 Riloff dataset

The Riloff dataset was published as a list of 3,200 tweet IDs. The corresponding tweets were manually labeled by third-party annotators. Three separate labels were collected for each tweet, and the dominant one was chosen as the final label.

Using the Twitter API[1], we attempted to collect the corresponding tweets, as well as the historical timeline tweets for each user, to be used later for building user embeddings. For a user with tweet t in Riloff, we collected those historical tweets posted before t. Only 701 original tweets, along with the corresponding user timelines, could be retrieved. Others have either been removed from Twitter, the corresponding user accounts have been disabled, or the API did not retrieve any historical tweets.

Table 3.1 shows the label distribution across this dataset. We divided the dataset into ten buckets, using eight for training, one for validation and one for testing. The division into buckets was stratified by users, i.e. all tweets from a user ended up in the same bucket. Stratification makes sure any specific embedding is only used during training, during validation, or during testing. We further ensured the overall class balance was represented in all of the three sets. Table 3.1 shows the size of each set.

3.2.2 Ptacek dataset

The Ptacek dataset consists of 50,000 tweet IDs labelled via distant supervision. Tags used as markers of sarcasm were #sarcasm, #sarcastic, #satire and #irony. These labels rely on signals provided by the users who posted the tweets.

In a similar scenario as with the Riloff dataset, we could only collect 27,177 tweets and corresponding timelines. We divided them into ten buckets and stratified by users. During preprocessing, we removed all sarcasm-marking tags from both the training tweets and the historical tweets. Table 3.1 shows statistics on both datasets.

[1] https://developer.twitter.com

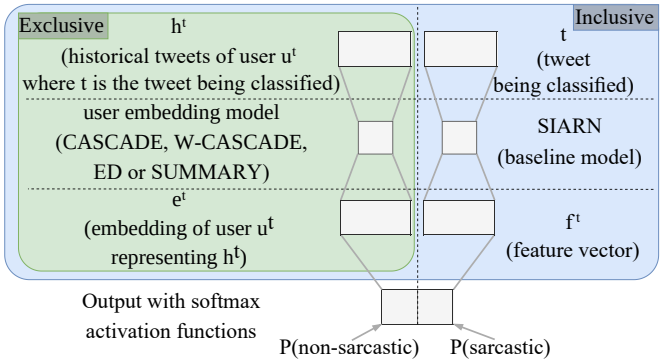

Figure 3.1: The architecture of the models used, as discussed in Section 3.3. Exclusive models do not use the current tweet being classified, prediction being based solely on user history. Inclusive models use both user history and the current tweet.

3.3 Contextual Sarcasm Detection Models

Let T be a set of tweets. For any $t \in T$, let u^t be the user who posted tweet t. Let h^t be a set of historical tweets of user u^t, posted before t, with $h^t \cap T = \emptyset$ and let e^t be the embedding of user u^t, i.e. a vector representation of h^t. Let $Y = \{\text{sarcastic}, \text{non-sarcastic}\}$ be the output space. Our goal is to find a model $m : \{(t, e^t) | t \in T\} \rightarrow Y$.

As a baseline, we implement the SIARN (Single-Dimension Intra-Attention Network) model proposed by Tay et al. (2018), since, at the time of conducting this work, achieves the best published results on both our datasets. SIARN only looks at the tweet being classified, that is $\text{SIARN}(t, e^t) = m'(t)$.

Further, we introduce two classes of models: exclusive and inclusive models. In exclusive models, the decision whether $t \in T$ is sarcastic or not is independent of t, i.e. $m(t, e^t) = m'(e^t)$. The content of the tweet being classified is not considered, prediction being based solely on user historical tweets. The architecture of such a model is shown in Figure 3.1. We feed the user embedding e^t to a layer with softmax activations to output a vector of probabilities over Y. We name these models EX-[emb], where [emb] is the name of the user embedding model.

Inclusive models account for both t and e^t, as shown in Figure 3.1. We start with the feature vector f^t extracted by SIARN from t. We then concatenate f^t with e^t and use an output layer with softmax activations. We name these models IN-[emb], where [emb] is the user embedding model. We now look at several user embedding models that build e^t for a user u^t as a representation of h^t.

CASCADE Embeddings To our knowledge, the user embedding model that has proven most informative in a sarcasm detection pipeline at the time of conducting this work was CASCADE (Hazarika et al., 2018). However, it has only been tested on a dataset of Reddit posts labelled via distant supervision. We test it on our datasets. Following original authors, we merge all tweets from h^t in a single document d^t, giving corpus $C = \{d^t | t \in T\}$. Using the Paragraph Vector model (Le and Mikolov, 2014) we generate a representation v^t of d^t. Next, we feed d^t to a neural network pre-trained on the personality detection corpus released by Matthews and Gilliland (1999). We merge the resulting hidden state p^t of the network with v^t using Generalized Canonical Correlation Analysis (GCCA) as described by Hazarika et al. (2018) to get e^t.

W-CASCADE Embeddings CASCADE treats all historical tweets in the same manner. However, long-term working memory plays an important role in verbal reasoning and textual comprehension (Kellogg, 2001). We therefore expect recent historical tweets to have a greater influence on the current behaviour of a user, compared to older ones. To account for this, we suggest the following model that accounts for the temporal arrangement of historical tweets. We first use CASCADE to build, for each historical tweet $r \in h^t$, two vectors v^t_r and p^t_r, and then merge them into e^t_r using GCCA. Recall that t is the tweet being classified and h^t are historical tweets preceding t. We then divide the sequence $\langle e^t_{r_1}, e^t_{r_2}, \ldots, e^t_{r_{|h^t|}} \rangle$ of representations of historical tweets $r_1, r_2, \ldots, r_{|h^t|}$ into ten contiguous partitions and multiply each representation with the index of the partition it belongs to. That is, we multiply $e^t_{r_i}$ by $i \ \% \ |h^t| + 1$, where $\%$ is the modulus operator. By convention, the tweet with the highest index is the most recent one. Finally, we sum the resulting vectors and normalize the result to get e^t.

ED Embeddings One of the main advantages of the encoder-decoder model (Sutskever et al., 2014), commonly used for sequence prediction tasks, is its ability to handle inputs and outputs of variable length. The encoder, a recurrent neural network, transforms an input sequence into an internal representation of fixed dimension. The decoder, another recurrent network, generates an output sequence using this representation. We use bi-directional LSTM cells (Schuster and Paliwal, 1997) and identify $e^t_{r_i}$, $1 \leq i \leq |h^t|$, with the internal state of the encoder after feeding in r_i. The training objective is to reconstruct the input r_i. We employ the same weighting technique as we did for W-CASCADE to construct e^t.

SUMMARY Embeddings We use an encoder-decoder model as in the previous paragraph, but change the objective from reconstructing the input to summarizing it. We use the model pre-trained on the Gigaword standard summarization corpus[2].

3.4 Experiments

In this section we review our experimental results and discuss their implications.

3.4.1 Experimental Setup

In a data preprocessing step, we filter out all tweets shorter than three words and replace all words that only appear once in the entire corpus with the UNK token. Then, we encode each tweet as a sequence of word vectors initialized using GloVe embeddings (Pennington et al., 2014). Following the authors SIARN, our baseline, we set the word embedding dimension to 100. We tune the dimension of all CASCADE embeddings to 100 on the validation set. For comparability, we set W-CASCADE embeddings to the same dimension. For CASCADE embeddings we make use of the implementation available at `https://github.com/SenticNet/cascade`. When training ED and SUMMARY, our decoder implements attention over the input vectors. We use the general global attention mechanism suggested by Luong et al. (2015). We implement both ED and SUMMARY using the OpenNMT toolkit (Klein et al., 2017).

For comparability with SIARN, our baseline, we follow its authors in setting a batch size of 16 for the Riloff dataset, and of 512 for the Ptacek dataset, and in training for 30 epochs using the RMSProp optimizer (Tieleman and Hinton, 2012) with a learning rate of 0.001.

3.4.2 Results

All results are reported in Table 3.2. We achieve state-of-the-art results on both datasets. However, user embeddings show high predictive power on the Ptacek dataset, but only marginal on the Riloff dataset. On the one hand, the EX-W-CASCADE model achieves higher performance (f1-score of 0.922) than the baseline (f1-score of 0.863) on Ptacek; that is without even looking at the tweet being classified. On the other hand, EX-W-CASCADE achieves an f1-score of only 0.478 on Riloff. Out of the exclusive models,

[2] `https://github.com/harvardnlp/sent-summary`

	Model	Riloff	Ptacek
	SIARN (baseline)	0.711	0.863
exclusive	EX-CASCADE	0.457	0.802
	EX-W-CASCADE	0.478	**0.922**
	EX-ED	**0.546**	0.873
	EX-SUMMARY	0.492	0.845
inclusive	IN-CASCADE	0.723	0.873
	IN-W-CASCADE	0.714	**0.934**
	IN-ED	**0.739**	0.887
	IN-SUMMARY	0.679	0.892

Table 3.2: F1-score achieved on the Riloff and Ptacek datasets for both exclusive and inclusive models, as discussed in Section 3.4.2.

Model	Riloff	#Riloff
EX-CASCADE	0.457	0.818
EX-W-CASCADE	0.478	0.797
EX-ED	**0.545**	**0.827**
EX-SUMMARY	0.492	0.772

Table 3.3: F1-score achieved by the exclusive models on the #Riloff dataset, compared to the Riloff dataset, as discussed in Section 3.4.2.

the highest f1-score of 0.546 on Riloff is achieved by EX-ED. By contrast, EX-ED achieves an f1-score of 0.873 on Ptacek.

As such, user embeddings are considerably less predictive on the Riloff dataset, labelled manually, than on the Ptacek dataset, labelled via distant supervision. However, one could wonder if there are differences between the two datasets other than the labelling method used, unaccounted for by us, that might have influenced our results. Fortunately, we noticed that many of the tweets in Riloff itself contain one or more of the tags that were used to mark sarcasm in Ptacek. For all tweets in Riloff, we checked the agreement between containing such a tag, and being manually annotated as sarcastic. The results are shown in Table 3.4. Note that the statistics shown are not for the entire dataset, as published by Riloff et al. (2013), but for the subset of tweets coming from users from which we could gather historical tweets, as discussed

	with tag	without any tag
labelled sarcastic	190	2
labelled non-sarcastic	217	292

Table 3.4: Disagreement between manual labels and the presence of sarcasm tags in the Riloff dataset, as discussed in Section 3.4.2.

in Section 3.2. We notice a large disagreement. In particular, 217 out of the 509 tweets that were annotated manually as non-sarcastic contained such a tag. To investigate further, we re-labelled the Riloff dataset via distant supervision considering these tags as markers of sarcasm, to create the #Riloff dataset. We then trained exclusive models on #Riloff. Results are reported in Table 3.3. As before, user embeddings are considerably less predictive of manual labels, present in Riloff, than on distant supervision labels, present in #Riloff.

3.5 Summary

We studied the performance of contextual models that use user embeddings for the task of textual sarcasm detection, considering both datasets labelled manually, and datasets labelled via distant supervision. We suggested neural models to build user embeddings, achieving state-of-the-art performance.

We conducted further analyses that lead to the following answers to our granular research questions introduced in Section 3.1. To RQ 1.1 we answer that when classifying a tweet as either sarcastic, or non-sarcastic, the user embedding is highly predictive of the sarcastic nature of that tweet, insofar as that nature is captured by the distant supervision criteria. This is to the extent that considering the tweet itself brings marginal improvement to the classification performance. This suggests that users have a prior disposition to being either sarcastic, or non-sarcastic, which can be deduced from historical behaviour, insofar as the distant supervision criteria captures their sarcastic intention. However, this behaviour could fluctuate over time, as suggested by the fact that we achieve higher performance when accounting for the temporal arrangement of historical tweets, as we do with W-CASCADE. To RQ 1.2 we answer that user embeddings are considerably more predictive of distant supervision labels than of manual labels. In this setting, to RQ 1 we answer that, insofar as context is represented by user embeddings, contextual sarcasm detection models perform considerably better on

datasets labelled via distant supervision, than on datasets labelled manually. A manual analysis of the Riloff dataset further underlined a discrepancy between the labels produced by the two methods.

Given the discrepancy, one might wonder which method should be used for producing accurate labels for a text. That is, labels that coincide with the sarcastic intention of the author of that text. One might argue that distant supervision is a viable candidate, considering that it relies on signals provided by the author. However, as we discuss in Section 4.2.1, relying on such signals might lead to noisy labels, in terms of both false positives and false negatives. In Chapter 4, we suggest a new labelling method and use it to construct a new dataset.

Chapter 4

iSarcasm: A Dataset of Intended Sarcasm

In this chapter we address RQ 2 of this thesis. The work presented herein, together with that presented in Chapter 3, count towards Contribution 2 of this thesis. This chapter is based on the paper "iSarcasm: A Dataset of Intended Sarcasm" that we published in the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Oprea and Magdy, 2020b).

4.1 Introduction

The effectiveness of sarcasm detection models that are built in a supervised learning paradigm depends on the availability and quality of training datasets. As discussed in Section 2.2, two methods of labelling texts for sarcasm have been suggested so far, manual labelling and distant supervision. In Section 3.4.2, when analysing the Riloff dataset, we showed that these methods could produce discrepant labels. Within the same chapter, in addressing RQ 1 of this thesis, we showed that contextual sarcasm detection models that use user embeddings perform considerably higher on distant supervision datasets, than on datasets labelled manually.

Given the discrepancy, it is unclear which labelling method, if any, should be used for producing accurate labels for a text. That is, labels that capture the sarcastic intention of the author of that text. In response, we formulate the second research question of this thesis.

RQ 2 How can we create a dataset of texts labelled for sarcasm, where the label of a text captures the sarcastic intention of the author of that text?

In this chapter, we answer RQ 2 and conduct further analyses, addressing the following, more granular, research questions:

RQ 2.1 What are the limitations of manual labelling and distant supervision, in terms of capturing the sarcastic intention of the authors of the texts that they are labelling?

RQ 2.2 What method of labelling a text for sarcasm would capture the sarcastic intention of the author of that text?

RQ 2.3 What is the performance of previous sarcasm detection models on a dataset labelled with such a method?

We begin with an overview of the ways in which previous labelling methods are sub-optimal. A main determinant is their reliance on labelling *proxies*, either in the form of third-party annotators, in the case of manual labelling, or predefined criteria, in the

case of distant supervision. This could limit not only their accuracy, producing both false positives and false negatives, but also their coverage. Here, by *coverage* we mean the number of different phenomena that are subsumed under the umbrella of sarcasm that these methods are able to capture. An overview of such phenomena is provided in Section 2.1, where we discuss linguistic theories of sarcasm.

We then suggest a labelling method that attempts to remove proxies; when labelling a text, our method involves the author of that text in the labelling process. Using this method, we build iSarcasm, a dataset of tweets labelled for sarcasm. More specifically, we publish a survey on a crowdsourcing platform. In the survey, we ask Twitter users to provide both sarcastic and non-sarcastic tweets that they have posted in the past. For each sarcastic tweet, we also ask them to explain why it was sarcastic and how they would convey the same meaning non-sarcastically.

Next, we collect third-party sarcasm labels for the tweets in iSarcasm from workers on the same crowdsourcing platform. Third-party annotation for sarcasm has been conducted before, as discussed in Section 2.2, but most such studies do not check the ability of the annotators to capture sarcasm in texts as intended by the authors of those texts. In our experiment, annotators achieve an f1-score of only 0.616, showing that sarcasm detection is challenging even for humans.

Next, we implement previous sarcasm detection models (Tay et al., 2018; Hazarika et al., 2018; Van Hee et al., 2018) that achieve state-of-the-art performance on previous datasets, and evaluate their performance on iSarcasm. While these models report f1-scores reaching 0.874 on previous datasets, their maximum f1-score on iSarcasm is 0.364, suggesting that previous datasets might be biased or obvious. This highlights the importance of developing new approaches for sarcasm detection.

iSarcasm contains 4,484 English tweets. Each tweet has an associated sarcasm label provided by its author, with a ratio of roughly 1:5 of sarcastic to non-sarcastic tweets.

The rest of this chapter is organised as follows. Section 4.2 discusses the limitations of current labelling methods, manual labelling and distant supervision, that make them suboptimal for capturing sarcasm in a text, as intended by the author of that text. Section 4.3 describes a new labelling method and how we used it to construct the iSarcasm dataset of tweets labelled for sarcasm. Section 4.4 shows an exploratory analysis of iSarcasm, and Section 4.5 reports and analyses the performance of third-party human annotators at detecting sarcasm in our dataset, and that of previous state-of-the-art sarcasm detection models. Section 4.6 summarises the chapter, states the answers to

the granular research questions of this chapter, and motivates the work conducted in the next chapter.

4.2 Limitations of Current Labelling Methods

In this section, we discuss limitations of current labelling methods that make them suboptimal for capturing sarcasm in a text, as intended by the author of that text. In doing so, we address RQ 2.1 of this chapter.

4.2.1 Limitations of Distant Supervision

We introduced distant supervision in Section 2.2.1, where we also mentioned representative datasets. Distant supervision labels a text based on signals provided by the author of that text, such as the inclusion of specific tags. As such, it might seem like a viable candidate for capturing the sarcastic intention of that author. However, relying on such signals might lead to noisy labels.

First, the tags might not necessarily mark sarcasm, but rather constitute the subject or object of conversation. For instance, when labelling tweets, a tag such as "#sarcasm" could be appended to clarify that a previous tweet should have been read sarcastically, as illustrated in the utterance "@user it was #Sarcasm" provided by Davidov et al. (2010). This could lead to false positive labels. Second, when using tags such as "#politics" and "#education", as done by Barbieri et al. (2014b), there is a strong underlying assumption that such tags are always accompanied by sarcasm, potentially generating further false positive labels. The assumption that certain social media accounts always publish sarcastic posts, as assumed by Barbieri et al. (2014a), is similarly problematic. Third, the sarcasm that distant supervision does capture might be of a specific flavor, such that the inclusion of a tag might be essential to ensure inferrability (Davidov et al., 2010). This could lead to false negative labels. A model trained on such a dataset might be biased towards that specific flavour of sarcasm, being limited in its ability to generalise to other flavours. Forth, under distant supervision, if a text does not meet the predefined criteria, it is considered non-sarcastic. This is a strong assumption that can lead to false negatives. Indeed, as discussed in Section 4.4.1, none of the sarcastic tweets in iSarcasm contain hashtags used as distant supervision criteria when constructing most previous datasets.

4.2.2 Limitations of Manual labelling

We introduced manual labelling in Section 2.2.2, where we also mentioned representative datasets. Under this labelling method, labels for a text are collected from third-party annotators. They are usually asked to discern sarcasm based solely on cues found within the text.

However, such cues are often not informative enough for discerning the sarcastic intention of the author of that text (González-Ibáñez et al., 2011). Sarcasm is a pragmatic phenomenon (Kreuz and Caucci, 2007) and its comprehension could lie in mechanisms that are outside pure language processing (Colston and Gibbs Jr., 2007). In such cases, in order to discern the sarcastic nature of a text, it needs to be interpreted in light of the contextual factors, such as factors that characterise the author of that text.

For instance, cultural factors might even influence the author's opinion on what phenomena are subsumed under the umbrella of sarcasm, that is, the definition of sarcasm. To illustrate this point, consider the following experiments. Dress et al. (2008) ask participants to define sarcasm. They notice that some participants define it as involving more humour than others. Veale and Hao (2007) harvest similes from the web, noticing that 18% of them contain irony. By contrast, Veale et al. (2008) show that, among Chinese similes, only 3% to 4% are ironic. Rockwell and Theriot (2001) compare the usage of sarcasm across members of collectivist cultures, such as Japanese, Chinese, or Thai, and members of individualist cultures, such as American, British, or German. They notice that collectivists tend to use sarcasm less often, and formulate it more subtly, compared to individualists. Joshi et al. (2016a) conduct an annotation experiment on the Riloff dataset (Riloff et al., 2013). They present the dataset, initially manually labelled for sarcasm by American annotators, to be labelled by Indian annotators who are trained linguists. They find higher disagreement between Indian and American annotators, than between annotators of the same nationality. Furthermore, they find higher disagreement between pairs of Indian annotators, indicating higher uncertainty, than between pairs of American annotators. They attribute these results to cultural differences between India and the United States. They conclude that sarcasm annotation expands beyond linguistic expertise and is dependent on considering such factors. Finally, there might even be cultures who do not use sarcasm at all, as mentioned by Haiman (1989) about the Hua, a group of people from Papua New Guinea.

As such, consider the scenario when an annotator, who is asked to discern the sarcastic nature of a text, does not share cultural common ground with the author of

that text. In this scenario, the annotator might not perceive the text as sarcastic, despite being intended as such by its author. Similarly, the annotator might perceive the text as sarcastic, in light of their cultural assumptions, despite not being intended as such by its author.

But even when the annotator does share cultural common ground with the author, there are personal factors that could make the annotator misjudge the sarcastic nature of the text. For instance, their mood and level of stress at the time of annotation (Reyes et al., 2012).

All in all, when an annotator labels a text for sarcasm, the annotator might have different cultural assumptions about sarcasm than the author of that text; or might be influenced by personal factors; both of which could interfere with their ability to recognise the sarcastic intention of the author. As such, annotator perception, i.e. *perceived sarcasm*, might differ from author intention, i.e. *intended sarcasm*. This could potentially lead to both false positive and false negative labels.

4.2.3 Removing Proxies

When labelling a text for sarcasm, both methods discussed above use a *proxy* through which they attempt to recover the sarcastic intention of the author of that text. This is in the form of predefined criteria, in the case of distant supervision, or third-party annotators, in the case of manual labelling. The proxy could induce noise in the form of both false positive and false negative labels. In what follows, we aim to remove such proxies, creating a dataset of tweets labelled for sarcasm by involving the authors of those tweets in the labelling process.

4.3 A Method of Capturing Intended Sarcasm

4.3.1 Collecting Sarcastic Tweets

The work described in this section addresses RQ 2.2 of this chapter. We designed an online survey where we asked Twitter users to provide links to one sarcastic and three non-sarcastic tweets that they had posted in the past, either on their timeline, or as replies to other tweets. The sarcasm label for each tweet was, thus, implicitly provided by the author of that tweet. We refer to such labels as *intended sarcasm* labels. We required that no retweets should be provided, and that tweets should not

include references to multimedia content or, if such content was referred, it should not be informative in judging sarcasm. For each sarcastic tweet, we asked users to also provide, in full English sentences, an *explanation* of why it was sarcastic and a *rephrase* that would convey the same message non-sarcastically. This way, we aimed to prevent them from misjudging the sarcastic nature of their previous tweets under experimental bias. Finally, we asked for their age, gender, birth country and region, and current country and region. We use the term *response* to refer to all data collected from one submission of the survey.

To ensure genuine responses, we came up with a list of quality control guidelines. This list was developed iteratively, after observing multiple survey submissions. As such, it is possible that not all tweets in our dataset come from submissions that met these guidelines. This is a limitation of the work described herein. Nevertheless, the final list is:

- The provided links should point to tweets posted no sooner than 48 hours before the submission, to prevent users from posting and providing tweets on the spot;
- All tweets in a response should come from the same account;
- Tweets cannot be from verified accounts or accounts with more than 30K followers, to avoid getting tweets from popular accounts and claiming to be personal tweets [1];
- Tweets should contain at least 5 words, excluding any hashtags and URLs; to decide on the number of words we used the `TweetTokenizer` in the NLTK toolkit (Loper and Bird, 2002);
- Links to tweets should not have been submitted in a previous response;
- Responses submitted in less than three minutes are discarded.

We published our survey on multiple crowdsourcing platforms, including Figure-Eight, known today as Appen; Amazon Mechanical Turk; and Prolific Academic[2]. We could not get high quality responses from Figure Eight, where by *high quality* we mean being in accordance with the criteria defined by the control steps above. On Mechanical Turk, we retrieved some high quality responses, but, unfortunately, they stopped our job, considering that getting links to personal tweets of participants violates their policy. We collected the majority of responses on Prolific Academic.

[1] The initial number was set to 5K, but some workers asked us to raise it since they had more followers.

[2] Appen: `www.appen.com`, Mechanical Turk: `www.mturk.com`, Prolific Academic: `prolific.ac`

4.3.2 Categorising Sarcastic Tweets

We then inspected each collected sarcastic tweet, along with the explanation provided by the author and the non-sarcastic rephrase, in order to further assign the tweet to one of the following categories:

1. sarcasm*: tweets that contradict the state of affairs and are critical towards an addressee;
2. irony*: tweets that contradict the state of affairs but are not obviously critical towards an addressee;
3. satire*: tweets that appear to support an addressee, but contain underlying disagreement and mocking;
4. understatement*: tweets that undermine the importance of the state of affairs they refer to;
5. overstatement*: tweets that describe the state of affairs in obviously exaggerated terms;
6. rhetorical question*: tweets that include a question whose invited inference (implicature) is obviously contradicting the state of affairs;
7. invalid: tweets for which the explanation provided by their authors is unclear / unjustified. These were excluded from the dataset.

This categorisation (excluding the *invalid* category) is the one defined by Leggitt and Gibbs (2000a). The symbol in the superscript attached to a word or phrase above is a visual indicator that the word or phrase denotes a category. Note the difference between "sarcasm" and "sarcasm*". The former denotes the higher-level phenomenon that encompasses all categories, while the latter refers to a specific category. That is, sarcasm* is a category of sarcasm.

4.3.3 Collecting Third-Party Labels

Next, we sampled a fraction of the tweets collected above, which we later consider to be the test set. For this fraction, we collected third-party sarcasm labels from human annotators. We refer to these labels as *perceived sarcasm* labels. We later compare these labels with the intended sarcasm labels, to estimate annotator performance at detecting sarcasm in tweets as intended by the authors of those tweets.

When collecting perceived sarcasm labels, we aimed to reduce noise caused by variations in how sarcasm is defined across annotators. As we discussed in Sec-

tion 4.2.2 annotator country could influence this definition. As such, we made sure all annotators lived in the same country. Specifically, we used Prolific Academic as the platform for publishing a third-party labelling survey, as it allows granular control over the target worker population. We selected workers from the United Kingdom, which yielded a large target population.

We collected three annotations for each tweet and used majority voting to choose the final label. This is the same procedure employed when the Riloff dataset was constructed (Riloff et al., 2013).

4.4 Exploratory Data Analysis

4.4.1 iSarcasm Dataset

We received 1,236 responses to our survey. Each response contained four tweets labelled for sarcasm by their author, one sarcastic and three non-sarcastic. As such, we received 1,236 sarcastic and 3,708 non-sarcastic tweets. We filtered tweets using quality control steps such as the ones described in Section 4.3.1. We also disregarded all tweets that fall under the *invalid* category mentioned in Section 4.3.2. The resulting dataset is what we call iSarcasm, containing 777 sarcastic and 3,707 non-sarcastic tweets. For each sarcastic tweet, we have its author's explanation as to why it is sarcastic, as well as how they would rephrase the tweet to be non-sarcastic. The average length of a tweet is around 20 words. Figure 4.1 shows the tweet length distribution across iSarcasm. The average length of explanations 21 words, and of rephrases 14 words. Over 46% of the tweets were posted in 2019, over 83% starting with 2017, and the earliest in 2008. Among the contributors who filled in our survey and provided the tweets, 56% are from the UK and 41% from the US, while 3% are from other countries such as Canada and Australia. 51% are females, and over 72% are less than 35 years old. Figure 4.2 shows the age and gender distributions across contributors. Table 4.1 shows the distribution of the sarcastic tweets into the categories introduced in Section 4.3.2. Sarcasm and irony are the largest two categories (73%), while understatement is the smallest one. Table 4.2 shows examples of the sarcastic tweets, along with the explanations and rephrases provided by the authors.

Recall that one of the limitations of distant supervision mentioned in Section 4.2 is the fact that it considers as non-sarcastic those texts that do not meet the predefined criteria. We investigated the presence of hashtags such as #sarcasm and #sarcastic in

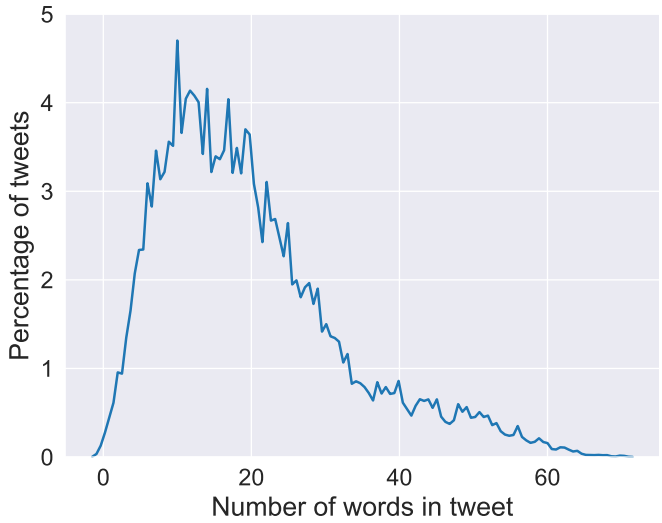

Figure 4.1: Tweet length distribution across iSarcasm, as discussed in Section 4.4.1.

overall		sarcasm category					
sarcastic	non-sarc.	sarcasm*	irony*	satire*	underst.*	overst.*	rhet. q.*
777	3,707	324	245	82	12	64	50

Table 4.1: Distribution of sarcastic tweets into the categories that were introduced in Section 4.3.2, as discussed in Section 4.4.1. The "*" symbol in the superscript attached to a word or phrase above is a visual indicator that the word or phrase denotes a category. Note the difference between "sarcasm" and "sarcasm*". The former denotes the higher-level phenomenon that encompasses all categories, while the latter refers to a specific category. That is, sarcasm* is a category of sarcasm.

iSarcasm, tags often used to mark sarcasm in previous distant supervision datasets. None of the tweets in iSarcasm contains such tags, confirming the potential of distant supervision to produce false negative labels.

We split iSarcasm into a training set and a test set, containing 80% and 20% of the examples chosen at random, respectively.

4.4.2 Perceived Sarcasm Labels

As we mentioned earlier, we collected three third-party labels for each tweet in the test set of iSarcasm. Using Cohen's kappa (κ; Cohen (1960)) as a measure, the pairwise inter-annotator agreement (IAA) scores were $\kappa_{12} = 0.37$, $\kappa_{13} = 0.39$ and $\kappa_{23} = 0.36$,

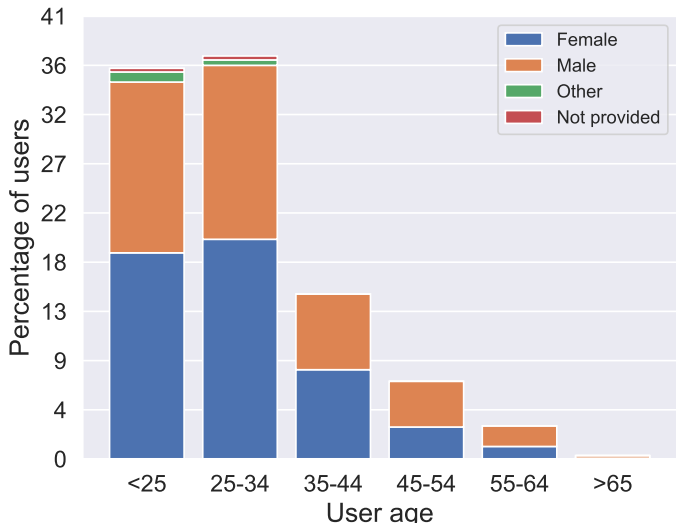

Figure 4.2: Age and gender distributions across the Twitter users who provided tweets in iSarcasm, as discussed in Section 4.4.1.

which highlights the high subjectivity of the task. We used majority voting to select a label each tweet, which we refer to as the *perceived sarcasm* label. Table 4.3 shows the agreement between the intended and perceived labels. As shown, 30% of the sarcastic tweets were unrecognised by the annotators, while 45% of the tweets perceived as sarcastic were not intended to be sarcastic by their authors. This further confirms that third-party annotation for sarcasm could produce noisy labels.

For more examples of tweets where intended and perceived labels are different, and further analyses in this direction, please consult our other related paper, Abu Farha et al. (2022b).

4.5 Evaluating Previous Sarcasm Detection Models

In the following, we examine the performance of previous sarcasm detection models on iSarcasm, specifically those models that have achieved state-of-the-art results on previous datasets. This addresses RQ 2.3.

4.5.1 Baseline Datasets

We consider four previously published datasets. Two of them, **Riloff** (Riloff et al., 2013) and **SemEval-2018** (Van Hee et al., 2018), were labeled via a hybrid approach

of distant supervision for initial collection, followed by manual labelling to produce the actual labels. We mentioned them in our discussion in Section 2.2.2. The other two datasets, **Ptacek** (Ptáček et al., 2014) and **SARC** (Khodak et al., 2018), were labeled using distant supervision.

The Riloff dataset consists of 3,200 tweet IDs. Out of these, we managed to collect only 1,832 using the Twitter API. Others have either been removed from Twitter, or the corresponding accounts have been disabled. Note that in Section 3.2 we reported being able to collect only 701 tweets, not 1,832. The discrepancy is due to the extra constraint enforced there, mainly that, besides the tweets themselves, historical tweets of the corresponding user should also be retrievable via the Twitter API. The next dataset, SemEval-2018, is a balanced dataset consisting of 4,792 tweet texts. The Ptacket dataset consists of 50,000 tweet IDs, out of which we were able to collect 27,177. Finally, The SARC dataset consists of Reddit comments. In a setting similar to Hazarika et al. (2018), who publish state-of-the-art results on this dataset, we consider two variants of SARC. SARC-balanced contains 154,702 comments with the same number of sarcastic and non-sarcastic comments, while SARC-imbalanced contains 103,135 comments, with a ratio of about 20:80 between the number of sarcastic and non-sarcastic comments.

4.5.2 Sarcasm Detection Models

Riloff and Ptacek datasets We replicate the models implemented in Tay et al. (2018), who report state-of-the-art results on Riloff and Ptacek. These models are:

- **LSTM** first encodes the tweet with a recurrent neural network with long-term short memory units (LSTM; Hochreiter and Schmidhuber (1997)), then adds a binary softmax layer to output a probability distribution over labels (sarcastic or non-sarcastic) and assigns the most probable label. It has one hidden layer of dimension 100.
- **Att-LSTM** adds an attention mechanism on top of the LSTM, in the setting specified by Yang et al. (2016). In particular, it uses the attention mechanism introduced by Bahdanau et al. (2014) of dimension 100.
- **CNN** encodes the tweet with a convolutional neural network (CNN) with 100 filters of size 3 and provides the result to feed-forward network with a final binary softmax layer, choosing the most probable label.
- **SIARN** (Single-Dimension Intra-Attention Network; Tay et al. (2018)) is the

model that yields the best published performance on the Riloff dataset. It relies on the assumption that sarcasm is caused by linguistic incongruity between words. It uses an intra-attention mechanism (Shen et al., 2018) between each pair or words to detect this incongruity.

- **MIARN** (Multi-Dimension Intra-Attention Network; Tay et al. (2018)) reports the best results on the Ptacek dataset. In addition to SIARN, MIARN allows multiple intra-attention scores for each pair of words to account for multiple possible meanings of a word when detecting incongruity. We use an implementation of MIARN similar to that described by its authors. We set the dimension of all hidden layers of **SIARN** and **MIARN** to 100.

SARC datasets Hazarika et al. (2018) report the best results on SARC-balanced and SARC-imbalanced, to our knowledge. However, they model both the content of the comments and the contextual information available about the authors of those comments. In this paper, we only focus on content modelling. For this, we use a convolutional network in a setting similar to what they describe. It uses three filter types of sizes 3, 4, and 5, with 100 filters for each size. We refer to this network as **3CNN**.

SemEval-2018 dataset The SemEval-2018 dataset contains two types of labels for each tweet: binary labels that specify whether the tweet is sarcastic or not; and labels with four possible values, specifying the category of sarcasm present. As such, there are two tasks for which this dataset is used, one corresponding to each label type. Wu et al. (2018) report the best performance at both tasks with their **Dense-LSTM** model. Given a tweet, the model uses a sequence of four LSTM layers to compute a hidden vector H. H is then concatenated with a tweet embedding S computed in advance by averaging embeddings of all words inside using the pre-trained embeddings provided by Bravo-Marquez et al. (2016). H and S are further concatenated with a sentiment feature vector of the tweet computed in advance using the *weka* toolkit (Mohammad and Bravo-Marquez, 2017), by applying the *TweetToLexiconFeatureVector* (Bravo-Marquez et al., 2014) and *TweetToSentiStrengthFeatureVector* (Thelwall et al., 2012) filters. The authors of Dense-LSTM train the network in a multitask setting on the SemEval-2018 dataset to predict three components: the binary sarcasm label, one of the four categories of sarcasm, and the corresponding hashtag, if any, that was initially used to mark the tweet as sarcastic, out of #sarcasm, #sarcastic, #irony and #not. They report an f1-score of 0.674. They further report an f1-score of 0.705 by averaging the performance of 10 Dense-LSTM models. We implement and Dense-LSTM to only

predict the binary sarcasm label, to make it applicable to iSarcasm and make the results on SemEval-2018 and iSarcasm comparable.

For each previous dataset, we implemented the models reported previously to achieve the best performance on that dataset, and made sure our implementations achieve similar performance to the published one. This is confirmed in Table 4.4, providing confidence in the correctness of our implementations.

Implementation Details In preprocessing, we use the spaCy library[3] for tweet tokenization. We then replace all tokens that only appear once in the corpus with ⟨unk⟩. We also replace all handles with a ⟨user⟩, ellipses with the ⟨ellipsis⟩ token, and numbers with the ⟨number⟩ token. We further remove all punctuation except ".", "!", "?", "(", and ")". We represent each tweet as a sequence of word vectors initialized using GloVe embeddings (Pennington et al., 2014) of dimension 100. We group our data into batches of 128 examples, train all models for 30 epochs using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 10^{-3}.

4.5.3 Results and Analysis

Table 4.5 reports precision, recall and f1-score results on the test set of iSarcasm using the detection models discussed, alongside third-party annotator performance. As shown, all the models perform considerably worse than humans, who achieve an f1-score of only 0.616. While MIARN is the best performing model, it achieves a low f1-score of 0.364, compared to the f-scores achieved on the Riloff and Ptacek datasets of 0.741 and 0.874, respectively. 3CNN achieves the lowest performance on iSarcasm with an F-Score of 0.286, compared to 0.675 and 0.788 on SARC-balanced and SARC-imbalanced, respectively. Similarly, Dense-LSTM achieves 0.318, compared to 0.666 on SemEval-2018.

Let us qualitatively consider those sarcastic tweets that Att-LSTM classifies correctly, but the human annotators do not. We noticed the attention weights were higher for some words that are commonly used to create hyperbole, such as *amazing*, *exciting* and *love*, or are associated with strong emotions, such as *proud*, *enjoy* and *anxiety*. On the other hand, some tweets that only humans classify correctly seem to require contextual information. One example is "Monday motivation: make it to friday!". Others tweets that only humans understand seem to allow more possible interpretations. One example is "I'm buzzing to get back to my double workouts tomorrow". Depending

[3] https://spacy.io/api/tokenizer

on the background of the person reading, it might be perceived as sarcastic (e.g., by a person who does not enjoy working out) or non-sarcastic (e.g., by a weightlifter).

All in all, models that achieved state-of-the-art performance on datasets labelled by third-party annotators, or labelled via distant supervision, fail dramatically on iSarcasm. This underlines the need for more effective models.

We believe the iSarcasm dataset, with its novel method of capturing sarcasm in tweets as intended by the authors of those tweets, has the ability to impact sarcasm detection research in the future. First, our experiments underline the need for more effective models, and iSarcasm is a platform where such models could be tested, platform free from the noise induced by the previous labelling methods. Second, iSarcasm opens the door towards exploring novel sarcasm-related tasks: sarcasm category prediction, using the category labels as ground truth (cf. Section 4.3.2); and sarcasm decoding and encoding, using the sarcastic tweets paired with their non-sarcastic rephrases, provided by the authors of those tweets (cf. Section 4.3.1).

4.6 Summary

The work conducted in this chapter addresses the granular research questions introduced in Section 4.1.

To RQ 2.1 we answer as follows. Distant supervision might produce false positive labels, because the predefined criteria can be met without necessarily implying the presence of sarcasm; it might only capture a specific flavour of sarcasm that is more subtle; and might produce false negatives due to the assumption that texts that do not meet predefined criteria are necessarily non-sarcastic. Manual labelling might produce both false positive and false negative labels. This is because, when labelling a text, the annotator might have different assumptions about what constitutes sarcasm, and how sarcasm is formulated, compared to the author of that text. Personal factors might also cloud the annotator's judgement, such as their mood and level of stress at the time of annotation.

In response to RQ 2.2, we introduced iSarcasm, a dataset of 4,484 tweets annotated for sarcasm by their authors. Sarcastic tweets have associated explanations as to why they are sarcastic; rephrases that convey the same message non-sarcastically; and labels that indicate the sarcasm category that they reflect. We believe iSarcasm will allow future work in sarcasm detection to progress in a setting free of the noise induced by previous labelling methods.

In response to RQ 2.3, we evaluated previous state-of-the-art sarcasm detection models on iSarcasm, observing a performance that is considerably lower than on previous datasets. We also collected third-party sarcasm labels from human annotators for the tweets in iSarcasm. Humans achieved a considerably higher performance, compared to models.

These results suggest further avenues of research, avenues that we explore in this thesis.

First, the low performance of state-of-the-art sarcasm detection models on iSarcasm, as well as the discrepancy between the performance of these and that of human annotators, suggests the need for more effective models. Indeed, more effective sarcasm detection in our tweets seems possible without further information external to the tweets, given that human annotators also lacked access to such information. In response, in Chapter 5 we report our work on crowdsourcing the task of building such models at the 16th International Workshop on Semantic Evaluation.

Second, despite being higher than model performance, in our experiments, human performance is less than 62% f-score. This could indicate that the task of detecting sarcasm in tweets is challenging even for humans. In Chapter 6 we reflect upon the factors that might contribute to making this task challenging for humans.

category	tweet text	explanation	rephrased
sarcasm*	Gotta love people who follow you and unfollow because you don't follow them within in an hour or 2. Sorry I don't stay on Twitter 24/7.	It is sarcastic because I dislike when people follow me only to unfollow me when I don't follow them back right away. I find it really annnoying and there is nothing I "love" about it.	I dislike people who follow me, only to unfollow me when I don't follow back right away. I'm not on Twitter that much to follow right away.
irony*	Staring at the contents of your fridge but never deciding what to eat is a cool way to diet	I wasn't actually talking about a real diet. I was making fun of how you never eat anything just staring at the contents of your fridge full of indecision.	I'm always staring at the contents of my fridge and then walking away with nothing cause I can never decide.
satire*	@user @user Totally didn't happen, it's a big conspiracy, video can be faked....after all, they've been faking the moon landings for years	It's an obvious subversion of known facts about mankind's space exploration to date that are nonetheless disputed by conspiracy theorists.	It's not a conspiracy, the video is real... after all, we've known for years that the moon landings happened.
underst.*	@user @user @user Still made 5 grand will do him for a while	The person I was tweeting to cashed out 5k in a sports accumulator - however he would've won 295k. "Still made 5k will do him for a while" is used to underplay the devastation of losing out.	He made 5 grand, but that will only last him a month.
overst.*	the worst part about quitting cigarettes is running into people you went to high school with at a vape shop	There are many things that are actually harder about quitting cigarettes than running into old classmates.	Running into old classmates at a vape shop is one of the easier things you have to deal with when you quit cigarettes.
rhetorical question*	@user do all your driver's take a course on how to #tailgate!	Drivers don't have to take a course on how to tailgate its just bad driving on their part.	Could you ask your drivers not to tailgate other people on the roads please?

Table 4.2: Examples of sarcastic tweets from iSarcasm, along with the explanations that the authors gave as to what made their tweets sarcastic (explanation), and the rephrase that they gave that would convey the same message non-sarcastically (rephrased), as discussed in Section 4.4.1. User handles were replaced with "@user".

	perceived sarcastic	perceived non-sarcastic
intended sarcastic	61	26
intended non-sarcastic	50	322

Table 4.3: The agreement between intended sarcasm labels, provided by the authors, and perceived sarcasm labels, provided by third-party annotators on the test set of iSarcasm, as discussed in Section 4.4.2.

Dataset	**Model**	**published**	**our impl.**
Riloff	LSTM	0.673	0.669
	Att-LSTM	0.687	0.679
	CNN	0.686	0.681
	SIARN	0.732	0.741
	MIARN	0.701	0.712
Ptacek	LSTM	0.837	0.837
	Att-LSTM	0.837	0.841
	CNN	0.804	0.810
	SIARN	0.846	0.864
	MIARN	0.860	0.874
SARC-balanced	3CNN	0.660	0.675
SARC-unbalanced	3CNN	0.780	0.788
SemEval-2018	Dense-LSTM	0.674	0.666

Table 4.4: F1-score yielded by our implementations of state-of-the-art models on previous datasets, compared to published results on those datasets, as discussed in Section 4.5.2.

Model	Precision	Recall	F1-score
Manual Labelling	0.550	0.701	**0.616**
LSTM	0.217	0.747	0.336
Att-LSTM	0.260	0.436	0.325
CNN	0.261	0.563	0.356
SIARN	0.219	0.782	0.342
MIARN	0.236	0.793	**0.364**
3CNN	0.250	0.333	0.286
Dense-LSTM	0.375	0.276	0.318

Table 4.5: Experimental results on iSarcasm, discussed in Section 4.5.3. *Manual Labelling* shows the results using the perceived sarcasm labels provided by third-party human annotators.

Chapter 5

iSarcasmEval: Intended Sarcasm Detection

In this chapter we address RQ 3 of this thesis. The work presented herein leads to Contribution 2 of this thesis. This chapter is based on the paper "iSarcasmEval, Intended Sarcasm Detection in English and Arabic" that we published in the Proceedings of the 16th International Workshop on Semantic Evaluation (Abu Farha et al., 2022a).

5.1 Introduction

In Chapter 4 we argued that previous methods of labelling texts for sarcasm are suboptimal. In response, we suggested a new labelling method and used it to create iSarcasm, a dataset of tweets labelled for sarcasm. We evaluated state-of-the-art sarcasm detection models on iSarcasm, and also collected third-party labels from human annotators. As seen by comparing Table 4.4 and Table 4.5, models achieved a considerably lower sarcasm detection performance on iSarcasm, compared to the performance they report on previous datasets. As further seen in Table 4.5, models also achieve a considerably lower performance compared to humans. Note that in the experiments shown therein, both models and humans were provided with the same information when asked to discern the sarcastic nature of a tweet. That is, the text of the tweet.

The low performance of state-of-the-art sarcasm detection models on iSarcasm, compared to both their performance on previous datasets, and to human performance, suggests the need for more effective models. In response, we formulate the third research question of this thesis.

RQ3 How can we build more effective sarcasm detection models?

To address this question, we crowdsourced the task of building such models at the 16th International Workshop on Semantic Evaluation. Every year, the workshop admits papers that summarise public competitions organised by the authors of those papers in the previous year. To organise a competition, one must first propose a task. In our case, this is sarcasm detection, with particular nuances, as discussed in Section 5.3. Once a board of reviewers admits the task, what follows is a public announcement encouraging the community to compete in solving the task. At the same time, the organisers usually make training and testing datasets available to the community. In our case, these include examples of tweets labelled for sarcasm.

In this chapter we overview the task that we describe in Abu Farha et al. (2022a), organised in collaboration with the other co-authors of that paper. Therein, we focused on two languages, English and Arabic, managing separate competitions for each lan-

guage. We refer to them as the English competition and the Arabic competition. My contributions towards the unit of work described in this chapter were: building the English training dataset; and helping write the paper above. In this thesis, we overview the data collection methodology for English, and report the sarcasm detection performance of the models submitted for English.

The rest of this chapter is organised as follows. Section 5.2 describes how we built a dataset of tweets labelled for sarcasm to be used for training and testing models. Section 5.3 introduces the task the we proposed, including three subtasks, indicates what constituted training and testing data for each subtask, and what metrics we used to evaluate model performance for each subtask. Section 5.4 reviews submissions, focusing on the top performing models for each subtask. Section 5.5 summarises the chapter, and reflects upon potential implications for future work.

5.2 Dataset Collection

We first collected a dataset of tweets labelled for sarcasm by their authors. For this, we used a survey published on the Prolific Academic crowdsourcing platform, asking questions similar to those asked by the survey described in Section 4.3.1. As a result, each tweet in the dataset has an associated label provided by its author, specifying the sarcastic nature of the tweet. Further, each sarcastic tweet also has an associated rephrase, provided by its author, that conveys the same message as the tweet, but does so without using sarcasm. Quality control was performed by manual inspection of the tweets and the rephrases. More systematic quality control could be performed in future work.

Next, we asked annotators to label each sarcastic tweet into the categories of ironic speech defined by Leggitt and Gibbs (2000b), the same categorisation that used in Section 4.3.2: sarcasm, irony, satire, understatement, overstatement, and rhetorical question. However, in contrast to Section 4.3.2, the labelling was not mutually exclusive. Intuitively, a tweet could belong to more than one category, e.g. it could be both sarcastic, and an understatement. Indeed, mutually exclusive categorisation is a limitation of the work described in Chapter 4 that we address here. We published the resulting dataset as the training dataset to those who participated in the English competition.

To construct a test set for English, we employed a slightly different approach to that described in Section 4.3.1. This is because the Prolific Academic crowdsourcing platform no longer permits us to ask workers to provide tweets that they themselves

have posted in the past. Such tweets can be used to personally identify workers, which the platform aims to prevent. Given the situation, to collect a test set, instead of asking workers to provide sarcastic tweets that they had posted in the past, we asked them to write a sarcastic text on the spot, along with a rephrase that would convey the same message non-sarcastically. To collect non-sarcastic texts, we proceeded as follows. We used the Twitter streaming API to collect an initial set of tweets. Next, we published these tweets in a survey on the Appen crowdsourcing platform[1], asking annotators to indicate the sarcastic nature of each tweet. We collected three such labels for each tweet, and used majority voting to decide on the final label. Out of these, we added to our test set those tweets labelled as non-sarcastic. In the process, we ensured that the distribution determined by the lengths of the non-sarcastic tweets matched that determined by the lengths of the sarcastic tweets in our test set. This was to ensure that length could not be used as a trivial discriminating signal between sarcastic and non-sarcastic tweets. Of course, using a different strategy to collect the test set is a limitation of this work.

The union of the training and testing datasets above is what we refer to as the iSarcasmEval dataset. It is published as a list of texts. Each text is accompanied by a binary sarcasm label, indicating whether it is sarcastic or non-sarcastic. Further, each sarcastic text is also accompanied by a rephrase that conveys the same message non-sarcastically, and a six further binary labels. Each of these binary labels corresponds to one of the categories of ironic speech mentioned above, indicating whether the text belongs to that category or not.

5.3 Task Description and Experimental Setting

The shared task that we published consists of the following three sub-tasks:

- **Subtask A - Sarcasm Detection**: Given a text, determine whether it is sarcastic or non-sarcastic;
- **Subtask B - Sarcasm Category Classification**: Given a text, determine which ironic speech categories it belongs to, if any;
- **Subtask C - Pairwise Sarcasm Identification**: Given a sarcastic text and its non-sarcastic rephrase, i.e. two texts that convey the same meaning, determine which is the sarcastic one.

The training set of iSarcasmEval contains 867 sarcastic and 2,601 non-sarcastic texts,

[1] https://appen.com

split	total	sarcastic	non-sarcastic
train	4,335	867	3,468
test (subtask A)	1,400	200	1,200
test (subtask C)	400	200	200

Table 5.1: Training and test set sizes for subtasks A and C, as discussed in Section 5.3.

split	sarcasm	irony	satire	underst.	overst.	rhet. quest.
train	713	155	25	10	40	101
test (subtask B)	180	20	49	1	10	11

Table 5.2: Training and test set sizes, along with the partition of texts into ironic speech categories, for subtask B, as discussed in Section 5.3.

respectively. Recall that each sarcastic text has an associated non-sarcastic rephrase. These 867 rephrases can be used as additional non-sarcastic examples. The test set contains 200 sarcastic and 1,200 non-sarcastic texts, respectively. The 1,200 non-sarcastic texts include the rephrases of the 200 sarcastic ones.

For subtasks A and C, the full test set is used. For subtask C, the test set contains the 200 sarcastic examples, along with their 200 non-sarcastic rephrases, without including the other non-sarcastic texts from the original test set described in Section 5.2. The sizes of the training and testing sets for subtask A and C are shown in Table 5.1. The sizes for subtask B, including the partition of sarcastic texts into ironic speech categories, are shown in Table 5.2.

Evaluation Metrics The evaluation metric for subtask A is the f-score of the sarcastic class, referred to as f_{sarc}. It is computed as $f_{\text{sarc}} = 2 \cdot (p_{\text{sarc}} \cdot r_{\text{sarc}})/(p_{\text{sarc}} + r_{\text{sarc}})$, where p_{sarc} and r_{sarc} are the precision and recall of the sarcastic class, respectively. For subtask B, the evaluation metric is the macro f-score over all categories of ironic speech, referred to as f. It is computed as $f = (1/N)\sum_{c=1}^{N} f_{\text{c}}$, where f_{c} is the f-score of category c of ironic speech, and N is the number of categories, that is, $N = 6$. For subtask C, the evaluation metric is accuracy. Accuracy is appropriate since we have an equal number of sarcastic and non-sarcastic examples in the test set of task C, as shown in Table 5.1. It is the ratio of the number of times the sarcastic text was correctly discriminated from its rephrase, to the size of the text set.

5.4 Submissions

There were 60 teams that participated in the shared task. Most popular was subtask A, with 43 submissions. Subtask B received 22 submissions and subtask C received 16 submissions. The following subsections provide an overview of the sarcasm detection approaches that performed best for each subtask.

5.4.1 Subtask A

Table 5.3 lists submissions for subtask A.

Baselines The table also lists two baseline results that we provide. The first one is referred to as *baseline-bert*. Given a text, it first uses a stack of transformer encoders (Vaswani et al., 2017), resulting in the BERT architecture (Devlin et al., 2019), to compute a vector representation of that text. This representation is also known as an *embedding*. We used the same tokenisation approach as the one used when training BERT. As such, the vector corresponding to the [CLS] token is considered to be the embedding of the entire input text. This embedding is provided to a classification head consisting of linear transformations and a softmax function. We interpret the output of the classification head as the probability that the input text is sarcastic. To implement baseline-bert used the transformers library Wolf et al. (2020), and initialised the encoders with the `bert-base-uncased` checkpoint published on the Huggingface model hub [2]. We fine-tuned the model for a maximum of 100 epochs, using early stopping regularisation with a patience of 3. We used a learning rate of $5e-5$, and clipped the norm of the gradients to 1. This resulted in a baseline f_{sarc} of 0.348. The second baseline is referred to as *baseline-svm*. Given an input text, it uses a support vector machine (SVM) with a polynomial kernel of degree 3, to classify the tf-idf representation of that texts. This resulted in a baseline f_{sarc} of 0.275. When training both baseline models described above, we considered the rephrases as additional non-sarcastic examples. In a preprocessing step, we remove all hashtags and urls, and replaced user handles with the token `@user`.

Consulting Table 5.3, we notice the team ranking first, `stce` (Yuan et al., 2022), achieved an f_{sarc} of 0.605. They use an ensemble learning approach with a combination of hard and soft voting between three models, all based on the transformer architecture: RoBERTa (Liu et al., 2019), initialised with the `roberta-large` check-

[2] https://huggingface.co

point from the Huggingface model hub; DeBERTa (He et al., 2021), initialised with the `deberta-v3-large` checkpoint; and XLM-RoBERTa (Conneau et al., 2020), initialised with the `xlm-roberta-large` checkpoint. They experiment with several strategies to achieve their results. First, in addition to the task dataset, they also consider public datasets, including: iSarcasm, described in Chapter 4 (Oprea and Magdy, 2020b); the dataset published by Van Hee et al. (2018), and a sample of texts from the multimodal sarcasm dataset[3]. Second, they extract statistical and text features that they concatenate to the text itself before providing it to the models above, such as part-of-speech information.

The team ranking second, `X-PuDu` (Han et al., 2022), achieved an f_{sarc} of 0.569. They ensemble two transformer-based models: ERNIE-M (Ouyang et al., 2021), and DeBERTa, mentioned above. After providing the input text to the models, they consider the vector representation corresponding to the [CLS] token as embedding of the text, which they provide to a classification head. The final ensemble considers not just the individual architectures above, but also the same architecture under different hyperparameter configurations.

The team ranking third, `TUG-CIC` (Aroyehun et al., 2022), achieved an f_{sarc} of 0.530. They use the BERT model mentioned above, but initialised with `BERTweet` checkpoints from the Huggingface hub, which they fine-tune on the SPIRS sarcasm dataset (Shmueli et al., 2020), before fine-tuning it on the task dataset.

We invite the interested reader to consult the respective papers for more details on the three approaches summarised above. These are Yuan et al. (2022), Han et al. (2022), and Aroyehun et al. (2022).

5.4.2 Subtask B

Table 5.4 lists submissions for subtask B, ranked by the macro f-score that they achieve.

Baselines We provide two baseline results for subtask B. The first one, referred to as *baseline-majority*, always predicts that the input text belongs to the ironic speech category of sarcasm, and not to any other category. This category was chosen as it is dominant in the training set, as seen in Table 5.2. The second beseline, referred to as *baseline-bert*, is similar to the baseline with the same name from Section 5.4.2. The difference is that the classification head has a 6-dimensional output, each corresponding to one of the six categories of ironic speech that we consider. We apply the sigmoid

[3] `https://github.com/headacheboy/data-of-multimodal-sarcasm-detection`

function to each dimension, interpreting the output as the probability that the input text reflects the ironic speech category corresponding to that dimension. We fine-tune this model in a similar setting as we did for baseline-bert from subtask A. This results in a baseline macro F-score of 0.0431.

Consulting Table 5.4, we notice the team ranking first, `PALI-NLP` (Du et al., 2022), achieved a macro f-score of 0.1630. They use an ensemble learning approach. The models they consider are BERT, initialised with the `BERT-base` checkpoint from the Huggingface hub; RoBERTa, initialised with the `RoBERTa-base` checkpoint from the hub; and BERTweet, initialised with the `BERTweet-base` checkpoint.

The team ranking second, `CS-UM6P` (El Mahdaouy et al., 2022), achieved a macro f-score of 0.0875. They use a model similar to GAN-BERT (Croce et al., 2020). It uses a generator that, conditioned on an ironic speech category, produces fake embeddings from a random noise that would resemble representations of examples from that ironic speech category. A discriminator is trained to recognise real examples from fake ones, while the generator is trained to cause the discriminator to classify fake examples as real. The discriminator is also trained to classify the real examples as either sarcastic, or non-sarcastic.

The team ranking third, `MaChAmp` (van der Goot, 2022), achieved a macro f-score of 0.0851. They first we pre-train a RemBERT (Chung et al., 2020) multi-task model across all the tasks. Then, they re-train a model for each task individually. They use the hyperparameters of `MaChAmp v0.3` (van der Goot et al., 2021), which were finetuned on the xTREME benchmark (Hu et al., 2020).

5.4.3 Subtask C

Table 5.5 lists submissions for subtask C, ranked by the accuracy that they achieve.

Baselines We provide baseline results computed using the models from subtask A, described in Section 5.4.3. However, we provided different inputs to the models. Specifically, given a sarcastic text and its rephrase, we produced two training examples. The first was the concatenation of the sarcastic text and the rephrase, in this order, separated by a [SEP] token. This example had label 0, indicating the position of the sarcastic text. The second example was the concatenation of the rephrase and the sarcastic text, in this order, and had label 1. The first baseline, referred to as *baseline-bert*, achieved an accuracy of 0.765, while the second one, referred to as baseline-svm, achieved 0.495.

Consulting Table 5.5, we notice the team ranking first, `X-PuDu` (Han et al., 2022), achieved an accuracy of 0.870. The same team ranked second for subtask A, and the approach here is rather similar, except for representing the input as we do above. The team ranking second, `Naive`, achieved an accuracy of 0.855. They used a RoBERTa model, initialised with the `RoBERTa-large` checkpoint from the Huggingface hub, with a classification head appended. The team ranking third, `YNU-HPCC` (Zheng et al., 2022), achieved an accuracy of 0.805. They also used a RoBERTa model. We suspect the difference in performance between the second and third teams to be, at least in part, the result of data preprocessing and hyperparameter optimisation.

5.5 Summary

This chapter provided an overview of the shared task we published at the 16th International Workshop on Semantic Evaluation, with the purpose of building more effective sarcasm detection models.

We created the iSarcasmEval dataset of texts labelled for sarcasm by their authors. Each sarcastic text is also accompanied by a rephrase, and six binary labels, each corresponding to a category of ironic speech. The shared task contained three subtasks: sarcasm detection, ironic speech category classification, and pairwise sarcasm identification. There were 60 teams that participated. We provided a high-level overview of the approaches of the top performing teams, for each subtask.

For all subtasks, the submissions that performed best used models based on the transformer architecture. The top performing models achieved: 0.605 f-score for subtask A; 0.1630 macro f-score for task B; and 0.870 accuracy for task C. This suggests that the task of detecting sarcasm in text remains challenging, and detecting the ironic speech category even more so. We hope our shared task will draw the attention of the community towards these tasks.

rank	team name	affilliation	f_{sarc}
1	stce	PALI Inc., China	0.605
2	X-PuDu	Baidu & Shanghai Pudong Development Bank, China	0.569
3	TUG-CIC	TU Graz, Austria	0.530
4	Plumeria	Indian Institute of Technology Kanpur, India	0.477
5	John Thomson	University of Alberta, Canada	0.456
6	Naive	Dalian University of Technology, China	0.452
7	MarSan_AI	Part AI Research Center, Iran	0.434
8	LISACTeam	Sidi Mohamed Ben Abdellah University, Morocco	0.429
9	LT3	Ghent University, Belgium	0.424
10	niksss	-	0.402
11	Amobee	-	0.401
12	YNU-HPCC	Yunnan University, China	0.392
13	Dartmouth	Dartmouth College, USA	0.386
14	underfined	Ping An Life Insurance Company of China, China	0.383
15	CS-UM6P	Mohammed VI Polytechnic University, Morocco	0.371
16	UTNLP	University of Tehran, Iran	0.369
17	Jumana-Safa	-	0.356
18	cnxup	University of Chinese Academy of Sciences, China	0.351
-	baseline-bert	-	0.348
19	IISERB Brains	Indian Institute of Science Education and Research, Bhopal, India	0.345
20	rematchka	Cairo University, Egypt	0.341
21	R2D2	Vellore Institute of Technology, India	0.328
22	AMI_UofA	University of Alberta, Canada	0.312
23	Amrita-CEN	Amrita Vishwa Vidyapeetham, India	0.308
24	DUCS	University of Delhi, India	0.307
25	Happy New Year	-	0.276
-	baseline-svm	-	0.275
26	Sarcastic weeps	FAST NUCES LHR, Pakistan	0.270
27	TechSSN	Sri Sivasubramaniya Nadar College of Engineering, India	0.264
28	NULL	Auburn University, USA	0.260
29	Cyborgs	-	0.248
30	I2C	Universidad de Huelva, Spain	0.245
31	MaChAmp	IT University of Copenhagen, Denmark	0.241
32	ISD	Stanford University, USA	0.240
33	SPDB	-	0.215
34	xuyt3	-	0.215
35	MACHON	Jerusalem College of Technology, Israel	0.215
36	FII_UAIC	University of Iasi, Romania	0.207
37	connotation_clashers	University of Tübingen, Germany	0.202
38	GetSmartMSEC	Meenakshi Sundararajan Engineering College, Chennai, India	0.201
39	UoR-NCL	University of Reading, UK	0.195
40	JCT	Jerusalem College of Technology, Israel	0.184
41	UMUTeam	Universidad de Murcia, Spain	0.180
42	MACHON	Jerusalem College of Technology, Israel	0.168
43	NARD@KGP	IIT Kharagpur, India	0.155

Table 5.3: Submissions for subtask A, in descending order, according to f_{sarc}, the f-score of the sarcastic class, as discussed in Section 5.4.1. Baselines results, baseline-bert and baseline-svm, are also listed. The affiliation of some teams is not specified.

rank	team name	affiliation	macro f-score
1	PALI-NLP	Ping An, China	0.1630
2	CS-UM6P	Mohammed VI Polytechnic University, Morocco	0.0875
3	MaChAmp	IT University of Copenhagen, Denmark	0.0851
4	Naive	Dalian University of Technology, China	0.0809
5	X-PuDu	Baidu & Shanghai Pudong Development Bank, China	0.0799
6	Plumeria	Indian Institute of Technology Kanpur, India	0.0778
7	R2D2	Vellore Institute of Technology, India	0.0760
8	IISERB Brains	Indian Institute of Science Education and Research, India	0.0751
9	MarSan_AI	Part AI Research Center, Iran	0.0743
10	I2C	Universidad de Huelva, Spain	0.0699
11	YNU-HPCC	Yunnan University, China	0.0646
12	John Thomson	University of Alberta, Canada	0.0601
13	AMI_UofA	University of Alberta, Canada	0.0601
14	Dartmouth	Dartmouth College, USA	0.0590
15	Amrita-CEN	Amrita Vishwa Vidyapeetham, India	0.0567
16	rematchka	Cairo University, Egypt	0.0560
17	TechSSN	Sri Sivasubramaniya Nadar College of Engineering, India	0.0465
18	NARD@KGP	IIT Kharagpur, India	0.0446
-	baseline-bert	-	0.0431
19	GetSmartMSEC	Meenakshi Sundararajan Engineering College, Chennai, India	0.0387
20	niksss	-	0.0380
-	baseline-majority	-	0.0380
21	Suhaib-Aburaidah	-	0.0346
22	Sarcastic weeps	FAST NUCES LHR, Pakistan	0.0313

Table 5.4: Submissions for subtask B, in descending order, according to the macro f-score, as discussed in Section 5.4.2. Baselines results, baseline-bert and baseline-majority, are also listed. The affiliation of some teams is not specified.

r	Team name	Affiliation	Accuracy
1	X-PuDu	Baidu, China	0.870
2	Naive	Dalian University of Technology, China	0.855
3	YNU-HPCC	Yunnan University, China	0.805
4	Plumeria	Indian Institute of Technology Kanpur, India	0.790
5	LISACTeam	Sidi Mohamed Ben Abdellah University, Morocco	0.775
6	UTNLP	University of Tehran, Iran	0.770
7	MarSan_AI	Part AI Research Center, Iran	0.765
-	baseline-bert	-	0.765
8	R2D2	Vellore Institute of Technology, India	0.750
9	NARD@KGP	IIT Kharagpur, India	0.735
10	rematchka	Cairo University, Egypt	0.720
11	CS-UM6P	Mohammed VI Polytechnic University, Morocco	0.695
12	Dartmouth	Dartmouth College, USA	0.660
13	IISERB Brains	Indian Institute of Science Education and Research, Bhopal, India	0.625
14	Sarcastic weeps	FAST NUCES LHR, Pakistan	0.495
-	baseline-svm	-	0.495
15	GetSmartMSEC	Meenakshi Sundararajan Engineering College, Chennai, India	0.340
16	MaChAmp	IT University of Copenhagen, Denmark	0.250

Table 5.5: Submissions for subtask C, in descending order, according to the accuracy, as discussed in Section 5.4.3. Baselines results, baseline-bert and baseline-svm, are also listed. The affiliation of some teams is not specified.

Chapter 6

The Influence of Socio-demographic Factors on Sarcastic Exchanges

In this thesis, we reflect upon previous computational investigations of sarcasm. In response to their limitations, we formulate our five research questions, listed in Section 1.2; these span two research directions, sarcasm detection and sarcasm understanding. The first three questions are related to the first direction, while the last two questions are related to the second direction. So far, we have explored the first direction. We now switch focus to the second one.

As such, in this chapter we address RQ 4 of this thesis. The work presented herein leads to the Contribution 3. This chapter is based on the paper "The Effect of Socio-demographic Variables on Sarcasm Communication Online" that we presented at the 23rd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW) (Oprea and Magdy, 2020a).

6.1 Introduction

In Chapter 4 we noticed that models achieved a considerably lower sarcasm detection performance on iSarcasm, compared to human annotators. However, as shown in Table 4.5, despite being higher than model performance, human performance was still less than 62%, quantified using the f-score. We first focused on model performance. Specifically, in an attempt to reduce the discrepancy between model and human performance, we crowdsourced the task of building more effective models, as described in Chapter 5.

Let us turn our attention to human performance. The low f-score could indicate that the task of detecting sarcasm in tweets is challenging even for humans. In this chapter we aim to determine factors that influence the ability of humans to perform this task. Our focus is on socio-demographic factors, such as age, gender, and country. More specifically, we are investigating whether interlocutors are more able to detect each other's sarcasm when their socio-demographic backgrounds are similar, compared to when their backgrounds are dissimilar.

Here, *interlocutors* are participants in a sarcastic exchange. At any point in the exchange, we differentiate between a *speaker*, and a *listener*. We assume each interlocutor is characterised by a socio-demographic background, which is a set of socio-demographic traits. Such a trait is an instance of a socio-demographic factor that determines a coarse partition over the space of potential interlocutors. That is, it does not identify any individual interlocutor, but a set of them. For instance, *gender* and *country* are factors; *female* and *United Kingdom* are traits that instantiate these fac-

tors; and the set {female, United States} is a background. Since we have defined a socio-demographic background as a set, we further employ set-theoretic terminology to compare and contrast backgrounds. As such, the similarity between two backgrounds is quantified as the cardinality of their intersection. The larger the cardinality, the more *similar* the backgrounds are; the smaller the cardinality, the more *dissimilar* they are. As special cases, we say two backgrounds are *identical* if they fully overlap, and are *disjoint* if their intersection is the empty set.

We proceed in our investigation by consulting previous studies of sarcasm in linguistics and sociolinguistics. According to such studies, we should indeed expect interlocutors with similar socio-demographic backgrounds to be more able to detect each other's sarcasm. There are at least two lines of reasoning that lead to this hypothesis.

First, as implied by linguistic studies (cf. Section 6.2.1), in a sarcastic exchange, the amount of common ground that the interlocutors share could be positively correlated with their ability to detect each other's sarcasm. Specific constituents of this common ground that are mentioned include shared social norms and shared expectations. In this case, in as much as socio-demographic factors determine common ground, we expect socio-demographic similarity between interlocutors to be positively correlated with their ability to detect each other's sarcasm.

Second, as mentioned in sociolinguistic studies (cf. Section 6.2.2), in a sarcastic exchange, certain socio-demographic traits that characterise the interlocutors could influence: when they choose to use sarcasm; how they formulate sarcastic utterances; and their predisposition to interpret utterances as sarcastic. This implies that interlocutors with similar traits could employ a similar mechanism to formulate and interpret sarcastic utterances. In turn, this could imply our hypothesis that socio-demographic similarity between interlocutors to be positively correlated with their ability to detect each other's sarcasm.

However, most of the linguistic and sociolinguistic studies consulted draw their conclusions by qualitatively reflecting upon sarcastic exchanges. There is a shortage of quantitative empirical evidence for our hypothesis, especially in the context of asynchronous exchanges of short textual utterances, such as tweets.

In response, we formulate the fourth research question of this thesis.

RQ 4 Are interlocutors with similar socio-demographic backgrounds more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds?

We divide RQ 4 into the following, more granular, research questions:

RQ 4.1 Are interlocutors with identical backgrounds more able to detect each other's sarcasm, compared to interlocutors with disjoint backgrounds?

RQ 4.2 If so, equality between what socio-demographic factors is most influential on the ability of interlocutors to detect each other's sarcasm?

RQ 4.3 Is socio-demographic similarity still influential when the listener has access to contextual information surrounding the speaker's utterance?

To address these questions, we emulate asynchronous exchanges via the process of third-party annotation. The subjects of annotation are tweets from the iSarcasm dataset introduced in Chapter 4.

We choose four socio-demographic factors for investigation: age, gender, country, and English language nativeness. There are two reasons for this choice. First, the information we have about the authors of the tweets in iSarcasm is limited to these four factors. Second, as discussed in Section 6.2.2, previous research supports the hypothesis that these factors have an influence in sarcastic exchanges.

We proceed as follows. First, we sample a subset from iSarcasm, considering the tweets, along with the *intended sarcasm* labels provided by the authors of those tweets. Next, we form several treatment groups of annotators. Each group contains representatives of one specific socio-demographic background, where a background is defined in terms of the four factors above. So, we have the same socio-demographic information about both tweet authors and annotators. Next, in two experimental settings, we ask each group of annotators to label the tweets in our sample as either sarcastic, or not sarcastic; we refer to any such resulting label as a *perceived sarcasm* label. In the first setting, the annotators are only shown the text of the tweets that they are asked to label. In the second setting, they are not shown the text, but are provided with the link to the tweet, and are asked to also consider surrounding tweets and profile information of the author of the tweet when deciding on the label. In each of the two experimental settings, and for each treatment group, we compare intended sarcasm labels with perceived labels. This allows us to quantify the ability of annotators to detect sarcasm in tweets, as intended by the authors of those tweets. To answer our research questions, we examine both the situation when annotator and author backgrounds are similar, and when their backgrounds are dissimilar.

Our results suggest an affirmative answer to RQ 4. In the first experimental setting, we find that similarity between tweet authors and annotators is significantly influential on the ability of the annotators to recognise sarcasm in tweets, as intended by the authors. This similarity is in terms of age, gender, and English language nativeness. In

the second setting, while the presence of contextual information alleviates the influence of English language nativeness and gender, age remains significantly influential.

The rest of this chapter is organised as follows. Section 6.2 reflects upon the implications of previous studies of sarcasm in linguistics and sociolinguistics, with the goal of determining their stance on RQ 4. Section 6.3 describes the methodology employed to quantitatively investigate RQ 4, with particular reference to each of the granular research questions of this chapter. Section 6.4 reports and analyses our quantitative investigation, and states the answers to the granular research questions. Section 6.5 provides these answers in a concise format, and discusses key takeaways and implications for future work. Section 6.6 summarises the chapter.

6.2 Socio-demographic Factors

In this section, we reflect upon previous studies of sarcasm in linguistics and sociolinguistics. Our goal in this section is to form, based on these studies, a stance on RQ 4. That is, on whether interlocutors with similar socio-demographic backgrounds more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds.

6.2.1 Sarcasm in Linguistics

We begin by looking at the linguistic theories of sarcasm introduced in Section 2.1. These theories do not make explicit statements about RQ 4. However, we may attempt to derive such statements from the assumptions of each theory.

Gricean Theory The Gricean view is introduced in Section 2.1.3. It does not make explicit statements about the socio-demographic dimensions of sarcastic exchanges. However, it does require the listener to point out the meaning implicated by the speaker. In this direction, research (Bouton, 1988, 1992) find that non-native speakers of English tend to differ in the meaning they attribute to conversational implicatures, compared to native speakers, in a setting where English language proficiency is controlled. If this is the case, we expect interlocutors who are both native speakers to be more able to detect each other's sarcasm, compared to those who are not native speakers. As such, English language nativeness is one of the socio-demographic factors that we consider.

Indirect Speech Act Theory The Indirect Speech Act theory is introduced in Section 2.1.2. For the instances of sarcasm that it does explain, let us reflect on RQ 4 in light of this view. In this direction, Searle (1975) notes that explaining indirect speech acts requires knowledge of the mutually shared background information of the interlocutors. This includes shared social norms and expectations. To create sarcasm, such information can be evoked and negated (Amante, 1981). Following our example in Section 2.1.2, the fact that cakes should not be served overcooked is a norm that Alice implicitly assumes. For her sarcasm to be understood, it is essential for Bob to also assume it. In this case, in as much as socio-demographic factors determine the norms that are assumed, we expect interlocutors with similar socio-demographic backgrounds to be more able to detect each other's sarcasm.

Echoic Theories Echoic theories are introduced in Section 2.1.3. They do not make explicit statements about the socio-demographic dimensions of sarcastic exchanges. However, as pointed out by Pexman (2005), some interlocutors, perhaps those more cynical, might attend more to failed expectations. Therefore, they might expect to see sarcasm more often in conversations with others. Of course, this might lead to sarcasm being undetected in exchanges unless all interlocutors engaged have similar ideas of what constitutes a failed expectation. As such, in as much as socio-demographic factors determine what expectations are assumed, listeners and speakers with similar socio-demographic backgrounds should be more able to detect each other's sarcasm, compared to those with dissimilar backgrounds.

Pretense Theories Pretense theories are introduced in Section 2.1.4. For instances of sarcasm that they do explain, socio-demographic factors might play a role in as much as they are consistent with pretense behaviour. First, as suggested by Pexman (2005), if the speaker exhibits traits that are consistent with insincerity and injudiciousness, then the listener might be more inclined to expect sarcasm from the speaker. Second, as further pointed out by Pexman (2005), if the listener shares these traits, they might be more able to detect the speaker's pretense. This seems to suggest that the most efficient setting for sarcastic exchanges is when the the speaker and listener have similar socio-demographic backgrounds. In the category of pretense theories there is also the more recent work of Cohn-Gordon and Bergen (2019). They suggest viewing sarcasm as a form of linguistic countersignaling: a communicative act where the interlocutors engage in a joint pretense about the state of the world, or the perspective that they hold, with the purpose of communicating about the common ground. The

ability of interlocutors to detect each other's sarcasm could thus be quantified as the amount of the shared knowledge that they have. This leads us to the same expectation of an increased detection ability between interlocutors of similar socio-demographic backgrounds, under the assumption that such backgrounds determine social partitions that share common ground.

Implicit Display Theory The Implicit Display theory is introduced in Section 2.1.5. Under this theory, interlocutor traits could play a role in sarcastic exchanges, as pointed out by Pexman (2005). Speaker traits could determine the expectations that the speaker assumes, as well as their predisposition to express negative attitudes in exchanges, and the manner in which they express such attitudes. As such, if the listener shares such traits with the speaker, that could make the listener more likely to judge the utterance of the speaker as being closer to prototypical sarcasm. In this case, in as much as such traits include socio-demographic traits, the setting that assures the maximum amount of accurate information transfer between the interlocutors is when they have similar socio-demographic backgrounds.

6.2.2 Sarcasm in Sociolinguistic

In Section 6.2.1 we introduced linguistic theories of sarcasm. We discussed how these theories might support an affirmative answer to RQ 4. That is, we expect interlocutors with similar socio-demographic backgrounds to be more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds.

As such, we are motivated to conduct our quantitative investigation of RQ 4 with this expectation in mind. However, it is still unclear *what* socio-demographic factors we should consider in our investigation—with the exception of English language nativeness, which we suggested when considering the Gricean theory above.

To discover more factors, in this section we turn to previous studies of sarcasm in sociolinguistics. We do not necessarily look for direct evidence towards RQ 4. Rather, we look at a broad range of studies that are concerned with the socio-demographic ecology of sarcastic exchanges.

To be more specific, consider the following thought experiment. Let's say we find a study which shows that interlocutors of similar ages understand each other's sarcasm better, compared to interlocutors of different ages. Such a study would provide direct evidence towards our research question, and suggest age as a socio-demographic factor for later quantitative investigation. However, continuing with the though experiment,

we also look at studies which, for instance, show that age determines the mechanism by which interlocutors formulate or comprehend sarcastic utterances—without making direct statements about the ability of interlocutors of different ages to understand each other's sarcasm. In such cases, we still consider age as a socio-demographic factor for later quantitative investigation. This is under the assumption that if age can influence formulation or comprehension, those of similar ages formulate and comprehend sarcasm similarly, thus understanding each other's sarcasm better.

In what follows, we mention the findings of some sociolinguistic studies, categorising them based on the socio-demographic factor that they mention. We then consider each such factor for later quantitative investigation, following the reasoning presented in the previous paragraph. More socio-demographic factors that the ones below are discussed in sociolinguistic literature. However, given our experimental setting outlined in Section 6.1, we only consider three factors: gender, age, and country. This is because we only have gender, age, and country information about the authors of the tweets in the iSarcasm dataset.

Gender Jorgensen (1996a) look at gender differences in emotional reactions to sarcasm. They notice males to be more likely than females to perceive humor in sarcasm, and females to be more likely to be offended or angered by sarcasm. Gibbs (2000) notices males to be more likely to use sarcasm in conversations with friends, compared to females. Taylor (2016) notice no correlation between gender and tendency to use sarcasm. Rather, they notice a preference for perceiving male behaviour as sarcastic.

Age Developmental literature suggests children begin to use personality traits to infer non-obvious meanings as young as age 4 (Heyman and Gelman, 1999). Harris et al. (2001) explore children's abilities to use the speaker traits as cues to sarcastic intent. They look at whether consistent personality trait information, such as being told that a sarcastic criticism was made by a mean speaker, would enable the detection of sarcastic intent. They notice that younger children rely more heavily on trait information, while older children have a stronger understanding of the phenomenon of sarcasm and a more complex way of integrating speaker traits into the discerning process. Phillips et al. (2015) compare the ability of adults of different ages to detect sarcasm in conversations that they observe. In their study, younger and middle-aged adults were significantly more effective at the task, compared to older adults. The latter were likely to take the literal meaning of sarcastic utterances, instead of the intended meaning concealed by sarcasm.

Country In a quantitative study, Joshi et al. (2016a) present a dataset of tweets, initially labelled for sarcasm by American annotators, to also be labelled by Indian annotators. They find higher disagreement between annotators of different nationalities, than between annotators of the same nationality. They consider cultural differences between India and the United States to be the cause of this disagreement. Note that their work is different from ours. They compare the labelling disagreement *between annotators* of different countries. By contrast, we compare the labelling disagreement *between annotators of tweets, and the authors of those tweets*. We do so both when annotator and author backgrounds are similar, and when their backgrounds are dissimilar. Nevertheless, compelled by the results of Joshi et al. (2016a), we include country as a factor of investigation. In our work, we choose to investigate two countries: the United Kingdom (UK) and the United States (US).

6.2.3 Sarcasm and Trolling

Before moving on to discussing our experimental setting, we make one final point. Buckels et al. (2014) suggest that sarcasm is a type of trolling. From here, one could further consult the work of Craker and March (2016), and that of Cheng et al. (2017), who show that two of the factors we have selected for investigation, mainly age and gender, are associated with trolling. Would this render our investigation of these factors redundant? Not quite. Based on linguistic and sociolinguistic studies of sarcasm, we find it problematic to consider sarcasm a type of trolling. To show why this is the case, we show that the intention to troll is not necessary for sarcasm to occur.

From a formal linguistic perspective, the argument is straightforward. Grice's theory only postulates the violation of a maxim and nothing about how that violation is achieved. Echoic theories have no claim over the manner in which dissociation from a previous proposition is achieved. While the Implicit Display Theory requires an expression of a negative attitude, it does not require that attitude to be trolling. In fact, it does not require that the expression should have an addressee at all. Indeed, it could well be directed at an object, or could be self-reflexive.

The fact that trolling is not necessary for sarcasm to occur is even more apparent if we look at sociolinguistic studies on the role of sarcasm in communication. Of particular relevance are the works of Jorgensen (1996b) and Pexman and Zvaigzne (2004a), who argue that reasons a speaker might choose to use sarcasm include: to demonstrate and enhance relationship closeness with the listener; as a linguistic code

between friends; or to show affection and appreciation.

As such, we have reasons to doubt the assumption that sarcasm is necessarily a type of trolling.

***

To sum up, in this section we introduced previous studies in linguistics and sociolinguistics. We discussed how linguistic theories might support the idea that interlocutors with similar socio-demographic backgrounds are more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds. Next, looking at both linguistic and sociolinguistic studies, we suggested age, gender, country, and English language nativeness as socio-demographic factors for investigation.

6.3 Data Collection and Analysis Methodology

In this section we describe the methodology employed to quantitatively investigate the granular research questions of this chapter.

6.3.1 Collecting Intended Sarcasm Labels

To address these questions, we emulate asynchronous exchanges via the process of third-party annotation. The subjects of annotation are tweets from the iSarcasm dataset introduced in Chapter 4. To collect the tweets in iSarcasm, along with the intended sarcasm labels, we used a survey published on the Prolific Academic crowdsourcing platform. Please consult Section 4.3 for more details.

In Section 6.2 we chose four socio-demographic factors for investigation: age, gender, country, and English language nativeness. In this context, a socio-demographic background is a set of four elements, each instantiating one of these factors. However, we have not yet specified what possible values we consider for each factor. To decide on this, we looked at the top socio-demographic backgrounds on Prolific Academic, in terms of the size of the partitions they determine over the space of workers. As such, we targeted workers of two backgrounds: females from the United Kingdom between 25- and 34-years-old; and males from the United States in the same age range. That is, the background of the first group is {age=between 25- and 34-years-old, gender=female, country=United Kingdom, English language nativeness=native}. For brevity, we denote this background as F_25-34_UK. Analogously, we denote the background of the latter group as M_25-34_US.

For the experiments in this chapter, we considered a total of 30 responses for each background, making a total of 240 tweets for both backgrounds, with a proportion of 1:4 of sarcastic to non-sarcastic tweets. This dataset is a subset of what constitutes the iSarcasm dataset. From this point onwards, in this chapter, whenever we use the term *iSarcasm*, we refer to this subset, rather than the entire dataset, unless otherwise specified.

6.3.2 Collecting Perceived Sarcasm Labels

We now describe how we collected perceived sarcasm labels for the tweets mentioned in the previous section. Our plan was to compare intended and perceived labels, as a way of tackling our research questions. We collected perceived labels from different treatment groups, and in different settings, depending on the granular research question addressed.

RQ 4.1

RQ 4.1 asks whether, in sarcastic exchanges, interlocutors with identical backgrounds more able to detect each other's sarcasm, compared to interlocutors with disjoint backgrounds?. In our experimental setting, an exchange consists of observing and annotating one of the tweets in iSarcasm. In this context, the interlocutors are the following: the speaker is the Twitter user who posted the tweet, and who provided the intended sarcasm label for that tweet. The listener is an annotator providing a perceived sarcasm label for that tweet.

To address RQ 4.1, we published a further online survey on Prolific Academic. The survey showed the texts of several tweets from iSarcasm and asked listeners, i.e. survey participants, to label each tweet as either sarcastic, or non-sarcastic. For each tweet, we collected such labels from two treatment groups, with 3 separate labels per tweet from each group, to alleviate labelling noise. The first group consisted of annotators who had the same socio-demographic background as the speaker of the tweet. The second group contained listeners with backgrounds disjoint to the background of the speaker. That is:

- If the tweet came from a female speaker from the United Kingdom who is between 25- and 34-years-old (F_25-34_UK), the first group contained listeners of the same background (F_25-34_UK), while the second group contained male listeners from the United States who are over 45-years-old (M_>45_US);

- Similarly, if the tweet came from a male speaker from the United States who is between 25- and 34-years-old (M_25-34_US), the first group contained listeners of the same background (M_25-34_US), while the second group contained female listeners from the United Kingdom who are over 45-years-old (F_>45_UK).

For brevity, we introduce the following condensed notation. Given a tweet, the specific background of the speaker could be irrelevant towards the point that the tweet makes, i.e. it can be any of the two speaker backgrounds that we consider (F_25-34_UK or M_25-34_US). In such a scenario, we use the notation *list=speak* to refer to the treatment group that contains listeners with the same background as that of the speaker, and *list≠speak* to refer to the treatment group that contains listeners with backgrounds disjoint from that of the speaker. For instance, if speaker background is F_25-34_UK, *list=speak* denotes the F_25-34_UK group of listeners, and *list≠speak* denotes the M_>45_US group. For reference, the treatment group notation is summarised in Table 6.1, and the condensed notation in Table 6.2.

We compared intended labels with *list=speak* and *list≠speak* perceived labels across our dataset, to see which of the two groups was best at capturing sarcasm as intended by the speakers. This allows us to make a statement about RQ 4.1, as discussed in Section 6.4.1.

RQ 4.2

RQ 4.2 asks for a more granular investigation into the influence of individual socio-demographic factors on the ability of interlocutors to detect each other's sarcasm.

In this purpose, we used the same survey as we did when collecting perceived sarcasm labels for addressing RQ 4.1. However, we collected perceived labels from four treatment groups, with three separate labels per tweet from each group. Each group corresponds to one of the four socio-demographic factors we chose for investigation: age, gender, country, and English language nativeness. Within each group, the listeners have the same background as the speaker, except for flipping value of the corresponding factor of study. That is:

- If the tweet came from a F_25-34_UK speaker, the first group, used for studying the influence of age, contained female listeners from the United Kingdom who were over 45-years-old (F_>45_UK). The second group, studying the influence of gender, contained M_25-34_UK listeners. The third one, studying

Group notation	Description
F_25-34_UK	Females between 25- and 34-years-old from the United Kingdom
F_25-34_US	Females between 25- and 34-years-old from the United States
F_>45_UK	Females over 45-years-old from the United Kingdom
F_25-34_!native	Females between 25- and 34-years-old, fluent, but non-native, speakers of English
M_25-45_US	Males between 25- and 34-years-old from the United States
M_>45_US	Males over 45-years-old from the United States
M_25-34_UK	Males between 25- and 34-years-old from the United Kingdom
M_25-34_!native	Males between 25- and 34-years-old, fluent, but non-native, speakers of English

Table 6.1: Summary of the notation used to denote listener treatment groups, as discussed in Section 6.3.2.

the influence of country, contained F_25-34_US listeners. Finally, the forth one, studying the influence of English language nativeness, contained female listeners, between 25- and 34-years old, whose first language was not English, but declared to be fluent in English. We denote them as F_25-34_!native;

- Similarly, if the tweet came from a M_25-34_US speaker, the four groups were M_>40_US (age flipped), F_25-34_UK (gender flipped), M_25-34_UK (country flipped), and M_25-34_!native (non-native, but fluent, speakers of English).

We introduce a condensed notation similar to the one employed when addressing RQ 4.1. In a sentence where the specific background of the speaker is not specified, we use *list=speak-[factor]* to refer to the treatment group that contains listeners that have the same background as the speaker, except for the specific *factor* being flipped. For instance, if speaker background is F_25-34_UK, *list=speak-age* denotes the F_>45_UK group of listeners, *list=speak-gender* denotes the M_25-34_UK group, *list=speak-country* denotes the F_25-34_US group, and *list=speak-native* denotes the F_25-34_!native group. Notational conventions are summarised in Table 6.1 (treatment group notation) and Table 6.2 (condensed notation). We looked at the performance of each of the four groups in capturing sarcasm as intended by the speakers. We then compared the performance of each group to that achieved by the *list=speak* group. For instance, for F_25-34_UK speakers, when the factor of investigation was *age*, we looked at how well the perceived labels provided by F_25-34_UK listeners (same background as the speakers) matched the intended labels, compared to those provided by F_>45_UK listeners (age flipped). This allows us to quantify the influ-

Speaker	list=speak	list≠speak
F_25-34_UK	M_ >45_US	F_>45_UK
M_25-34_US	F_>45_UK	M_>45_US

(a)

Speaker	list=speak-age	list=speak-gender	list=speak-country	list=speak-native
F_25-34_UK	F_25-34_UK	M_25-34_UK	F_25-34_US	F_25-34_!native
M_25-34_US	M_25-34_US	F_25-34_US	M_25-34_UK	M_25-34_!native

(b)

Table 6.2: Summary of the condensed notation used to refer to listener treatment groups across speaker backgrounds, as discussed in Section 6.3.2.

ence of age in whether sarcasm is perceived as intended. Doing this for all factors of interest allows us to make a statement about RQ 4.2, as discussed in Section 6.4.2.

RQ 4.3

RQ 4.3 asks whether, in a sarcastic exchange, the socio-demographic similarity between interlocutors is still influential on the ability of the listener to detect the speaker's sarcasm, in the specific situation when the listener has access to contextual information surrounding the speaker's utterance.

To investigate this, we first modified our label collection survey. The new version no longer showed tweet texts, but showed links to the corresponding tweets on Twitter. When labelling a tweet, we invited the listeners, i.e. survey participants, to consider not only the text of the tweet pointed to by the link, but also: surrounding tweets; and any other contextual information that they might find, either on the timeline, or on the profile of the speaker, i.e. the Twitter user who posted the tweet being labelled. We used two strategies to verify that participants actually looked at contextual information. First, note that the modified survey only showed tweet links. They had to click the link, which would open a new tab in their web browser where they could see the text of the tweet on Twitter. Second, we manually checked the average response time per survey, which was around seven minutes longer for the modified survey, compared to the original survey.

We published the modified survey on Prolific Academic and collected perceived

sarcasm labels for all tweets in our dataset from all six treatment groups mentioned thus far, with 3 labels per tweet from each group. When the specific background of a speaker is not relevant, we refer to listener treatment groups using a similar naming convention as above, while adding the prefix "cont:", as an abbreviation of "context". Then, the six groups are *cont:list=speak* (listener group with the same background as the speaker, irrespective of what the background of the speaker is), *cont:list≠speak* (listener group with background disjoint to that of the speaker), *cont:list=speak-age*, *cont:list=speak-gender*, *cont:list=speak-country*, and *cont:list=speak-native* (listener groups with the same background as the speaker except for flipping the factor of study, i.e. age, gender, country, or English language nativeness). Note that, for instance, *list=speak* and *cont:list=speak* refer to the same treatment group. The "cont:" prefix simply underlines that the group was asked to label tweets while also considering contextual information. As such, our group naming convention now includes an extra semantic layer, mainly a specification of the experimental setting in which labels were collected from that group: the absence of "cont:" indicates that the listeners were only shown tweet texts, while the inclusion of "cont:" indicates that they were shown tweet links and were asked to consider contextual information.

We compared intended labels with perceived labels from the *cont:list=speak* and *cont:list≠speak* groups across our dataset, to see if listener socio-demographic background still made a difference when contextual cues to speaker intent, found on speaker Twitter profiles, were considered by listeners. Next, to study the potential influence of individual factors, we compared *cont:list=speak* perceived labels to perceived labels from the *cont:list=speak-age*, *cont:list=speak-gender*, *cont:list=speak-country*, and *cont:list=speak-native* groups. This allows us to make a statement about RQ 4.3, as discussed in Section 6.4.3.

***

In summary, there are granular research questions that we address, RQ 4.1, RQ 4.2, and RQ 4.3. To address the first two of these, for each of the two speaker backgrounds, we collect labels from six treatment groups with 3 labels per tweet collected from each group to account for labelling noise. That amounts to 36 labels for each of the 120 tweets in our dataset. To address the third of these questions, we collect 36 further labels per tweet from the same treatment groups, in a different experimental setting, where listeners are shown tweet links instead of tweet texts. That amounts to 72 labels for each tweet, giving a total of 8,640 labels collected.

6.3.3 F-score to Quantify Performance

As we saw, we have several treatment groups, each group including listeners of a specific socio-demographic background. Answering the granular research questions of this chapter requires us to quantify the performance of each treatment group in detecting sarcasm in our dataset, as intended by the speakers. Quantifying this performance reduces to checking the match between the perceived labels that the group provides, and the intended labels provided by the speakers. Consider the following example.

Assume we have collected 4 sarcastic and 2 non-sarcastic tweets in our dataset. Let $l^{(i)} = [1,0,0,1,1,1]^T$ be the vector of intended sarcasm labels for these tweets, each position corresponding to a tweet, where 0 denotes the absence of sarcasm and 1 the denotes its presence in that tweet. We then collect, for these tweets, the vector $l^{(p)} = [1,0,1,0,0,0]^T$ of perceived sarcasm labels from a listener (annotator).

In this scenario, a straightforward measure of performance is accuracy. That is, the ratio of the number of correct perceived labels (i.e. perceived labels that are 1 for tweets intended sarcastic and 0 for those intended non-sarcastic) to the number of tweets in the dataset. In our example, that would be $2/6$, as the intended and perceived labels only match for the first two tweets. However, accuracy can be misleading in scenarios such as ours, when dealing with imbalanced data, i.e. data where not all classes have the same number of representatives. Recall that we have a ratio of 1:4 of sarcastic to non-sarcastic tweets in our dataset. To see this, say a listener carelessly labels all tweets they see as sarcastic. In such a scenario, the accuracy achieved by that listener in our example above would be $4/6$, giving us false confidence in the ability of that listener to recognise intended sarcasm.

To avoid such a scenario, we use f-score instead of accuracy as a measure of the match between intended and perceived labels. F-score is, to our knowledge, the most popular metric used to measure the performance of classification systems in machine learning and natural language processing literature, due to its robustness to imbalanced data.

We now describe how f-score is computed. We start by defining the precision of sarcasm detection as:

$$p = \frac{\sum_{n=1}^{6} l_n^{(i)} l_n^{(p)}}{\sum_{n=1}^{6} l^{(p)}}.$$

That is, the ratio of the number of times the listener said the a tweet was sarcastic and was correct (i.e. the perceived label was the same as the intended label), divided by the number of times they said it was sarcastic. In our example above $p = 1/2$. Next we

define recall as:

$$r = \frac{\sum_{n=1}^{6} l_n^{(i)} l_n^{(p)}}{\sum_{n=1}^{6} l^{(i)}}.$$

That is, out of the total number of tweets intended as sarcastic, how many the listener got right. In our example $r = 1/4$. Finally, we define f-score as:

$$f = \frac{2pr}{p+r}.$$

That is, the f-score is the harmonic mean of precision and recall. Note that f-score penalises large differences between precision and recall, making it robust to imbalanced data. In our example, $f = 0.333$. The higher the f-score, the better the listener is at perceiving sarcasm as intended by the speaker.

6.3.4 Randomization Test to Compare Performance

As we just saw, for each treatment group, we can compute the f-score between intended sarcasm labels and the perceived labels provided by that group. We interpret that f-score as a numerical summary of the performance of the group in capturing intended sarcasm. As such, given two groups, we can compute the corresponding f-scores, and quantify the difference in performance between the two groups as the numerical difference between the f-scores. However, in our experiments, while we do control the socio-demographic background of each treatment group, there are many other factors that we are unable to control, for instance the level of focus of each listener, or their honesty. One attempt to account for such sources of noise is the fact that we collect three separate labels from each group. However, this does not provide sufficient grounds to believe noise is no longer a threat. As such, we would like to make a rigorous statement about the significance of the difference in f-score, in a framework that deals with uncertainty by design. The standard way to accomplish such a task is to use a statistical test of significance. In this work we use a randomisation test (Noreen, 1989). Yeh (2000) argue for the appropriateness of the randomisation test when the metric of investigation is f-score. We encourage the interested reader to consult their work. Following them, we use a p-value threshold of 0.05, and our null hypothesis states that the difference is not significant.

For brevity, in the rest of the paper, we employ the convention of omitting references to the randomisation test when characterising the difference between the performance of two treatment groups as significant or not. As such, we will say “the

difference is significant" to mean "the difference is statistically significant under a randomisation test with $0.01 < p \leq 0.05$". Similarly, we will say "the difference is very significant" when $p \leq 0.01$.

6.4 Results and Analysis

Our results are reported in two tables. Table 6.3 shows, for each speaker, the precision, recall, and f-score achieved by the six treatment groups that we consider in the first experimental setting, when shown tweet texts for labelling. Table 6.4 shows, for each speaker, the f-score achieved by the same treatment groups in the second setting, when only shown tweet links and asked to consult contextual information on Twitter. As discussed in Section 6.3.2, information from Table 6.3 will help us address RQ 4.1 and RQ 4.2, while that in Table 6.4 will be used to address RQ 4.3.

The first row in each table shows the results achieved by the group *list=speak*, i.e. the group of listeners who have the same socio-demographic background as the speakers. The next five rows show the results achieved by the other groups. In these five rows, the value of a metric (precision, recall, or f-score) could be shown with an "*" symbol appended. This indicates a significant difference between the value achieved by the corresponding treatment group and the value achieved by *list=speak* for that metric (c.f. discussion on statistical significance, Section 6.3.4). If two "*" symbols are appended, the difference is very significant. For instance, in Table 6.3, the second row shows the results for the treatment group *list≠speak*. When speaker background is F_25-34_UK, *list≠speak* denotes the group M_>45_US. We notice that this group achieves a precision of 0.455, which is very significantly different to that achieved by the group *list=speak*, i.e. F_25-34_UK, of 0.648.

Below there are three subsections, one for each of our granular research questions. Each subsection discusses in detail those results that are relevant for addressing the corresponding granular research question.

6.4.1 Answering RQ 4.1: Does Socio-demographic Background Identity Have an Influence?

To address this question, we consider the first two rows from Table 6.3, corresponding to treatment groups *list=speak* and *list≠speak*.

Consider speaker background F_25-34_UK first. In this case, *list=speak* denotes

group F_25-34_UK of listeners, and *list≠speak* denotes group M_>45_US. We notice a very significant drop in precision from 0.648 to 0.455 between the first and the second group ($p = 0.00005$), but a small, insignificant drop in recall from 0.633 to 0.622 ($p = 0.499$). This amounts to a very significant drop in f-score from 0.640 to 0.526 (0.002). This suggests a very significant influence of the socio-demographic factors of investigation on the ability of listeners to perceive sarcasm as intended by female speakers from the UK between 25- and 34-years-old. The very significant variation in precision, with insignificant variation in recall, could suggest a higher predisposition of M_>45_US listeners to classifying a tweet as sarcastic, compared to F_25-34_UK listeners.

Next, consider speaker background M_25-34_US. In this case, *list=speak* denotes group M_25-34_US of listeners, and *list≠speak* denotes group F_>45_UK. We notice a significant drop in precision from 0.460 to 0.356 (p=0.011) between the first and the second group, but an insignificant drop in recall from 0.511 to 0.467 ($p = 0.307$). This amounts to a drop in f-score from 0.484 to 0.404 that is still insignificant. Overall, there seems to be a significant influence of the socio-demographic factors of investigation when precision is the metric of interest. Similarly to the previous paragraph, the lower precision and the insignificant variation in recall could suggest a higher predisposition of F_>45_UK listeners to classifying a tweet as sarcastic, compared to M_25-34_US listeners. Considering the information in both paragraphs, it seems that listeners over 45-years-old, irrespective of gender and country, show a higher predisposition to considering a tweet sarcastic.

Comparing the results across the two speaker backgrounds, for *list=speak* listeners, we notice a further aspect. Mainly, F_25-34_UK listeners labelled tweets coming from F_25-34_UK speakers with a higher f-score of 0.640, compared to only 0.484 achieved by M_25-34_US listeners when labelling tweets coming from M_25-34_US speakers. UK females seems to be more able to detect each other's sarcasm, compared to US makes.

To sum up, interlocutors with identical socio-demographic backgrounds seem significantly more able to detect each other's sarcasm, compared to interlocutors of dissimilar backgrounds. This provides significant statistical ground for positively answering RQ 4.1. Furthermore, as side effects of our experiment, we noticed a higher predisposition of older listeners to interpret a tweet as sarcastic, and more sarcasm detection ability between UK females, than between US males.

6.4.2 Answering RQ 4.2: Which Socio-demographic Factors are Most Influential?

To address this question, we consider the performance of each of the treatment groups *list=speak-age*, *list=speak-gender*, *list=speak-country*, and *list=speak-native*, found in the last four rows in Table 6.3, to that of the group *list=speak*, found in the first row. We are interested in how the performance changes as we flip each of the factors of interest.

Consider speaker background F_25-34_UK first. In this case, *list=speak* denotes treatment group F_25-34_UK, and *list=speak-age* denotes group F_>45_UK. We notice a very significant drop in precision from 0.648 to 0.483 ($p = 0.0005$) between the two groups, an equal recall of 0.633, amounting to a significant drop in f-score from 0.640 to 0.548 ($p = 0.017$). Here, listener age seems to exert a significant influence on their sarcasm detection ability. Looking at the next treatment groups, *list=speak-gender* which denotes M_25-34_UK, and *list=speak-country* which denotes F_25-34_US, we do not notice any significant difference. We find the lack of a significant effect of county particularly intriguing. It seems that US females are statistically just as able to recognise the sarcasm of UK females as other UK females are. UK females may be using a flavour of sarcasm that is more apparent to listeners of both nationalities. English language nativeness, on the other hand, seems significantly influential. Looking at the last row, we notice a very significant drop in the precision achieved by F_25-34_UK listeners, compared to that achieved by F_25-34_!native listeners, from 0.648 to 0.491 ($p = 0.0003$). The change in recall is insignificant, from 0.633 to 0.622. The overall drop in f-score from 0.640 to 0.549 is very significant ($p = 0.01$).

Next, consider speaker background M_25-34_US. In this case, *list=speak* denotes treatment group M_25-45_US, and *list=speak-age* denotes group F_>45_UK. Interestingly, for tweets posted by speakers of the current background, we do not notice any significant influence of listener age. Sarcasm detection ability seems similar between younger US males, and between younger and older US males. Gender, on the other hand, seems to have a significant influence when speaker background is M_25-34_US. Indeed, comparing the performance of *list=speak* which here denotes M_25-34_US, to that of *list=speak-gender*, which here denotes F_25-34_US, we notice no significant change in precision, but a significant increase in recall from 0.511 to 0.633 ($p = 0.034$). Young US females seem to be better at pointing out the sarcasm of young US males than other young US males are. Country does not seem to have an influence. The

next factor with a significant influence is English language nativeness. Indeed, we notice a very significant drop in precision between M_25-34_US and M_25-34_!native treatment groups, from 0.460 to 0.355 ($p = 0.01$).

To sum up, age seems to very significantly impact sarcasm detection ability between UK females, but not between US males. That is, among UK females, age determines a social partitioning, perhaps each partition being characterised by a specific flavour of sarcasm. This does not seem to be the case among US males. On the other hand, sarcastic communication between genders seems to be more efficient in the UK compared to the US. Country seems to not be influential. English language nativeness, on the other hand, does have a significant impact, irrespective of the speaker background considered. Our results provide statistical grounds for answering RQ 4.2 in the following way. In an exchange, similarity between age, gender, and English language nativeness of the interlocutors, could have a significant influence on their ability to detect each other's sarcasm. Consulting the corresponding p-values, in our experiment, age was the most influential, followed by English language nativeness, and gender.

6.4.3 Answering RQ 4.3: Are Socio-demographic Factors Influential When Context is Provided?

To address this question, we consult Table 6.4. We compare the performance achieved by the treatment groups *cont:list≠speak*, *cont:list=speak-age*, *cont:list=speak-gender*, *cont:list=speak-country*, and *cont:list=speak-native*, found in the last five rows, to that of the group *cont:list=speak*, found in the first row. We are interested in whether there is any significant performance variation between *cont:list=speak* and any of the other five treatment groups. If there is, this would indicate that socio-demographic factors may still have an influence, even in the second experimental setting where listeners were only shown tweet links and were asked to consider contextual information found on Twitter.

Consider speaker background F_25-34_UK first. In this case, *cont:list=speak* denotes treatment group F_25-34_UK, and *cont:list≠speak* denotes group M_>45_US. We notice a significant drop in precision between the two groups from 0.575 to 0.504 ($p = 0.04$), with no significant changes in recall and f-score. The drop in precision is less, however, that it was in the first experimental setting, when listeners were shown tweet texts. The availability of contextual information seems to have alleviated, but not eliminated, the influence of listener socio-demographic traits on their ability to recog-

nise sarcasm as intended by the speakers. Let us consult the last four rows of Table 6.4 to see which traits remain influential. Comparing *cont:list=speak-age*, which here denotes F_>45_UK, to *cont:list=speak*, we notice a very significant drop in precision, from 0.575 to 0.471 ($p = 0.005$), an insignificant drop in recall, amounting to a very significant drop in f-score from 0.640 to 0.540 ($p = 0.003$). While the drop in precision is still less than it was in the first experimental setting, age remains a decisive factor. As in the first setting, gender and country are not significant. Unlike the first setting, however, the influence of English language nativeness of the listeners seems to have been eliminated by allowing listeners access to contextual information. Indeed, the change in precision, recall, and f-score, between *cont:list=speak* and *cont:list=speak-native* is no longer statistically significant in this experimental setting.

Next, consider speaker background M_25-34_US. In this case, we notice that the presence of contextual information has, statistically, eliminated the influence of socio-demographic factors. Listener background does not seem to significantly influence the listener's ability to understand the sarcasm of M_25-34_US speakers when context is present. Granted, no listener does a particularly remarkable job, as the maximum f-score achieved by any group is less than 0.6. The important note is, however, that contextual information seems significantly more indicative of sarcasm produced by M_25-34_US speakers, than of that produced by F_25-34_UK speakers, as it is able to eliminate the influence of all socio-demographic factors. Perhaps Twitter users from the United States disclose more public information on their profiles than users from the United Kingdom do.

To sum up, when context is available, age seems to very significantly impact the ability of UK females to detect each other's sarcasm, but not that of US males. The impact of all the other socio-demographic factors investigated seems to be eliminated by the presence of context. This is the answer that our experiment suggests to RQ 4.3.

6.5 Discussion

In this section we summarise the answers that Section 6.4 suggests to our research questions, discuss what implications these answers could have for future work, and conclude with what we believe to be key takeaways from this paper.

6.5.1 Answers to Research Questions

In Section 6.1 we introduced three granular research questions. To our knowledge, our work is the first to provide a quantitative investigation into these questions. Furthermore, we believe to also be the first to quantitatively investigate such questions through the lens of social media data.

RQ 4.1 asks if interlocutors with identical backgrounds more able to detect each other's sarcasm, compared to interlocutors with disjoint backgrounds. The investigation in Section 6.4.1 suggests a positive answer to this question. The socio-demographic factors that we investigated, age, gender, country, and English language nativeness, had a statistically significant influence. As a side effect of that investigation, we noticed a higher sarcasm detection ability between UK females than between US males. Furthermore, we argued that UK females may use a more apparent flavor of sarcasm, recognised better by all listeners. One could view this as evidence in support of Utsumi's Implicit Display Theory (Utsumi, 2000) (c.f. Section 6.2.1) in that sarcasm is prototype-based category. That is, there is a concept of prototypical sarcasm which utterances can express to varying degrees. In other words, an utterance can be more or less sarcastic.

RQ 4.2 asks about the influence of individual socio-demographic factors. The investigation in Section 6.4.2 suggests that the most influential factor is age, followed by English language nativeness, and gender.

RQ 4.3 asks whether the presence of contextual information alleviates the influence of the factors discussed. The investigation in Section 6.4.3 suggests that age similarity remains influential on the ability of UK females to detect each other's sarcasm, but not on that of US males. The influence of all other socio-demographic factors seems to be eliminated. We also noted that contextual information seems to be more indicative of the sarcasm produced by US males, than of that produced by UK females, perhaps suggesting that US males disclose more information on their Twitter profiles than UK females do.

6.5.2 Key Takeaways

Here we summarise what we believe to be the key takeaways. Our results indicate that, interlocutors with similar socio-demographic backgrounds more able to detect each other's sarcasm, compared to interlocutors with dissimilar backgrounds. This suggests that such background information should be considered in the design of future

social analysis tools that either study sarcasm directly, or look at related phenomena where sarcasm may have an influence (Maynard and Greenwood, 2014), such as the expression of sentiment, emotion, and hate-speech.

We provided a statistical methodology for comparing the significance of specific socio-demographic factors. The most influential factors were age, English language nativeness, and gender, in this order. We also showed that public Twitter information can provide enough contextual cues to speaker intent to eliminate the influence of all socio-demographic factors investigated except for age. Again, this suggests that such contextual cues should be considered in the design of future social investigations of sarcasm or related phenomena.

We made observations regarding the online social ecology surrounding sarcastic discourse. However, we believe future qualitative investigation that is out of the scope of this paper (i.e. not directly related to our research questions) is necessary to verify these observations. Mainly, we noted a higher sarcasm detection ability between UK females, than between US males. We also noted that UK females may use a more apparent form of sarcasm than US males, that is easier to detect for listeners of both nationalities. Consistent with this, we observed contextual information to be more indicative of the sarcasm of US males. Our results also suggested a higher sarcasm detection ability across genders in the UK, compared than in the US; and a higher ability across age groups in the US, than in the UK.

Finally, the fact that sarcasm used by UK females seemed easier to detect for listeners of both nationalities could be an argument in favour of Utsumi's theory of sarcasm (Utsumi, 2000) (cf. Section 6.2) in that there is a concept of prototypical sarcasm which utterances can express to varying degrees. As in the previous paragraph, however, we believe these observations require further qualitative investigation that is out of the scope of this paper.

6.5.3 Implications for Future Work

We discuss two main ways in which we believe our work could inform future research, and suggest potential ways forward.

Design of Social Analysis Tools As discussed in the previous section, our findings indicate that both the socio-demographic factors investigated, and public social information, may be informative in design of social analysis tools that investigate sarcasm or related phenomena. These tools include, but are not limited to, sarcasm detection

models. We suggest a few ways in which all this information may be procured when exploring the Twitter network, given the popularity of tweet datasets. Public social information is easily accessible manually, or programatically, using the Twitter Application Programming Interface (API). The socio-demographic factors are usually either available, or can be inferred from, public profile information. If inference is necessary, Chamberlain et al. (2017) suggest how to infer age of Twitter users based on whom they follow. Li et al. (2018) use a Bayes model coupled with a convolutional network to infer the location of timeline tweets. The country from which most timeline tweets originate may be considered the user's country. Sayyadiharikandeh et al. (2016) use a boosted stacked classifier to detect gender of Twitter users. English language nativeness could be deduced from the language of most timeline tweets, in conjunction with (available or inferred) user location. Once these factors are inferred, they can be either manually explored, or encoded in a computational framework. If encoding is required, one could identify a certain trait with the embedded representation of a set of tweets that come from users who posses that trait. For instance, the trait of being female could be encoded as the joint embedding of a set of tweets that all come from female users. The embedding could be built, for instance, using the ParagraphVector model (Le and Mikolov, 2014).

Usage of the Experimental Setup for Analysing Other Phenomena Our experimental setup could be used to study how the socio-demographic traits of interlocutors influence the usage and interpretation of other linguistic phenomena, such as metaphors; or of social phenomena, such as hate speech and fake news. To this end, we provide the web application we developed that host our surveys for data collection and labelling[1].

6.6 Summary

In this chapter we have considered how sarcastic exchanges can be influenced by the socio-demographic backgrounds of the interlocutors engaged. We asked whether identical backgrounds lead to higher sarcasm detection ability between interlocutors, which socio-demographic factors have the most influence on this ability, and whether the influence is alleviated by the presence of contextual information.

Consulting linguistic theories of sarcasm, as well as sociolinguistic studies of sar-

[1] https://github.com/silviu-oprea/f9.

castic communication, we chose four factors for investigation: gender, age, country, and English language nativeness. For our experiments, we collected sarcastic tweets from Twitter users who posted them, users whom we referred to as speakers. Such tweets were implicitly labelled by the users themselves, labels that we referred to as intended sarcasm labels. We then had third-party annotators, whom we referred to as listeners, further label these tweets for sarcasm. We referred to the resulting labels as perceived sarcasm labels. We compared intended and perceived labels using f-score as a quantifier for similarity. Our results indicate that age, English language nativeness, and gender are statistically influential. The influence of age is maintained even when contextual information is available. We suggest that these factors, along with public social information, should be included in the future design of social analysis tools that either investigate sarcasm directly, or look at related phenomena where sarcasm may have an influence, such as the expression of sentiment, emotion, and hate-speech. We also made observations regarding social behaviour. We noted a higher sarcasm detection ability across genders in the UK, than in the US, and a higher ability across ages in the US, than in the UK. Furthermore, we noted that UK females may use a more apparent form of sarcasm, compared to the more subtle sarcasm of US speakers. Finally, contextual information seemed more indicative of the sarcasm of US males than of that of UK females.

	speaker F_25-34_UK			
	listener	**precision**	**recall**	**f-score**
list=speak	F_25-34_UK	0.648	0.633	0.640
list≠speak	M_>45_US	0.455**	0.622	0.526*
list=speak-age	F_>45_UK	0.483**	0.633	0.548*
list=speak-gender	M_25-34_UK	0.610	0.678	0.642
list=speak-country	F_25-34_US	0.582	0.633	0.606
list=speak-native	F_25-34_!native	0.491**	0.622	0.549**

(a)

	speaker M_25-34_US			
	listener	**precision**	**recall**	**f-score**
list=speak	M_25-34_US	0.460	0.511	0.484
list≠speak	F_>45_UK	0.356*	0.467	0.404
list=speak-age	M_>45_US	0.477	0.578	0.523
list=speak-gender	F_25-34_US	0.483	0.633*	0.548
list=speak-country	M_25-34_UK	0.422	0.544	0.476
list=speak-native	M_25-34_!native	0.355**	0.544	0.430

(b)

Table 6.3: Experimental results addressing RQ 4.1 and RQ 4.2. In the first column of each subtable above we show the name of each treatment group. Each subtable corresponds to one speaker background. For each background, we shown precision, recall, and f-score results achieved by each treatment group. "*" indicates a significant difference (p-value threshold of 0.05) between the value achieved by the corresponding treatment group and the one achieved by *list=speak*. "**" indicates a very significant difference (p-value threshold of 0.01).

	speaker F_25-34_UK			
	listener	**precision**	**recall**	**f-score**
cont:list=speak	F_25-34_UK	0.575	0.722	0.640
cont:list≠speak	M_>45_US	0.504*	0.744	0.601
cont:list=speak-age	F_>45_UK	0.471**	0.633	0.540**
cont:list=speak-gender	M_25-34_UK	0.583	0.622	0.602
cont:list=speak-country	F_25-34_US	0.606	0.700	0.649
cont:list=speak-native	F_25-34_!native	0.500	0.722	0.591

(a)

	speaker M_25-34_US			
	listener	**precision**	**recall**	**f-score**
cont:list=speak	M_25-34_US	0.431	0.589	0.498
cont:list≠speak	F_>45_UK	0.406	0.644	0.498
cont:list=speak-age	M_>45_US	0.403	0.578	0.475
cont:list=speak-gender	F_25-34_US	0.483	0.644	0.552
cont:list=speak-country	M_25-34_UK	0.451	0.567	0.502
cont:list=speak-native	M_25-34_!native	0.408	0.667	0.506

(b)

Table 6.4: Experimental results addressing RQ 4.3. In the first column of each subtable above we show the name of each treatment group. Each subtable corresponds to one speaker background. For each background, we shown precision, recall, and f-score results achieved by each treatment group. "*" indicates a significant difference (p-value threshold of 0.05) between the value achieved by the corresponding treatment group and the one achieved by *cont:list=speak*. "**" indicates a very significant difference (p-value threshold of 0.01).

Chapter 7

Should a Chatbot be Sarcastic? Understanding User Preferences

In this chapter we address RQ 5 of this thesis. The work presented herein leads to the Contribution 4 of this thesis. This chapter is based on two papers that we published. The first one is "Chandler: An Explainable Sarcastic Response Generator"[1], published in the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Oprea et al., 2021). The second one is "Should a Chatbot be Sarcastic? Understanding User Preferences Towards Sarcasm Generation", published in the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Oprea et al., 2022).

7.1 Introduction

The presence of sarcasm in online communication has motivated an increasing number of computational investigations of sarcasm across the community in recent years. The survey of Băroiu and Trăușan-Matu (2022) shows this trend. Most investigation focus on sarcasm detection. We have overviewed some of these in Section 2.3.

A recent set of investigations consider the task of sarcasm generation. That is, the task of creating chatbots, i.e. dialogue systems, able to generate sarcastic utterances. Approaches to sarcasm generation introduced so far (Joshi et al., 2015a; Mishra et al., 2019; Chakrabarty et al., 2020) are mainly motivated by the potential to create more approachable, human-like conversational agents, considering that sarcasm is a natural part of human discourse. In this chapter, we suggest reconsidering this motivation, as a community, for two reasons.

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

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

[1] Since publishing this paper, we renamed our sarcastic response generator from *Chandler* to *Max*.

In response, we suggest it is imperative, not least from an ethical perspective, to consider the following research question.

RQ5 In what conversational context is it appropriate for a chatbot respond sarcastically, and how should it formulate sarcasm such that it is understood by humans?

We divide this question into the following, more granular, research questions:

RQ5.1 When should a chatbot be sarcastic?

(a) When do humans consider sarcasm appropriate?

(b) When do humans prefer sarcasm, over non-sarcasm?

RQ5.2 How should a chatbot formulate sarcasm?

(a) What linguistic devices do humans associate with sarcasm?

(b) What sarcasm flavour do humans prefer?

Here, by *flavour*, we mean a specific conjunction of linguistic devices that humans may associate with sarcasm, such as intensifiers and emotional markers, as expanded upon in Section 7.3.

To address our research questions, we suggest the following approach. First, given a set of input utterances, generate several sarcastic responses. Each response should be of a specific sarcasm flavour, i.e. should display a specific conjunction of linguistic devices. Next, create a survey that asks human participants: to indicate how appropriate it was to respond sarcastically to the input; to select their preferred response; and to rate the sarcasticness of each response, investigating whether they associate the linguistic devices in the response with sarcasm.

To achieve this, we require a sarcastic response generator that provides control over the linguistic devices used. We argue that previous generators are suboptimal. This is because they are grounded in linguistic theories which imply that the presence of sarcasm in an utterance is signalled by linguistic incongruity. However, as discussed in Section 2.1, incongruity is a device that is not sufficient, for sarcasm to occur. In response, we introduce Max, a novel modular sarcastic response generator. It is grounded in the Implicit Display Theory (IDT), introduced in Section 2.1.5. The IDT focuses specifically on making the distinction between sarcasm and non-sarcasm. It specifies devices whose presence is both necessary and sufficient for sarcasm to occur. These are allusion to a failed expectation, pragmatic insincerity, and emotional markers. Max can generate sarcasm of different flavours, and allows control over flavour its output should reflect. Herein, we also compare Max's outputs to those of previous generators,

to examine the preference of participants towards an even greater range of sarcasm flavours.

Our results indicate that people find sarcastic responses inappropriate for most input utterances. When sarcasm was considered appropriate, the inputs commonly had a positive sentiment, and often had elements of humour. Further, even when considered appropriate, people still did not usually *prefer* sarcastic responses over non-sarcastic ones. Sarcasm was typically preferred when it was also considered funny and not too specific. Finally, we identified pragmatic insincerity and emotional markers (cf. Section 2.1.5) as crucial linguistic devices to include in generating recognizable sarcasm.

The rest of this chapter is organised as follows. Section 7.2 overviews previous sarcasm generators and discusses their limitations. Section 7.3 introduces the methodology we employ to address the granular research questions of this chapter. This includes a description of Max, and of the experiments we conducted. Section 7.4 reports the results of the experiments, answering our research questions in light of these results. Section 7.5 summarises these answers. Section 7.6 summarises the chapter.

7.2 Previous Sarcasm Generators

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

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

module constructs the sarcastic paraphrase from $u^{(+)}$ and $v^{(-)}$.

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

Previous generators share a limitation that make them unfit for our purposes. Mainly, they identify sarcasm with linguistic incongruity. Thus, they only provide this single device for investigation, device that is not sufficient for sarcasm to occur, as discussed in Section 2.1, particularly in Section 2.1.1 and Section 2.1.2. A further limitation, shared by Mishra et al. (2019) and Chakrabarty et al. (2020), is that their generators only work with input utterances of negative sentiment. However, as discussed earlier, sarcastic communication can have many goals, including to praise, or to strengthen friendships.

7.3 Methodology

In this section we look at the methodology employed to address our research questions. Specifically, we first select a set of input utterances, as discussed in Section 7.3.1. Next, for each input, we generate four sarcastic responses of different flavours using Max, as discussed in Section 7.3.2. and three more responses using other systems, as discussed in Section 7.3.3. Finally, for each input, in a survey, we ask human participants to rate the responses across several dimensions, to understand their preference towards the appropriateness of sarcasm, and which linguistic devices they associate with sarcasm, as further discussed in Section 7.3.3.

7.3.1 Selecting Input Texts

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

ing positive or neutral; and *positive* and *neutral*, partitions that we formed analogously to how we formed the *negative* partition. Our final input dataset contains 250 texts.

7.3.2 Generating Sarcastic Responses

As discussed in Section 2.1.5, the Implicit Display Theory (IDT) focuses specifically on making the distinction between sarcasm and non-sarcasm. Because of this, we chose it to serve as a grounding for Max, our generation system. The IDT directly suggests an algorithm for sarcasm generation that first identifies an ironic environment, then creates an utterance that implicitly displays it. We now discuss how we implement each step.

Ironic Environment As discussed in Section 7.3.1, each input text U_{in} describes an action. In this scenario, herein, we assume the expectation Q that is part of the ironic environment negates that action. For instance, say U_{in} expresses the event $P = [$<user> wins the marathon$]$. We assume $Q = \neg P = [$<user> does not win the marathon$]$. As we shall see, the algorithm we suggest will not, in fact, require us to formulate Q, but it relies on the above assumption.

Allusion to Q Following Utsumi (2000), we define allusion in terms of coherence relations, similar to the relations of rhetorical structure theory (RST) (Mann and Thompson, 1987). That is, if U_α is an utterance that expresses proposition α, we say U_α alludes to the expectation Q if and only if there is a chain of coherence relations from α to Q[2]. So, we need to first select a proposition α to either start or end the coherence chain, then specify the chain between α and Q, and formulate U_α such that it expresses α. We suggest defining such α as objects of if-then relations, where the subject is P, the proposition expressed by input text U_{in}. That is, relations of the form "if P then α" should hold.

To infer α given U_{in}, we use COMET (Bosselut et al., 2019), an adaptation framework for constructing commonsense knowledge. More specifically, COMET is a language model fine-tuned on the task of completing triples from commonsense knowledge bases, such as ATOMIC (Sap et al., 2019) and ConceptNet (Speer et al., 2017). Here, commonsense triple is a representation of a commonsense relation; it consists of a subject, a relation type, and an object. For instance, the subject could be "eating", the relation type could be "requires", and the object could be "being hungry". This relation

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

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

Table 7.1: Coherence chains between the object α of an if-then relation and the failed expectation Q, for each relation type, as discussed in Section 7.3.2. Here, P is the proposition expressed by the input text U_{in}. In the examples, $U_{\text{in}} =$'<user> won the marathon'.

expresses the commonsense fact that eating requires being hungry. COMET is fine-tuned to input the subject of a relation, such as "eating", along with the relation type, such as "requires", and output the relation object, such as "being hungry". We use the COMET variant fine-tuned on ATOMIC, a knowledge base of if-then commonsense relations. In our case, the subject is U_{in}, and we set α to the relation object.

In the examples that follow, assume the input to Max is $U_{\text{in}} =$'<user> won the marathon'. We leverage four relation types found in ATOMIC:

1. *xNeed*: the object α of a relation of this type specifies an action that the user needed to perform before the event took place, e.g. "if U_{in} then $\alpha = [$*xNeed* to train hard]";
2. *xAttr*: the object α specifies how a user that would perform such an action is seen, e.g. "if P then $\alpha = [$*xAttr* competitive]";
3. *xReact*: the object α specifies how the user could feel as a result of the event, e.g. "if P then $\alpha = [$*xReact* happy]"; and
4. *xEffect*: the object specifies a possible effect that the action has on the user, e.g. "if P then $\alpha = [$*xEffect* gets congratulated]".

In Table 7.1 we show, for each relation type, the coherence chains between the relation object α and the failed expectation Q. Under these conditions, to generate an utterance U_α that alludes to Q, we need to choose any U_α that expresses α.

Pragmatic insincerity The second requirement for implicit display is that the utterance generated should include pragmatic insincerity. In this thesis, we focus on violating Grice's maxim of quality (Grice, 1975), where we aim for the propositional content of the generated utterance to be incongruous to that of U_{in}, the input utterance.

Algorithm 1: Sarcastic response generation algorithm, as discusses in Section 7.3.2.

input: utterance U_{in};

ironic environment

Let $Q := \neg P$ be the failed expectation;

implicit display

Choose an if-then relation type τ from *xNeed*, *xAttr*, *xReact*, and *xEffect*;

Let $\alpha = \text{COMET}(U_{\text{in}}, \tau)$;

return response U_{out} that expresses $emotion(\neg\alpha)$;

To achieve this, we first choose an if-then relation type, then infer the relation object α from U_{in} using COMET, and construct an utterance that expresses $\neg\alpha$. For instance, if U_{in} ='<user> won the marathon', and we have chosen the *xAttr* relation type, the constructed utterance could express $\neg\alpha$ = [<user> is not competitive].

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

Logical form and explainability At this point we formulate Algorithm 1 for generating a sarcastic response U_{out}, given an input utterance U_{in} that expresses proposition P. We refer to $emotion(\neg\alpha)$ as the *logical form* of the sarcastic response we generate. Here, *emotion* is a function that augments $\neg\alpha$ to express a negative attitude. Note that the logical form, together with the coherence chain between α and the failed expectation Q, provide a complete explanation for *how* and *why* sarcasm occurs. The explanation is $\varepsilon = (\text{emotion}(\neg\alpha), \mathcal{C})$, where is C the coherence chain from α to Q. The coherence chain for each relation type can be selected from Table 7.1.

Logical Form to Text To convert the logical form to text, we rely on predefined patterns for each if-then relation type. As a running example, assume the input utterance U_{in} ='<user> won the marathon' and the chosen relation type is *xAttr*. Say $\alpha = \text{COMET}(U_{\text{in}}, xAttr) = [xAttr$ competitive$]$. The logical form is $emotion(\neg[xAttr$ competitive$])$. We first construct an intermediate utterance U_α using the following rule: *<user> <verb> competitive*. Here, *<verb>* is a verb specific to each relation type. In our example, U_α could be '<user> is competitive'.

Next, for each input U_{in}, we generate three responses. The first response U_{out}^{-e} only includes pragmatic insincerity, i.e. it expresses $\neg$[*xAttr* competitive]. To construct it, we apply a rule-based algorithm to generate the negation of U_{α} in a manner similar to (Chakrabarty et al., 2020), discussed in Section 7.2. U_{out}^{-e} could be '<user> is not competitive'. The second response U_{out}^{-i} does not include pragmatic insincerity, but only markers that express an emotional attitude, i.e. it expresses *emotion*([*xAttr* competitive]). To achieve this, in a pattern-based manner, we augment U_{α} with hyperbole and interjections, as indicated by Utsumi (2000). U_{out}^{-i} could be '<user> is definitely competitive, yay!'. The third response U_{out} includes both devices, i.e. it expresses *emotion*($\neg$[*xAttr* competitive]). U_{out} could be '<user> is definitely not competitive, yay!'. A full list of patterns is shown in Appendix A.

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

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

7.3.3 Measuring User Preferences

We built three surveys, labelled (a)–(c), that we published on the Prolific Academic[3] crowdsourcing platform, one for each output type, out of U_{out}^{-e}, U_{out}^{-i}, and U_{out}. As such, in the survey corresponding to U_{out}, we presented participants with the input text U_{in}, along with the responses produced by Max-xNeed, Max-xAttr, Max-xReact, and Max-xEffect.

In each survey, we also enclosed a response from DialoGPT (Zhang et al., 2020), a recent dialogue system that is not built to be sarcastic; a response produced by SarcasmBot, the sarcastic response generator of Joshi et al. (2015a)[4]; and a response produced by R^3, the state-of-the-art sarcastic paraphrase generator of Chakrabarty et al.

[3] https://prolific.co

[4] https://github.com/adityajo/sarcasmbot

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

Table 7.2: Responses generated by all systems to the utterance "I ran out of characters :drooling_face:", as discussed in Section 7.3.3.

(2020)[5].

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

Table 7.2 shows an example input utterance, along with responses from all systems.

All in all, each survey instance contained a specific input text, and seven responses generated as mentioned above and presented in a random order. In the survey, we asked participants to evaluate each response across four dimensions:

1. Sarcasm: How sarcastic is the response?
2. Humour: How funny is the response?

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

3. Coherence: How coherent is the response to the input? It is coherent if it sounds like sensible response that a person might give in a real conversation; and
4. Specificity: How specific is the response to the input? It is not specific if it can be used as a response to many other inputs.

Each dimension ranged from 0 to 4, in line with previous work (Chakrabarty et al., 2020). Next, we asked participants to select their preferred response out of the seven, i.e. the one that they would personally use. Finally, we asked them to judge, on a scale from 0 to 4, how appropriate it was to respond sarcastically to the shown input text.

Each survey instance was presented to three different participants. However, we did not use a voting scheme to aggregate the three survey instances into one. Rather, aggregation was conducted per-system. This is because of the inherently subjective nature of our metrics. For instance, recall from Chapter 6 that the level of perceived sarcasticness of an utterance varies with the socio-demographic background of the annotator. As such, there is no objective standard by which to judge that a certain answer provided by an annotator is *correct* or *incorrect*. Indeed, the inter-participant agreement was low, but not surprisingly so, given that participants could have come from different sociocultural backgrounds. However, this does not entail that population statistics are not informative. As related work in this direction, consider that of Amidei et al. (2018), who make the point "an unchecked focus on reduction of disagreement among annotators runs the danger of creating generation goals that reward output that is more distant from, rather than closer to, natural human-like language." (Amidei et al., 2018) Consider also the work of Davani et al. (2021), who discuss the issue of disagreement in subjective tasks. We do, however, encourage more work in this direction.

7.4 Evaluation

We now look at the responses that the participants provided in our survey. Based on these responses, we formulate answers to the research questions asked in Section 7.1.

7.4.1 When should a chatbot be sarcastic?

Let us first reflect upon RQ 5.1a: When do humans consider sarcasm appropriate?

Figure 7.1 shows the mean appropriateness score for each of the five sentiment categories. A one-way ANOVA test between the means yielded a p-value ≈ 0.001.

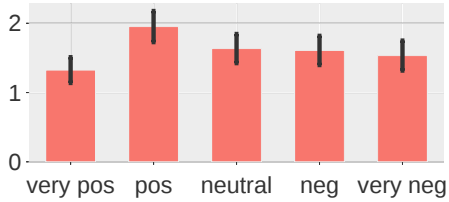

Figure 7.1: Mean sarcasm appropriateness score for each sentiment category, as discussed in Section 7.4.1. The error bars represent 95% confidence intervals.

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

Table 7.3: Example inputs with low sarcasm appropriateness score, as discussed in Section 7.4.1.

We therefore proceeded with Tukey's range test (Tukey, 1949), to find the means that are significantly different from one another. We noticed that sarcasm was considered significantly more appropriate by survey participants in responses to positive inputs, compared to very positive, and very negative inputs, respectively. This supports our statement from Section 7.2: the assumption of previous state-of-the-art generators that sarcasm should *only* be generated for negative inputs is problematic. However, even for the positive class, the mean appropriateness is less than 2. This makes it difficult to recommend responding sarcastically based on sentiment only.

To gain more insight, we proceeded with a qualitative inspection of the inputs that yielded the highest and lowest appropriateness scores, respectively. We noticed a few main themes, that we labelled *joke*, *family*, *school*, *leisure* and *death*. We then asked two humans to label all inputs across these dimensions. A third human resolved all disagreements. Finally, we computed the Pearson correlation coefficient of each theme with the sarcasm appropriateness score, across all inputs. We noticed a significant ($p < 0.05$) positive correlation between appropriateness and the category *joke*, and significant negative correlation with belonging to the *family* theme. We show some examples of the theme *family* with low appropriateness scores in Table 7.3.

Thus, according to our analysis, sarcasm seems to be most appropriate for positive inputs, and for humorous inputs, which may invite more sarcastic responses. In other situations, however, sarcasm might be interpreted as inappropriate and even of-

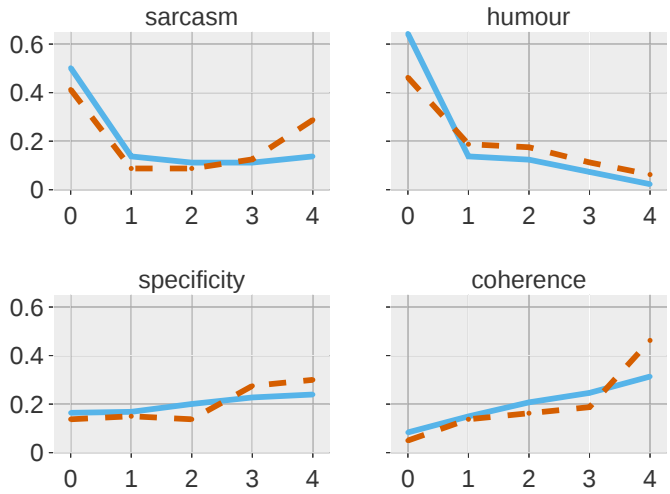

Figure 7.2: Distribution of the sarcasm, humour, specificity, and coherence scores of the *preferred* response; across all survey instances (continuous blue line) and across instances with a high sarcasm appropriateness (dashed red line), as discussed in Section 7.4.1.

fensive (Meaney et al., 2021).

Let us now reflect upon RQ 5.1b: When do humans prefer sarcasm, over non-sarcasm?

We first consider the overall preference towards either sarcasm or non-sarcasm. Recall that participants also specified their preferred response for each input. The distribution of the sarcasm, humour, specificity, and coherence scores of this *preferred* response, across all survey instances, is illustrated in Figure 7.2 with a blue, continuous, line. The red, dashed, line illustrates the distribution across the 80 survey instances where the sarcasm appropriateness score of the input was higher than the midpoint, i.e. at least 3. The same figure also shows the analogous distributions for humour, specificity, and coherence scores of the preferred response.

We notice considerably higher preference towards non-sarcastic and non-humorous responses. As indicated by the blue lines, over 50% of the preferred responses were those considered non-sarcastic and non-humorous by participants, the rest of the distribution being highly skewed towards the lower sarcasm and humour regions. Furthermore, note that even when sarcasm was considered highly appropriate, participants still preferred non-sarcastic responses, as indicated by the red, dashed, line in the top-left of Figure 7.2. Although there is a shift in the distribution towards sarcasm in this case, the skew is still towards the non-sarcastic region. Looking at the bottom row of Figure 7.2, on the other hand, we notice a negative skew, indicating an overall preference

towards higher coherence. This is slightly the case for specificity as well.

To investigate further, we fit a logistic regression model to predict whether a response is preferred based on its sarcasm, humour, specificity, coherence scores, and two-way interactions between these variables. All coefficients are listed in Table 7.4. We noticed noticed a significant ($p < 0.05$) positive relationship between coherence

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

Table 7.4: Detailed results of logistic regression described in section Section 7.4.1.

and preference, as well as the interaction between sarcasm and humour. The term representing the product of sarcasm and specificity had a significant negative effect on preference. In terms of the specific systems, we notice DialoGPT was preferred about 44% of the time, followed by Max-xAttr^{-i} (20%), and SarcasmBot (15%), which corresponds exactly to the coherence ranking in Table 7.5.

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

7.4.2 How Should a Chatbot Formulate Sarcasm

Let us now consider RQ 5.2a: What linguistic devices do humans associate with sarcasm?

In Table 7.5 we show mean sarcasm, humour, specificity, and coherence scores

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

Table 7.5: Means of the sarcasm, humour, specificity, and coherence scores provided by participants, for each variant of Max, as discussed in Section 7.4.1 and Section 7.4.2. "*" indicates statistically significant difference from row (a) within the same numbered group (t-tests with Bonferroni correction, $p < 0.001$).

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

Allusion We have four strategies for alluding to the failed expectation, depending on the relation type considered. We notice the highest sarcasm score is achieved by Max-xAttr (row 2a), followed by Max-xNeed (row 1a), Max-xReact (row 3a) and Max-xEffect (row 4a). The same ranking holds for variants of Max that do not include pragmatic insincerity or emotional markers. Out of the allusion strategies selected, the responses perceived as most sarcastic are those that mention attributes of the user. Similarly, we notice that variants of Max that use the xAttr relation are also perceived

and the most coherent, specific to the input, and achieve the highest humour score.

Pragmatic Insincerity To assess whether annotators associate pragmatic insincerity with sarcasm, we compare mean sarcasm scores of complete versions of Max, denoted by Max-<relation>, with mean sarcasm scores of those versions that omit pragmatic insincerity, denoted by Max-<relation>$^{-i}$. Comparing Max-xAttr (row 2a), with Max-xAttr^{-i} (row 2b), we notice a significant drop in average sarcasm score. We observe a similar trend in group 3 for Max-xReact^{-i}, indicating the importance of pragmatic insincerity. However, this did not hold for the other two relation types. Humour scores do not change significantly when omitting pragmatic insincerity. Additionally, both specificity and coherence seem to significantly increase when removing pragmatic insincerity, irrespective of the relation type considered.

Emotional Markers Comparing complete versions of Max with those that omit emotional markers, we notice that the omission of such markers leads to significantly lower perceived sarcasm for all relation types. For instance, comparing Max-xAttr (row 2a) with Max-xAttr^{-e} (row 2c), we notice a significant drop in average sarcasm score from 2.1 to 1.1. Humour is also significantly impacted by the omission of emotional markers for all relation types considered except for *xEffect* (row 4). Oh the other hand, coherence and specificity are not significantly influenced.

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

Finally, let us consider RQ 5.2b: What sarcasm flavour do humans prefer?

While we established that participants typically preferred non-sarcatic responses, we next look at the sarcasm people preferred in our experiments when they *did* prefer sarcasm. We proceed as follows. We consider the set of survey instances that showed the complete versions of Max, where the sarcasm score given by the participant to their preferred response was at least 3, leaving us with 107 (around 14%) of the 750 survey instances. We divide these instances into five categories, based on input sentiment.

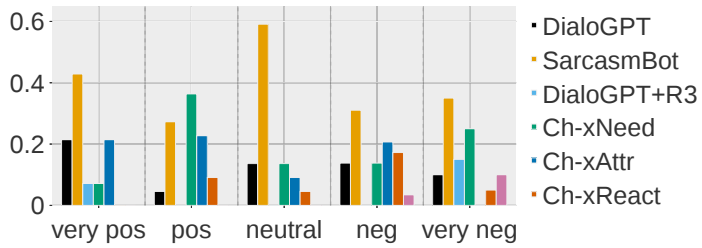

Figure 7.3: Normalized number of times each system was preferred for instances were the participant preferred a response that they also considered sarcastic, as discussed in Section 7.4.2. The ordering of systems shown on the right side of the chart corresponds to the ordering of each bar, within each of the five groups, very positive ("very pos"), positive ("pos"), neutral, negative ("very neg"), and very negative ("very neg").

Within each category, for each generation system, we count the number of times that a response produced by that system was preferred. Figure 7.3 shows the normalised counts across all systems, for each sentiment category.

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

7.5 Recommendations

We now summarise the answers to our granular research questions introduced in Section 7.1 that the work presented in this chapter suggests. We hope that future work on sarcasm generation will benefit from these answers.

To RQ 5.1a we answer that people think sarcasm is *inappropriate* as a response to most inputs. However, if it is to be used, it is seen as most appropriate when the input is positive, but not extremely positive. People also found sarcasm to be a suitable response to jokes. To RQ 5.1b we answer that even when they consider sarcasm appropriate, people usually do not prefer sarcasm. Rather, coherence is the most important factor in explaining their response preferences. When people do prefer sarcastic re-

sponses, it is mainly when they also consider those responses amusing. Further, they generally dislike sarcasm that is very specific. To RQ 5.2a we answer that when generating sarcasm, pragmatic insincerity and emotional markers are important to include as they have a high influence of sarcasm perception. To RQ 5.2b we answer that people commonly prefer the simple, stereotypical sarcastic responses of SarcasmBot, compared to more complex responses. This suggests that a simple solution rule-based solution, such as SarcasmBot, might suffice. However, such solutions might be unable to produce diverse outputs. More investigation is required to examine if repeating similar responses is desirable in long conversations.

7.6 Summary

We introduced Max, a linguistically grounded framework for sarcasm generation. Using it, we generated sarcastic responses with a variety of flavours that we presented to humans, asking them to annotate the responses on four dimensions: sarcasm, humour, coherence, and specificity. Observing the responses allowed us to answer the granular research questions of this chapter.

Annotators found sarcastic responses from chatbots inappropriate in most situations. They might be acceptable when a joking environment has already been established. However, even when deemed appropriate, annotators did not usually prefer sarcastic responses to non-sarcastic ones. Among the sarcastic responses that they did prefer, notable were simple responses that exhibit stereotypical sarcasm. Finally, annotators associated the presence pragmatic insincerity and of emotional markers of with the presence of sarcasm. We hope that the findings of this chapter will inform future work on sarcasm generation, not only in terms of the methodology, but also in terms of the motivation.

We make one final point. We would like to emphasise that our goal in this chapter was not to build a “state-of-the-art” sarcasm generator that “does better” than previous generators. That goal could be the object of future research, pursued while accounting for the findings summarised in the previous paragraph. Besides, sarcasm being a subjective phenomenon, it is really rather unclear what ‘doing better” means. We suggest that future work is needed to clarify this point.

Rather, our focus is stated by the granular research questions of this chapter. Max, the sarcasm generator that we did suggest here, is merely a tool that we used, among other tools, to help us answer these questions. It is not the main contribution of this

chapter. It is faithful to a previous formal linguistic theory of sarcasm, the Implicit Display Theory. While we argued that this theory has advantages over previous ones, it might still have its own limitations that could have propagated into our generator.

Chapter 8

Conclusion

In this thesis, we investigated the phenomenon of sarcasm, using online textual utterances as a lens. We conducted quantitative, computational investigations, but placed our efforts in the context of previous studies of sarcasm in linguistics and sociolinguistics. Our efforts were guided by five research questions. In the process of answering these questions, we brought four main contributions that span two research directions: sarcasm detection, and sarcasm understanding. We summarised these contributions in six published research articles.

Sections 8.1 to 8.4 summarise the contributions. The information presented therein is also illustrated in Figure 1.1 in Chapter 1. Section 8.5 discusses limitations and provide suggestions for future work. Section 8.6 ends the thesis with final thoughts.

8.1 Contribution 1: iSarcasm Dataset and Analysis

In Chapter 3 we suggested contextual sarcasm detection models that we evaluated on two datasets of tweets labelled for sarcasm. Tweets in the first dataset were accompanied by distant supervision labels, while the second one were accompanied by both manual labels, provided by human annotators, and distant supervision labels. While we achieved state-of-the-art results in all experiments, we noticed that, when detecting sarcasm in a tweet, contextual information about the author of that tweet was considerably more predictive of distant supervision labels, than of manual labels. This discrepancy motivated the work presented in Chapter 4. Therein, we argued that, when labelling a tweet, both labelling methods could produce noisy labels that do not coincide with the sarcastic intention of the author of that tweet. In response, we suggested a method that involves the author in the labelling process. Specifically, published a survey on a crowdsourcing platform where we asked Twitter users to provide us with tweets that they themselves had posted in the past. Using this method, we built iSarcasm, a dataset of tweets labelled for sarcasm. Each sarcastic tweet in iSarcasm is accompanied by: an explanation of why it is sarcastic, provided by its author; a rephrase that conveys the same message non-sarcastically, also provided by its author; a further label specifying the category of ironic speech that it belongs to. We then evaluated previous state-of-the-art sarcasm detection models on our dataset, showing they achieve a much lower performance on iSarcasm, compared to the performance reported on previous datasets. We also collected third-party sarcasm labels for the tweets in iSarcasm from human annotators. Human performance was considerably higher than model performance, but still les than 62% f-score. This could indicate that the task of detecting sarcasm in

tweets in challenging even for humans. This work amounts Contribution 1, which we summarise as follows.

Summary of Contribution 1 We showed that both manual labelling and distant supervision could lead to noisy labels. In response, we suggested a new method of labelling tweets for sarcasm, along with iSarcasm, a dataset created using this method. We then showed that previous sarcasm detection models underperform on iSarcasm, and that detection might be challenging even for humans.

We published this work in two articles, one at ACL 2019 (Oprea and Magdy, 2019), describing the work conducted in Chapter 3, and another at ACL 2020 (Oprea and Magdy, 2020b), describing the work conducted in Chapter 4.

8.2 Contribution 2: iSarcasmEval Task and Analysis

The low performance of previous sarcasm detection models on iSarcasm, compared to both their performance on previous datasets, and to human performance, suggested the need for more effective models. In response, as reported in Chapter 5, we crowdsourced the task of building such models at the 16th International Workshop on Semantic Evaluation. In this purpose, we created the iSarcasmEval dataset of texts labelled for sarcasm by their authors. Each sarcastic text is also accompanied by a rephrase, provided by its author, and six further binary labels, indicating the categories of ironic speech that the text belongs to. Note that, in iSarcasm, each tweet was assigned to only one category of ironic speech. However, we subsequently noticed those categories were not mutually exclusive. We addressed this limitation of iSarcasm when building iSarcasmEval, allowing a text to be assigned to multiple categories. The shared task contained three subtasks: sarcasm detection, ironic speech category classification, and pairwise sarcasm identification. Sarcasm identification refers to differentiating the sarcastic text from it's non-sarcastic rephrase. There were 60 teams that participated. We provided a high-level overview of the approaches of the top performing teams, for each subtask. These approaches outperformed our baselines by a margin. However, we argued that the task of detecting sarcasm in text remains challenging, and detecting the ironic speech category even more so. This work amounts to Contribution 2, which we can summarise as follows.

Summary of Contribution 2 We crowdsourced the task of building more effective sarcasm detection models at the 16th International Workshop on Semantic Evaluation.

In this purpose, we built the iSarcasmEval dataset of texts labelled for sarcasm. iSarcasmEval addresses one limitation of iSarcasm, mainly that each tweet in iSarcasm was assigned to only one of the categories of ironic speech, despite these categories not being mutually exclusive.

We published a summary of this work in an article at SemEval 2022 (Abu Farha et al., 2022a).

8.3 Contribution 3: Socio-demographic Insight

As mentioned in Section 8.1, human annotators achieved a considerably higher sarcasm detection performance on iSarcasm, compared to models. However, human performance was still less than 62%, quantified using the f-score. In response, as reported in Section 8.2, we first focused on model performance. Specifically, in an attempt to reduce the discrepancy between model and human performance, we crowdsourced the task of building more effective models.

In Chapter 6 we turned our attention to human performance. The low f-score achieved by humans could indicate that sarcasm detection in text is challenging even for humans. Motivated by this observation, we switched our focus to studying sarcastic exchanges between human interlocutors. We studied the socio-demographic ecology of such exchanges. We found that interlocutors with identical socio-demographic backgrounds were more able to detect each other's sarcasm, compared to those with disjoint backgrounds. Investigating specific factors, we found similarity between age, English language nativeness, and gender, to be significantly influential on the ability of interlocutors to understand each other's sarcasm. This work amounts to Contribution 3, which we can summarise as follows.

Summary of Contribution 3 We conducted a quantitative analysis of the socio-demographic ecology of sarcastic exchanges between human interlocutors. In our experiments, interlocutors with similar socio-demographic backgrounds were more able to detect each other's sarcasm, compared to those with dissimilar backgrounds. Age, English language nativeness, and gender, were most influential on this ability. We suggest that future social analysis tools, including sarcasm detection models, should account for such factors.

We presented this work at CSCW 2020 (Oprea and Magdy, 2020a).

8.4 Contribution 4: Guidelines for Sarcasm Generation

In Chapter 7 we focused on a task recently introduced in the community, mainly sarcasm generation. Previous approaches to sarcasm generation introduced were mainly motivated by the potential to create more approachable, human-like chatbots, considering that sarcasm is a natural part of human discourse. We argued that the community should reconsider this motivation, given the potentially offensive nature of sarcasm. We then set out to determine in what conversational context it would be appropriate for a chatbot to be sarcastic, in what context humans would prefer sarcasm over non-sarcasm coming from chatbots, and how a chatbot should formulate sarcasm such that it is understood by humans, and it is in accordance to their preferences. In this purpose we introduced Max, a linguistically grounded framework for sarcasm generation. Using it, we generated sarcastic responses of different flavours, that we presented to human annotators. Analysing their responses, we noted the following. First, annotators found sarcastic responses inappropriate for most conversational contexts when coming from chatbots. When sarcasm was considered appropriate, it commonly occurred in a context with positive sentiment that often had elements of humour. However, even when considered appropriate, people did not usually prefer sarcastic responses over non-sarcastic ones. When sarcasm was preferred, it commonly had elements of humour, and was more general, as opposed to addressing specific aspects of the conversational context. We also found that when including pragmatic insincerity and emotional markers, people considered the responses more sarcastic, compared to when such linguistic devices were missing from the responses. This work amounts to Contribution 4, which we can summarise as follows.

Summary of Contribution 4 We introduce Max, a novel sarcastic response generator grounded in a formal linguistic theory of sarcasm. We use Max to generate sarcastic responses to a given set of input utterances, and ask human annotators to label each response on several dimensions, including appropriateness and sarcasticness. Studying their responses, we provide guidelines for dialogue systems concerning the appropriateness of generating sarcastic responses, and the formulation of such responses.

We published part of this work in an article at EMNLP 2021 (Oprea et al., 2021), where we demonstrated Max, our sarcasm generation framework. We then published the rest at ACL 2022 (Oprea et al., 2022).

8.5 Limitations and Future Work

In this section we discuss limitations of the work presented in this thesis and suggest future work directions that could address these limitations.

Limitations of the iSarcasm Dataset

First, recall that each tweet in iSarcasm is accompanied by one extra label indicating the category of ironic speech that it belongs to. However, after building the dataset, we noticed that a tweet could exhibit traits that can be associated with more than one category. That is, the categories are not mutually exclusive. We addressed this limitation of iSarcasm when building the iSarcasmEval dataset, as discussed in Chapter 5, and summarised in Section 8.1.

Second, when asking survey participants to provide us with examples of sarcastic tweets that they themselves had posted in the past, the term *sarcasm* was not defined in the survey. We relied on their intuitive understanding of the term. This could be problematic, in light of our conversation from Section 4.2.2. Therein, we pointed out that cultural factors that characterise a person might even influence the their opinion on what phenomena are subsumed under the umbrella of sarcasm, that is, their definition of sarcasm. However, we did not restrict the reach of our survey to participants of one particular cultural or socio-demographic background. As such, it could be that tweets coming from different participants exhibit slightly different phenomena. Perhaps not all of them are sarcastic in some universal, cross-cultural sense. This issue could be alleviated if we published the socio-demographic information that we have collected about each participant. Sarcasm detection models could consider such information in the detection pipeline and make judgement that are particular to specific socio-demographic backgrounds. Unfortunately, while we are free to use this information internally, we are unable to publish it to the community. This is due to the restrictions imposed by the privacy policy that survey participants have agreed to. Future work could build further datasets with more relaxed privacy constraints. Of course, the privacy policy should be reviewed by an ethics committee and participants should be clearly informed about the public availability of their socio-demographic data.

Third, we argued that what constitutes an *accurate* label for a tweet is that label which reflects the sarcastic intention of the author of that tweet. However, apart from detecting author intention, it might also be important to consider the task of deter-

mining whether a particular observer of that tweet would consider the tweet sarcastic, irrespective of authorial intention. Authorial intention is important to detect for downstream tasks such as opinion mining and sentiment analysis. The sarcastic perception of a given observer is important to detect for downstream tasks such as discerning the offensive nature of that tweet to that observer. This requires future sarcasm detection models that are able to account for the cultural and socio-demographic factors that characterise the author, and the observer, respectively, and judge the tweet in that context. More effort is needed to understand what specific factors should be considered. The work that we presented in Chapter 6, summarised in Section 8.3 provides a starting point.

Limitations of the iSarcasmEval Task

First, the iSarcasmEval dataset that we provided for training and testing did not contain any further contextual information. Participants were expected to build models that judged the sarcastic nature of the texts in iSarcasmEval without having been provided with any contextual information. However, as we have discussed previously, for instance in Section 4.2.2, sarcasm is a pragmatic phenomenon. The sarcastic nature of a text could only be made apparent when interpreting it in light of contextual information external to the text. Such information could include the conversational context in which the text occurred, such as surrounding tweets, in the case of tweet datasets; and cultural and socio-demographic information about the author of that text. We suggest that future datasets should include such information. The insight from Chapter 6, summarised in Section 8.3, could aid the development of such datasets, in terms of what information about the author should be enclosed.

Second, the low performance achieved by the models on subtask B requires further investigation. More work is needed to understand the relevance of the chosen categorisation. Also, the ironic speech category labels should either be provided by the authors themselves, to avoid any bias introduced by annotators, or more emphasis should be placed on annotator training and annotation guideline clarity. This is to mitigate labelling noise that might indeed account, at least in part, for the low performance of the models submitted for subtask B.

Third, when we organise future competitions on sarcasm detection, we, as a community, might want rethink the metrics that we use to quantify success. It might be less desirable to encourage the submission of complex ensembles of models with little

introspection into their functionality and accountability, for the sole purpose of gaining a small performance margin on the leaderboard. Rather, as a community, we might want to reward deriving insight into the phenomenon of sarcasm, and its cultural and socio-demographic determinants.

Limitations of the Socio-demographic Investigation

First, our quantitative investigation involved emulating online sarcastic exchanges via third-party annotation. This had the advantage of providing us with granular control over the socio-demographic background of the annotator population. However, annotators might have behaved under experimental bias differently to how they would behave when participating in real online sarcastic exchanges. More work is needed to understand how such experimental biases could be minimised.

Second, we grouped all participants from a specific country, such as the United Kingdom, or the United States, into one socio-demographic category. More granular socio-demographic categorisations could be considered in future work. This is particularly the case for the United States, which spans a large geographical surface. Comparing different regions might exhibit cultural differences that could have implications on how sarcasm is understood.

Third, we have shown that socio-demographic similarity between human interlocutors could influence their ability to detect each other's sarcasm. However, it is unclear why socio-demographic similarity is influential. In Section 6.2 we suggested that one explanation could be that socio-demographic factors that characterise an interlocutor determine the set of social norms and expectations that they assume. Future work could investigate whether this suggestion holds. Such work could also enquire into the specific norms and expectations that a particular socio-demographic factor determines. This could be vital for informing how such socio-demographic factors should be encoded in computational models.

Limitations of Max and the Annotation Procedure

First, when building Max, our aim was to generate sarcastic responses to input texts that described actions. In this scenario, given such a text, we assumed that the expectation—a required component of the ironic environment—was the proposition that negated the action described in that text. See Section 7.3.2 for more details. More complex notions of expectation could be considered in the future. For instance, one might consider

commonsense expectations that are implied by the action described in the text.

Second, Max is faithful to the Implicit Display Theory of sarcasm. While this theory might have advantages over previous ones, discussed in Section 2.1.5, it might still have its own limitations that could have propagated into our sarcasm generation framework. More work is needed to understand these limitations. Being grounded in this theory, our investigation was also limited to the linguistic devices that the theory claims are associated with sarcasm, such as pragmatic insincerity. As such, the conclusions presented only hold with regards to the responses that Max is able to generate. Future work could investigate where the conclusions hold for other flavours of sarcasm. One option, that would remove the noise induced by synthetic sarcasm generators, is the following. One could ask humans to produce sarcastic responses, and present these to third party annotators, asking the same questions that we have asked in this chapter, while only informing the annotators that the responses were generated by humans once the annotators have completed the survey.

Third, we asked human annotators to judge the sarcastic nature of the responses generated by Max. However, recall from Chapter 6 that the level of perceived sarcasticness of an utterance can vary with the socio-demographic background of the annotator. However, we did not compare how annotators of different background would differ in their perceived levels of sarcasticness. As we argue in Section 7.3.3, population-level conclusions are still informative. However, future work could be conducted to understand if preferences regarding sarcastic responses from chatbots vary across annotators of different backgrounds.

8.6 Final Thoughts

Dear reader, thank you very much for taking the time to browse through this thesis. Whether you read all of it, or a small fragment of it, I hope that it was a good use of your time and you found something interesting to think about. I'll leave you with this final thought, which I hope to be considered by future computational investigations of sarcasm.

Sarcasm is a computationally underexplored category of linguistic phenomena that is difficult to encompass into a single explicative theory. These are pragmatic phenomena. Detecting sarcasm in text might require contextual information external to that text. This includes information about the cultural and socio-demographic background of the author of that text. Such information could not only influence how sarcastic texts

are formulated, but the author's very definition of what sarcasm is. Therefore, these are tasks that should be approached from a computational social science perspective. Sarcasm detection is not intrinsically a natural language processing task. Using larger, general-purpose models, with internals difficult to interpret, can indeed find practical applications. But such models do not necessarily advance our understanding of sarcasm, nor scientific pursuit as such. Similarly, sarcasm generation is also a computational social science task. We should not build general-purpose sarcastic chatbots, but chatbots that account for the cultural and socio-demographic characteristics of their human interlocutor. Such characteristics could determine when humans prefer sarcastic responses, and what flavour of sarcasm they best understand.

Appendix A

Patterns Used by Max to Generate Sarcastic Responses

Here we show the patterns used by Max to convert the logical form of sarcasm to text, as discussed in Section 7.3.2. We show patterns for each if-then relation type, *xNeed*, *xAttr*, *xReact*, and *xEffect*.

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

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

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

A.1 Patterns for the Complete Version of Max

xNeed patterns:

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

xAttr patterns:

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

xReact patterns:

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

xEffect patterns:

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

A.2 Patterns for Max without Pragmatic Insincerity

xNeed patterns:

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

xAttr patterns:

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

xReact patterns:

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

xEffect patterns:

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

A.3 Patterns for Max without Emotional Markers

xNeed patterns:

- You didn't <obt>.

xAttr patterns:

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

xReact patterns:

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

xEffect patterns:

- You're not going to obt_inf right now.

Appendix B

Sarcastic Responses Generated by SarcasmBot

Here we show the 28 unique responses produced by SarcasmBot to our set of 250 inputs, as discussed in Section 7.4.2.

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

Bibliography

Abercrombie, G. and Hovy, D. (2016). Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of Twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany. Association for Computational Linguistics.

Abu Farha, I., Oprea, S. V., Wilson, S., and Magdy, W. (2022a). iSarcasmEval, intended sarcasm detection in english and arabic. In *Proceedings of The 16th International Workshop on Semantic Evaluation*, Seattle, Washington. Association for Computational Linguistics.

Abu Farha, I., Wilson, S., Oprea, S., and Magdy, W. (2022b). Sarcasm detection is way too easy! an empirical comparison of human and machine sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5284–5295, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amante, D. J. (1981). The theory of ironic speech acts. *Poetics Today*, 2(2):77–96.

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

Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., and Silva, M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Aroyehun, S., Angel, J., and Gelbukh, A. (2022). Tug-cic at semeval-2021 task 6: Two-stage fine-tuning for intended sarcasm detection. In *Proceedings of the 16th*

International Workshop on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics.

Austin, J. L. (1962). *How to do things with words*, volume 88. Oxford University Press, Oxford, UK.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Bamman, D. and Smith, N. (2015). Contextualized sarcasm detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Barbieri, F., Ronzano, F., and Saggion, H. (2014a). Italian irony detection in twitter: a first approach. In *CLiC-it*, page 28. AILC.

Barbieri, F., Saggion, H., and Ronzano, F. (2014b). Modelling sarcasm in Twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, Baltimore, Maryland. Association for Computational Linguistics.

Benamara, F., Grouin, C., Karoui, J., Moriceau, V., and Robba, I. (2017). Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du défi fouille de textes deft2017. *Atelier TALN 2017: Défi Fouille de Textes*.

Bharti, S. K., Babu, K. S., and Jena, S. K. (2015). Parsing-based sarcasm sentiment recognition in twitter data. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1373–1380.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Bouazizi, M. and Ohtsuki, T. (2015). Opinion mining in twitter: How to make use of sarcasm to enhance sentiment analysis. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1594–1597.

Bouton, L. F. (1988). A cross-cultural study of ability to interpret implicatures in English. *World Englishes*, 7(2):183–196.

Bouton, L. F. (1992). The interpretation of implicature in english by nns: Does it come automatically–without being explicitly taught? *Pragmatics and language learning*, 3:53–65.

Bravo-Marquez, F., Frank, E., Mohammad, S. M., and Pfahringer, B. (2016). Determining word-emotion associations from tweets by multi-label classification. In *WI*, pages 536–539. IEEE.

Bravo-Marquez, F., Mendoza, M., and Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69:86 – 99.

Bruntsch, R. and Ruch, W. (2017). Studying irony detection beyond ironic criticism: Let's include ironic praise. *Frontiers in Psychology*, 8. Original Research.

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

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102.

Bueno, R. O., Pardo, F. M. R., Farías, D. I. H., Rosso, P., y Gómez, M. M., and Medina-Pagola, J. (2019). Overview of the task on irony detection in spanish variants. In *IberLEF@SEPLN*.

Campbell, J. D. and Katz, A. N. (2012). Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480.

Chakrabarty, T., Ghosh, D., Muresan, S., and Peng, N. (2020). Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Chamberlain, B. P., Humby, C., and Deisenroth, M. P. (2017). Probabilistic inference of twitter users' age based on what they follow. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 191–203, Skopje, Macedonia. Springer International Publishing.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *CSCW*, pages 1217—1230, Portland, OR, USA. Association for Computing Machinery.

Chung, H. W., Févry, T., Tsai, H., Johnson, M., and Ruder, S. (2020). Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821.

Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P., et al. (2018). Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohn-Gordon, R. and Bergen, L. (2019). Verbal irony, pretense, and the common ground.

Colston, H. L. (2000). On necessary conditions for verbal irony comprehension. *Pragmatics & Cognition*, 8(2):277–324.

Colston, H. L. and Gibbs Jr., R. W. (2007). *A brief history of irony.*, pages 3–21. Irony in language and thought: A cognitive science reader. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

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

Colston, H. L. and O'Brien, J. (2000b). Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole. *Discourse Processes*, 30(2):179–199.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Craker, N. and March, E. (2016). The dark side of facebook®: The dark tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, 102:79–84.

Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2021). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dews, S. and Winner, E. (1995). Muting the meaning a social function of irony. *Metaphor and Symbolic Activity*, 10(1):3–19.

Dress, M. L., Kreuz, R. J., Link, K. E., and Caucci, G. M. (2008). Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

Du, X., Hu, D., ZHI, J. M., Jiang, L., and Shi, X. (2022). Pali-nlp at semeval-2022 task 6: isarcasmeval- fine-tuning the pre-trained model for detecting intended sarcasm. In

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics.

El Mahdaouy, A., EL MEKKI, A., Essefar, K., Skiredj, A., and Berrada, I. (2022). Cs-um6p at semeval-2022 task 6: Transformer-based models for intended sarcasm detection in english and arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol*, 15(1-2):5–27.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

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

Haiman, J. (1989). Alienation in grammar. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 13(1):129–170.

Han, Y., Chai, Y., Wang, S., Sun, Y., Huang, H., Chen, G., Xu, Y., and Yang, Y. (2022). X-pudu at semeval-2022 task 6: Multilingual learning for english and arabic sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Harris, M., Ivanko, S., Jungen, S., Hala, S., and Pexman, P. (2001). You're really nice: Children's understanding of sarcasm and personality traits. Poster presented at the second biennial meeting of the Cognitive Development Society, Virginia Beach, VA, USA.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. (2018). CASCADE: Contextual sarcasm detection in online discussion forums.

In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Heyman, G. D. and Gelman, S. A. (1999). The use of trait labels in making psychological inferences. *Child Development*, 70(3):604–619.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

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

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

Joshi, A., Bhattacharyya, P., Carman, M., Saraswati, J., and Shukla, R. (2016a). How do cultural differences impact the quality of sarcasm annotation?: A case study of Indian annotators and American text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99, Berlin, Germany. Association for Computational Linguistics.

Joshi, A., Kunchukuttan, A., Carman, M. J., and Bhattacharyya, P. (2015a). Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *Workshop on Issues of Sentiment Discovery and Opinion Mining at the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery.

Joshi, A., Sharma, V., and Bhattacharyya, P. (2015b). 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)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., and Carman, M. (2016b). Are word embedding-based features useful for sarcasm detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011, Austin, Texas. Association for Computational Linguistics.

Kellogg, R. T. (2001). Long-term working memory in text production. *Memory & Cognition*, 29(1):43–52.

Khodak, M., Saunshi, N., and Vodrahalli, K. (2018). A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kingma, D. P. and Ba, J. (2015). Adam: a method for stochastic optimization. *ICLR*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Kreuz, R. and Caucci, G. (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

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

Kreuz, R. J., Long, D. L., and Church, M. B. (1991). On being ironic: Pragmatic and mnemonic implications. *Metaphor and Symbolic Activity*, 6(3):149–162.

Kumon-Nakamura, S., Glucksberg, S., and Brown, M. (1995). How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*, 124(1):3–21.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Leggitt, J. S. and Gibbs, R. W. (2000a). Emotional reactions to verbal irony. *Discourse Processes*, 29(1):1–24.

Leggitt, J. S. and Gibbs, R. W. (2000b). Emotional reactions to verbal irony. *Discourse Processes*, 29(1):1–24.

Li, P., Lu, H., Kanhabua, N., Zhao, S., and Pan, G. (2018). Location inference for non-geotagged tweets in user timelines. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1150–1165.

Liebrecht, C., Kunneman, F., and van den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

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

Matthews, G. and Gilliland, K. (1999). The personality theories of h. j. eysenck and j. a. gray: A comparative review. *Personality and Individual Differences*, 26(4):583–626.

Maynard, D. and Greenwood, M. (2014). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*,

pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., and Magdy, W. (2021). SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Mishra, A., Tater, T., and Sankaranarayanan, K. (2019). A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6144–6154, Hong Kong, China. Association for Computational Linguistics.

Mohammad, S. and Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Moores, B. and Mago, V. (2022). A survey on automated sarcasm detection on twitter. *CoRR*, abs/2202.02516.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses*. Wiley, New York, NY, USA.

Oprea, S. V. and 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*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Oprea, S. V. and Magdy, W. (2020a). The effect of sociocultural variables on sarcasm communication online. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Oprea, S. V. and Magdy, W. (2020b). iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

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

Oprea, S. V., Wilson, S., and Magdy, W. (2022). Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Ouyang, X., Wang, S., Pang, C., Sun, Y., Tian, H., Wu, H., and Wang, H. (2021). ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pexman, P. M. (2005). Social Factors in the Interpretation of Verbal Irony: The Roles of Speaker and Listener Characteristics.

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

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

Phillips, L. H., Allen, R., Bull, R., Hering, A., Kliegel, M., and Channon, S. (2015). Older adults have difficulty in decoding sarcasm. *Developmental psychology*, 51(12):1840—1852.

Ptáček, T., Habernal, I., and Hong, J. (2014). Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference*

on Computational Linguistics: Technical Papers, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Rajadesingan, A., Zafarani, R., and 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*, WSDM '15, page 97–106, New York, NY, USA. Association for Computing Machinery.

Reyes, A. and Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4):754–760. 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on PACIS 2010.

Reyes, A., Rosso, P., and Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng.*, 74:1–12.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and 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*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Rockwell, P. and Theriot, E. M. (2001). Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18(1):44–52.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. (2019). Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Sayyadiharikandeh, M., Ciampaglia, G. L., and Flammini, A. (2016). Cross-domain gender detection in twitter. In *Proceedings of the Workshop on Computational Approaches to Social Modeling*, Bellevue, WA, USA. SocInfo.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Searle, J. R. (1975). Indirect speech acts. In *Speech acts*, pages 59–82. Brill, Leiden, Netherlands.

Searle, J. R. and Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, Cambridge, UK.

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., and Zhang, C. (2018). Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*. AAAI Press.

Shmueli, B., Ku, L.-W., and Ray, S. (2020). Reactive Supervision: A New Method for Collecting Sarcasm Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2553–2559, Online. Association for Computational Linguistics.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Sperber, D. and Wilson, D. (1981). Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Tay, Y., Luu, A. T., Hui, S. C., and Su, J. (2018). Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Taylor, C. (2016). Women are bitchy but men are sarcastic? investigating gender and sarcasm. *Gender and Language*, 11(3).

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning. Accessed: 2019-03-01.

Tsur, O., Davidov, D., and Rappoport, A. (2010). Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):162–169.

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

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

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

van der Goot, R. (2022). MaChAmp at SemEval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task multi-lingual learning for a pre-selected set of semantic datasets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1695–1703, Seattle, United States. Association for Computational Linguistics.

van der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., and Plank, B. (2021). Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Van Hee, C., Lefever, E., and Hoste, V. (2018). SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Veale, T. and Hao, Y. (2007). Making lexical ontologies functional and context-sensitive. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 57–64, Prague, Czech Republic. Association for Computational Linguistics.

Veale, T. and Hao, Y. (2010). Detecting ironic intent in creative comparisons. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 765–770, NLD. IOS Press.

Veale, T., Hao, Y., and Li, G. (2008). Multilingual harvesting of cross-cultural stereotypes. In *Proceedings of ACL-08: HLT*, pages 523–531, Columbus, Ohio. Association for Computational Linguistics.

Wallace, B. C., Choe, D. K., and Charniak, E. (2015). Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Wilson, D. (2006). The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.

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

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., and Huang, Y. (2018). THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56, New Orleans, Louisiana. Association for Computational Linguistics.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *COLING*, pages 947–953, Saarbrücken, Germany. ACL.

Yuan, M., Mengyuan, Z., Jiang, L., Mo, Y., and Shi, X. (2022). stce at semeval-2022 task 6: Sarcasm detection in english tweets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zheng, G., Wang, J., and Zhang, X. (2022). Ynu-hpcc at semeval-2022 task 6: Transformer-based model for intended sarcasm detection in english and arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.