iSarcasm: A Dataset of Intended Sarcasm

Silviu Vlad Oprea
University of Edinburgh
silviu.oprea@ed.ac.uk

Walid Magdy
University of Edinburgh
wmagdy@inf.ed.ac.uk

Abstract

This paper considers the distinction between intended and perceived sarcasm in the context of textual sarcasm detection. The former occurs when an utterance is sarcastic from the perspective of its author, while the latter occurs when the utterance is interpreted as sarcastic by the audience. We show the limitations of previous labelling methods in capturing intended sarcasm and introduce the iSarcasm dataset of tweets labeled for sarcasm directly by their authors. We experiment with sarcasm detection models on our dataset. The low performance indicates that sarcasm might be a phenomenon under-studied computationally thus far.

1 Introduction

Sarcasm is a form of irony that occurs when there is some discrepancy between the literal and intended meanings of an utterance. This discrepancy is used to express dissociation towards a previous proposition, often in the form of contempt or derogation (Wilson, 2006).

Sarcasm is omnipresent in social media text and can be highly disruptive of systems that harness this data for sentiment and emotion analysis (Maynard and Greenwood, 2014). It is therefore imperative to devise models for textual sarcasm detection. The effectiveness of these models depends on the availability and quality of labelled data used for training. Collecting such data is challenging due to the subjective nature of sarcasm. Dress et al. (2008) notice a lack of consistence in how sarcasm is defined by people of different socio-cultural backgrounds. As a result, an utterance that is *intended* as sarcastic by its author might not be *perceived* as such by audiences of different backgrounds (Rockwell and Theriot, 2001).

There are two methods used so far to label texts for sarcasm: distant supervision, where texts are considered sarcastic if they meet predefined criteria, such as including specific hashtags; and manual labelling by human annotators. We believe both methods are suboptimal for capturing the sarcastic intention of the authors of the texts. As such, existing models might be optimized to capture the noise induced by these labelling methods.

In this paper, we present the iSarcasm dataset of tweets labelled for sarcasm by their authors. To our knowledge, this is the first attempt to create noise-free examples of intended sarcasm. In a survey, we asked Twitter users to provide us with both sarcastic and non-sarcastic tweets that they have posted in the past. Tweet labels are thus implicitly specified by the authors themselves. We implement a restrictive quality control algorithm to avoid noisy input, using both survey responses and metadata from their Twitter profiles.

We investigate how third-party annotators perceive the sarcastic intention of the authors reflected in iSarcasm. Third-party annotation for sarcasm has been conducted before (Filatova, 2012; Riloff et al., 2013; Abercrombie and Hovy, 2016), but no studies checked the ability of the annotators to capture the actual sarcasm meant by the authors. We collect the third-party labels from workers on crowdsourcing platforms. These labels capture the intention of the authors with an F-score of 0.616, indicating that sarcasm is a subjective phenomenon that is challenging even for humans to detect. This suggests future research into what influences sarcasm perception.

We also implement sarcasm detection models suggested previously (Tay et al., 2018; Hazarika et al., 2018; Van Hee et al., 2018) and test them on our dataset. While these models achieve F-scores reaching 0.87 on existing datasets, they yield a maximum F-score of 0.356

on iSarcasm, suggesting that previous datasets might be biased or obvious. iSarcasm seems to reflect a category of sarcasm less explored computationally thus far. This highlights the importance of developing new, more effective approaches, for sarcasm detection.

iSarcasm contains 4,484 English tweets, each with an associated intended sarcasm label provided by its author, with a ratio of roughly 1:5 of sarcastic to non-sarcastic labels. We publish the dataset publicly for research purposes[1].

2 Background

2.1 Intended and Perceived Sarcasm

Sarcasm is a form of irony marked by a discrepancy between the literal and intended meanings of an utterance, through which the speaker usually manifests a hostile, derogatory, or contemptuous attitude (Wilson, 2006). There are many reasons why a speaker might use sarcasm. It can provide a way of feeling safer when conveying a negative message (Norrick, 1994) or when expressing anger (Ducharme, 1994). It can also showcase dominance and control (Dews et al., 1995).

The way sarcasm is defined varies greatly across socio-cultural backgrounds. Dress et al. (2008) notice that members of collectivist cultures tend to express sarcasm in a more subtle way than individualists. They also point out gender differences. Those who identify as females seem to have a more self-deprecating attitude when using sarcasm. Rockwell and Theriot (2001) find some cultures to associate sarcasm with humour more than others. There also seem to be cultures who do not use sarcasm at all, such as the Hua, a group of New Guinea Highlanders (Attardo, 2002). Because of these differences, an utterance that is intended sarcastic by its author might not be perceived as such by the audience (Jorgensen et al., 1984). The converse could also be true. The audience could perceive the utterance as sarcastic, even if it was not intended as such by the author.

The distinction between intended and perceived sarcasm, also referred to as encoded and decoded sarcasm, respectively, has been pointed out in previous research (Kaufer, 1981; Rockwell and Theriot, 2001). However, it has not been considered in a computational context thus far when building datasets for textual sarcasm detection. We believe accounting for it is essential, especially nowadays. Consider social media posts that can reach audiences of unprecedented sizes. It is important to consider both the communicative intention of the author, as well as possible interpretations by audiences of different socio-cultural backgrounds.

2.2 Sarcasm Datasets

Two methods were used so far to label texts for sarcasm: distant supervision and manual labelling.

Distant supervision This is by far the most common labelling method for sarcasm in text. Texts are considered positive examples (sarcastic) if they meet predefined criteria. The criteria is typically containing specific tags, such as #sarcasm for Twitter data (Ptáček et al., 2014), and /s for Reddit data (Khodak et al., 2018), or being generated by specific accounts (Barbieri et al., 2014a). Negative examples are usually randomly selected posts that do not match the criteria. Table 1 gives an overview of datasets constructed this way, with the tags or accounts they consider sarcastic.

The main advantage of distant supervision is that it allows the collection of large datasets labeled automatically with no manual effort. However, the labels it produces can be highly noisy. For instance, it considers all tweets that lack predefined tags as non-sarcastic. We discuss limitations in details in Section 3.

Manual labelling An alternative to distant supervision is collecting texts and presenting them to human annotators. 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 a set of 2,240 Twitter conversations, ending up with 448 positive and 1,732 negative labels, respectively. Riloff et al. (2013) use a hybrid approach for labeling, where they collect a set of 1,600 tweets that contain #sarcasm or #sarcastic, and another 1,600 without these tags. They remove the tags from all tweets and present them to a group of human annotators for final labelling. We call this the *Riloff dataset*.

The main limitation of manual labelling is the absence of evidence on the intention of the author of the texts that are being labelled. Annotator perception may be different to author intention, con-

[1] Available at https://github.com/silviu-oprea/isarcasm

Sarcasm labeling method	Source	Details / Tags / Accounts
Distant supervision		
Davidov et al. (2010)	Twitter	#sarcasm, #sarcastic, #not
Barbieri et al. (2014b)	Twitter	#sarcasm, #education, #humor, #irony, #politics
Ptáček et al. (2014)	Twitter	#sarcasm, #sarcastic, #irony, #satire
Bamman and Smith (2015a); Joshi et al. (2015)	Twitter	#sarcasm, #sarcastic
González-Ibáñez et al. (2011); Reyes and Rosso (2012); Liebrecht et al. (2013); Bouazizi and Ohtsuki (2015); Bharti et al. (2015)	Twitter	#sarcasm
Barbieri et al. (2014a)	Twitter	tweets posted by *@spinozait* or *@LiveSpinoza*
Khodak et al. (2018)	Reddit	/s
Manual annotation / Hybrid		
Riloff et al. (2013); Van Hee et al. (2018)	Twitter	tweets
Abercrombie and Hovy (2016)	Twitter	tweet-reply pairs
Filatova (2012)	Amazon	product reviews

Table 1: Datasets previously suggested for sarcasm detection, all annotated using either distant supervision or manual labelling, as discussed in Section 2.2.

sidering the subjective nature of sarcasm. Hybrid methods can share limitations of both distant supervision and manual labelling.

2.3 Sarcasm Detection Models

Based on the information considered when classifying a text as sarcastic or non-sarcastic, we identify two classes of models across literature: text-based models and contextual models.

Text-based models These models only consider information available within the text being classified. Most work in this direction considers linguistic incongruity (Campbell and Katz, 2012) to be a marker of sarcasm. In this direction, Riloff et al. (2013) look for a positive verb in a negative sentiment context. Bharti et al. (2015) search for a negative phrase in a positive sentence. (Hernández Farías et al., 2015) measure semantic relatedness between words using Wordnet-based similarity. Joshi et al. (2016b) use the cosine similarity between word embeddings. Recent work captures incongruity using a neural network with an intra-attention mechanism (Tay et al., 2018).

Contextual models These models utilize information both from the text and from the context of its disclosure, such as author information. Using Twitter data, Bamman and Smith (2015a) represent author context as manually-curated features extracted from their historical tweets. Amir et al. (2016) merge all historical tweets into one document and use the Paragraph Vector model (Le and Mikolov, 2014) to build a representation of that document. Building on this, Hazarika et al. (2018) extract additional personality features from that document using a model pre-trained on a personality detection benchmark corpus.

Despite reporting encouraging results, all discussed models are trained and tested on datasets annotated via manual labelling, distant supervision, or a mix between them. We believe both labelling methods are limited in their ability to capture intended sarcasm without noise. In the following section, we discuss how noise can occur.

3 Limitations of Current Labelling Methods

In this section, we discuss limitations of current labelling methods that make them suboptimal for capturing intended sarcasm. We demonstrate some of these empirically on the Riloff dataset (Riloff et al., 2013), which uses a hybrid approach for labelling.

3.1 Limitations of Distant Supervision

Since it is based on signals provided by the authors, distant supervision might seem like a candidate for capturing intended sarcasm. However, we identify a few fundamental limitations with it. First, the tags may not mark sarcasm, but may constitute the subject or object of conversation, e.g. *#sarcasm annoys me!*. This could lead to false positives in the training data. Second, when using tags such as #politics and #education (Barbieri et al., 2014b), there is a strong underlying assumption that these tags are accompanied by sarcasm, potentially generating further

	with tag	without tag
annotated sarcastic	345	26
annotated non-sarcastic	486	975

Table 2: The agreement between manual annotation and the presence of sarcasm tags in the Riloff dataset

false positives. The assumption that some accounts always generate sarcasm (Barbieri et al., 2014a) is similarly problematic. In addition, the intended sarcasm that they do capture might be of a specific flavor, either when tweeted by a specific account or when carrying a given tag. Building a model trained on this dataset might, therefore, be biased to a flavour of sarcasm, being unable to capture other flavours, increasing the risk of false negatives. Finally, if a text does not contain the predefined tags, it is considered non-sarcastic. This is a strong and problematic assumption that can lead to false negatives in the training data. Indeed, no tweet in iSarcasm, sarcastic or non-sarcastic, includes any of the predefined tags traditionally associated with sarcasm.

3.2 Limitations of Manual labelling

In manual labelling texts are collected and presented to human annotators (Filatova, 2012; Riloff et al., 2013; Abercrombie and Hovy, 2016; Van Hee et al., 2018). This is problematic in terms of capturing intended sarcasm in light of studies that point out how sarcasm perception varies across socio-cultural contexts (Rockwell and Theriot, 2001; Dress et al., 2008).

Joshi et al. (2016a) provide more insight into this problem on the Riloff dataset. They present the dataset, initially labelled by Americans, to be labelled by Indians 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 socio-cultural differences between India and the United States. They conclude that sarcasm annotation expands beyond linguistic expertise and is dependent on considering such factors.

Labels provided by third-party annotators might therefore not reflect the sarcastic intention of the authors of the texts that are being labelled, making this labelling method suboptimal for capturing intended sarcasm. To investigate this further, we looked at the Riloff dataset discussed in Section 2.2 which uses a hybrid approach for labelling. The dataset is published as a list of labelled tweet IDs We could only retrieve 1,832 tweets, the others being removed from Twitter. We looked at the agreement between the presence of tags and manual annotation. Table 2 shows the results. We notice that 58% of the tweets with tags were labeled non-sarcastic. This disagreement between distant supervision and manual annotation provides further evidence to doubt the ability of the latter to capture intended sarcasm, at least not the flavor that distant supervision might capture.

As we saw, both labelling methods use a proxy for detecting intended sarcasm, in the form of predefined tags, predefined sources, or third-party annotators. As such, they may create noisy labels, in terms of both false positives and false negatives. Our objective is to create a noise-free dataset labelled for intended sarcasm. To accomplish this we collect labels from the authors themselves.

4 Data Collection

In the following we describe the process of collecting our iSarcasm dataset and the third-party labels.

4.1 Collecting Sarcastic Tweets

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, on their timeline, or as replies to other tweets. We made it clear that the tweets had to be their own and no retweets were allowed. We further required that the 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, users had to recall and explain, in full English sentences, why it was sarcastic and what they would say to 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 implemented the following quality control steps:

- The provided links should be for 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. The initial number was set to 5K, but some workers asked us to raise it since they had more followers.
- Tweets should not consist of only hashtags or URLs;
- Links to tweets should not have been submitted in a previous response;
- Responses submitted in less than three minutes are discarded.

Our survey shows an initial consent form to be accepted before allowing any contributor to provide responses. It contains instructions and information about how the data will be handled and published. We informed contributors that only tweet IDs will be made public, to allow them to take down their tweet anytime they want as a control to their privacy. They have agreed that we may publish the tweet IDs and the labels as part of open science, and that we may collect public information from their profile.

We published our survey on multiple crowdsourcing platforms, including Figure-Eight (F8), Amazon Mechanical Turk (AMT) and Prolific Academic (PA)[2]. We could not get any quality responses from F8. In fact, most of our quality control steps were developed over multiple iterations on F8. On AMT, we retrieved some high quality responses, but, unfortunately, AMT stopped our job, considering that getting links to personal tweets of participants violates their policy. We collected the majority of responses on PA.

4.2 Labelling Sarcasm Categories

In the next stage we asked a human trained in linguistics to further label each sarcastic tweet as belonging to one of the following categories of *ironic speech*:

1. *sarcasm*: tweets that contradict a knowable state of affairs and are critical towards an addressee;
2. *irony*: tweets that contradict a knowable 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 that do not exhibit ironic speech and constitute noise.

This categorisation (excluding the *invalid* category) is similar to the one presented by **?**.

4.3 Collecting Third-Party Labels

In this part, we decided to replicate the manual annotation approach presented in previous research (Riloff et al., 2013; Abercrombie and Hovy, 2016; Van Hee et al., 2018) on our dataset and compare the resulting *perceived sarcasm* labels to the *intended sarcasm* labels collected from the authors of the tweets. Our aim is to estimate the human performance in detecting sarcasm as intended by the authors.

When collecting perceived sarcasm labels, we aimed to reduce noise caused by variations in how sarcasm is defined across socio-cultural backgrounds. Previous studies have shown gender (Dress et al., 2008) and country (Joshi et al., 2016a) to be the variables that are most influential on this definition. Based on their work, we made sure all annotators shared the same values for these variables. We used PA as the platform for publishing a third-party labelling survey, as it allows the most granular control over the target worker population. Following the work of Riloff et al. (2013) on the Riloff dataset, we collected three separate labels for each tweet and considered the dominant one.

5 Data Statistics and Analysis

5.1 iSarcasm Dataset

We received 1,236 responses to our survey. About 84% of the contributors were from the UK and the

[2]AMT: www.mturk.com, PA: prolific.ac, F8: www.figure-eight.com

US (52% from the UK and 32% from the US), and the others from countries such as Canada and Australia. 48% of them were females, and over 68% were less than 35 years old.

Each response contained four tweets labelled for sarcasm by their author. In total we got 1,236 sarcastic and 3,708 non-sarcastic tweets. We applied the quality control steps described in Section 4.1 and disregarded all tweets that fall under the *invalid* category. 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's sarcastic, as well as how they would rephrase it to be non-sarcastic. The average length of a tweet is around 20 words, of explanations 22 words, and of rephrases 16 words. Table 3 shows the distribution over the categories. Table 4 shows sarcastic examples, one for each category, along with the explanations and rephrases.

We checked the presence of the tags #sarcasm, #sarcastic, #irony and #not in iSarcasm, tags commonly used to mark tweets as sarcastic in distant supervision. None of our tweets contains any of those tags, which confirms one of our discussed limitations of this approach, that the lack of tags should not be associated with lack of sarcasm.

We publish iSarcasm as two files, a training set and a test set, containing 80% and 20% of the examples chosen at random, respectively. Each file contains tweet IDs along with corresponding intended sarcasm labels. For sarcastic tweets we also specify the category of ironic speech they belongs. This is in accordance with the consent form that the contributors have agreed to, whose privacy we take very seriously. At the same time, we are aware of the fact that corresponding tweets might become unavailable at any time. Further, the explanations, rephrases, and user demographic information, might prove invaluable for future modelling and analysis purposes. As such, we are happy to provide all these to researchers who contact us, under an agreement to protect the privacy of the contributors according to the consent form.

5.2 Third-Party Labels

We collected three third-party labels for the test set in iSarcasm. We computed inter-annotator agreement (IAA) among third-party annotators using Cohen's kappa (κ; Cohen (1960)). The pairwise IAA scores were $\kappa_{12} = 0.37$, $\kappa_{13} = 0.39$ and $\kappa_{23} = 0.36$. This only indicates a *fair* agreement among annotators, according to Cohen's categorisation. We used majority voting to select the final perceived sarcasm label for each tweet. Table 5 shows the agreement between the intended and perceived labels. We notice that 26 tweets intended as sarcastic were not perceived as such. Similarly, 50 tweets that were non intended as sarcastic were perceived sarcastic.

There are several potential causes for the inter-annotator disagreement, as well as the disagreement between intended and perceived labels. It could be that the annotators lacked contextual information important for ameliorating their uncertainty. On the other hand, perhaps even if given such information, there are further demographic variables that influence sarcasm perception that should be controlled, besides gender and country which we controlled. These variables might influence the way any information is processed, including contextual information. While this hypothesis is backed by research in psycholinguistics (Pexman, 2005), further computational investigation is necessary that is outside the scope of this paper. We hope that our results here will motivate this work. Nevertheless, these results increase our belief that relying on third party labels can be suboptimal in terms of capturing intended sarcasm.

6 Detecting Intended Sarcasm

In the following, we train models to detect sarcasm on our dataset. Our aim is to investigate the ability of these models to detect intended sarcasm rather than sarcasm labeled using distant supervision or manual annotation, which we believe to produce noisy labels. We first consider the datasets which have been used most commonly in sarcasm detection, up to our knowledge. For each dataset, we implement the model that are reported to achieve good results that dataset. We then test each model on iSarcasm and compare the results.

6.1 Baseline Datasets

The datasets we consider are Riloff (Riloff et al., 2013) and Ptacek (Ptáček et al., 2014). Riloff consists of 3,200 tweet IDs labelled manually for sarcasm. Out of these, we manage to collect 1,832 tweets, the rest have been removed from Twitter. Similarly, for the Ptacket dataset labelled via distant supervision, we collect 27,177 tweets out of

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

Table 3: Distribution of sarcastic tweets into the categories introduced in Section 4.2.

category	**tweet text**	**explanation**	**rephrased**
sarcasm	Thank @user for being so entertaining at the Edinburgh signings! You did not disappoint! I made my flight so will have plenty time to read @user	I went to a book signing and the author berated me for saying I was lying about heading to Singapore straight after the signing	I would have said 'here is the proof of my travel, I am mad you embarassed me in front of a large audience'!
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	@mizzieashitey @PCDPhotography Totally didnt happen, its a big conspiracy, video can be faked....after all, theyve 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.
rhet. 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: Examples of sarcastic tweets from our datasets, one for each category discussed in Section 4.2. We also the explanations that authors gave as to what made their tweets sarcastic (explanation) and how they rephrased them to be non-sarcastic (rephrased).

	perc. sarc.	perc. non-sarc.
int. sarc.	61	26
int. non-sarc.	50	322

Table 5: The agreement between intended labels (*int.*), provided by the authors, and perceived labels, provided by third-party annotators, (*perc.*) on the test set of iSarcasm, as discussed in Section 5.2.

the 50K published tweet IDs.

6.2 Sarcasm Detection Models

We replicate some of the models reported by (Tay et al., 2018) to achieve good results on Riloff and Ptacek. These are: **LSTM** first encodes the tweet with a recurrent neural network (RNN; Elman (1990)) with long-term short memory units (LSTM; Hochreiter and Schmidhuber (1997)), then adds a binary softmax layer to output a probability distribution over labels and assigns the most probable label. **Att-LSTM** builds upon the LSTM model with a neural attention mechanism in the setting specified by Yang et al. (2016). **CNN** encodes the tweet with a convolutional neural network (CNN) and provides the result to feed-forward network with a final binary softmax layer, choosing the most probable label. We represent each tweet as a sequence of word vectors initialized using GloVe embeddings (Pennington et al., 2014) of dimension 100. We use an LSTM with 100 units and a CNN with max-pooling with 100 filters of size 3. We set the dimension of all hidden layers of SIARN and MIARN to 100. To check that we have a similar setting to Tay et al. (2018), we compare our implementations with theirs on Riloff and Ptacek and report similar performance, using the same data splits as they do, as reported in Table 6.

6.3 Results and Analysis

Table 7 reports precision, recall and f-score results on iSarcasm.

CNN is the best performing model with an F-score of 0.356, however all models achieve a low performance. Human performance significantly higher, with an F-score of only 0.616. All models

Dataset	Model	published	our impl.
Riloff	LSTM	0.673	0.669
	Att-LSTM	0.687	0.679
	CNN	0.686	0.681
Ptacek	LSTM	0.837	0.837
	Att-LSTM	0.837	0.841
	CNN	0.804	0.810

Table 6: F-score yielded by our implementations on the Riloff and Ptacek, compared to published results.

Model	Precision	Recall	F-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

Table 7: Experimental results on iSarcasm. *Manual Labelling* shows the results using the perceived sarcasm labels provided by third-party human annotators.

achieve a much higher recall than precision, predicting a large number of false positives. It is the higher precision that sets humans apart.

Let us look at sarcastic tweets that Att-LSTM classifies correctly, but the human annotators do not. We noticed the attention weights were high 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 on the background of the person reading, it might be perceived as sarcastic (e.g. by a sedentary person) or non-sarcastic (e.g. by an athlete).

These results underline the complexity of the intended sarcasm expressed in our dataset and motivate the need to develop more effective methods for detecting it. These methods might account for socio-cultural traits of the authors (available on, or inferred from, their Twitter profiles), or might look at what contextual elements are needed to judge the sarcasm in our dataset. Previous research has considered certain contextual elements (Bamman and Smith, 2015b; Rajadesingan et al., 2015; Amir et al., 2016; Hazarika et al., 2018), but only on sarcasm captured by previous labelling methods.

7 Conclusion

The difference between intended and perceived sarcasm has not been considered when labelling texts for sarcasm or building detection models. In this paper, we presented iSarcasm, the first dataset that contains sarcastic and non-sarcastic tweets as intended by their authors. We believe this dataset will allow future work in sarcasm detection to progress in a setting free of the noise found in existing datasets. We saw that computational models perform poorly in detecting sarcasm in the new dataset, indicating that the sarcasm detection task might be more challenging compared to how it was seen in earlier research. We aim to promote research in sarcasm detection, and to encourage future investigations into sarcasm in general and how it is perceived across cultures. In the future we also plan to collect further third-party labels from annotators of different backgrounds. This could provide more insight into how sarcasm is perceived across sociocultural contexts, which could yield better prediction models.

References

Gavin Abercrombie and Dirk Hovy. 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. ACL.

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

Salvatore Attardo. 2002. Talk is cheap: Sarcasm, alienation, and the evolution of language. *Journal of Pragmatics*, 34.

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

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

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

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014b. Modelling sarcasm in twitter, a novel approach. In *WASSA*, pages 50–58. ACL.

S. K. Bharti, K. S. Babu, and S. K. Jena. 2015. Parsing-based sarcasm sentiment recognition in twitter data. In *ASONAM*, pages 1373–1380. ACM.

S. K. Bharti, K. S. Babu, and S. K. Jena. 2015. Parsing-based sarcasm sentiment recognition in twitter data. In *ASONAM*, pages 1373–1380. ACM.

Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Opinion mining in twitter how to make use of sarcasm to enhance sentiment analysis. In *ASONAM*, pages 1594–1597. ACM.

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

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

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *CoNLL*, pages 107–116. ACL.

Shelly Dews, Joan Kaplan, and Ellen Winner. 1995. Why not say it directly? the social functions of irony. *Discourse Processes*, 19(3):347–367.

Megan L. Dress, Roger J. Kreuz, Kristen E. Link, and Gina M. Caucci. 2008. Regional variation in the use of sarcasm. *JLS*, 27(1):71–85.

Lori J. Ducharme. 1994. Sarcasm and interactional politics. *Symbolic Interaction*, 17(1):51–62.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*. ELRA.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *HLT*, pages 581–586. ACL.

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

Delia Irazú Hernández Farías, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2015. Valento: Sentiment analysis of figurative language tweets with irony and sarcasm. In *SemEval*, pages 694–698. ACL.

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

Julia Jorgensen, George A Miller, and Dan Sperber. 1984. Test of the mention theory of irony. *Journal of Experimental Psychology: General*, 113(1):112–120.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016a. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *LaTeCH*, pages 95–99. ACL.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *IJCNLP*, pages 757–762. ACL.

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

David S. Kaufer. 1981. Understanding ironic communication. *Journal of Pragmatics*, 5(6):495 – 510.

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

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML*, pages 1188–1196. PMLR.

Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *WASSA*, pages 29–37. ACL.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC*. ELRA.

Neal R Norrick. 1994. Involvement and joking in conversation. *Journal of Pragmatics*, 22(3):409 – 430.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.

Penny M Pexman. 2005. Social Factors in the Interpretation of Verbal Irony: The Roles of Speaker and Listener Characteristics.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING*, pages 213–223. ACL.

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

Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4):754–760.

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

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

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading inbetween. In *ACL*, pages 1010–1020. ACL.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *SemEval*, pages 39–50. ACL.

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

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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.