iSarcasm: A Dataset of Intended Sarcasm

Citation for published version:
Oprea, SV & Magdy, W 2020, iSarcasm: A Dataset of Intended Sarcasm. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics (ACL), pp. 1279–1289, 2020 Annual Conference of the Association for Computational Linguistics, Virtual conference, United States, 5/07/20. https://doi.org/10.18653/v1/2020.acl-main.118

Digital Object Identifier (DOI):
10.18653/v1/2020.acl-main.118

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

iSarcasm: A Dataset of Intended Sarcasm

Silviu Vlad Oprea
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
silviu.oprea@ed.ac.uk

Walid Magdy
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
wmagdy@inf.ed.ac.uk

Abstract

We consider 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. Examining the state-of-the-art sarcasm detection models on our dataset showed low performance compared to previously studied datasets, which indicates that these datasets might be biased or obvious and sarcasm could be a phenomenon under-studied computationally thus far. By providing the iSarcasm dataset, we aim to encourage future NLP research to develop methods for detecting sarcasm in text as intended by the authors of the text, not as labeled under assumptions that we demonstrate to be sub-optimal.

1 Introduction[1]

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 sarcasm detection. The effectiveness of such 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. For instance, Dress et al. (2008) notice a lack of consistence in how sarcasm is used by people of different socio-cultural backgrounds. As a result, an utterance *intended* sarcastic by its author might not be *perceived* as such by audiences of different backgrounds (Rockwell and Theriot, 2001; Oprea and Magdy, 2020).

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 sub-optimal for capturing the sarcastic intention of the authors of the texts. As a result, existing models trained on such datasets 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 both sarcastic and non-sarcastic tweets that they had posted in the past. For each sarcastic tweet, we asked them to explain why it was sarcastic and how they would convey the same meaning non-sarcastically. Labels were thus implicitly specified by the authors themselves. We implemented restrictive quality control to exclude spurious survey responses. We then asked a trained linguist to manually check the sarcastic tweets and further label them into the subcategories of sarcasm defined by Leggitt and Gibbs Jr. (2000).

We further collected third-party sarcasm labels for the tweets in iSarcasm from workers on a crowdsourcing platform. 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. On iSarcasm, annotators recognise author labels with an F-score of 0.616. This indicates that sarcasm is a subjective phenomenon, challenging even for humans to detect. Further, it demonstrates

[1] **This article is a preprint of an article accepted for publication at ACL 2020.**

that using third-party annotators to label texts for sarcasm can lead to inaccurate labels.

We implemented state-of-the-art sarcasm detection models (Tay et al., 2018; Hazarika et al., 2018; Van Hee et al., 2018) and tested them on iSarcasm, to investigate their effectiveness in capturing sarcasm as intended by the authors. While these models achieve F-scores reaching 0.874 on existing datasets, they yield a maximum F-score of 0.364 on iSarcasm, suggesting that previous datasets might be biased or obvious. This highlights the importance of developing new approaches for sarcasm detection that are more effective at capturing author intention.

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 tweets. Each sarcastic tweet has an extra label indicating the category of sarcasm it belongs to. We publish the dataset publicly for research purposes[2].

2 Background

2.1 Intended and Perceived Sarcasm

The way sarcasm is used can vary across sociocultural 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. Females seem to have a more self-deprecating attitude when using sarcasm than males. Rockwell and Theriot (2001) find some cultures to associate sarcasm with humour more than others. There are also 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 intended sarcastic by its author might not be perceived as such by the audience (Jorgensen et al., 1984). Conversely, the audience could perceive the utterance as sarcastic, even if it was not intended as such.

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, for tasks such as opinion mining, as well as possible interpretations by audiences of different sociocultural backgrounds, for tasks such as hate-speech detection.

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 method. Texts are considered positive examples (sarcastic) if they meet predefined criteria, such as containing specific tags, such as #sarcasm for Twitter data (Ptáček et al., 2014), and /s for Reddit data (Khodak et al., 2018), or being posted by specific social media accounts (Barbieri et al., 2014a). Negative examples are usually random posts that do not match the criteria. Table 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 no manual effort. However, as we discuss in Section 3, the labels produced can be very noisy.

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) who recently presented their dataset as part of a SemEval shared task for sarcasm detection. It is a balanced dataset of 4,792 tweets. We call it the *SemEval-2018 dataset*.

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.

[2] 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); Benamara et al. (2017); Cignarella et al. (2018); Van Hee et al. (2018); Bueno et al. (2019)	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.

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. 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 (Tay et al., 2018) uses a neural intra-attention mechanism to capture incongruity.

Contextual models These models utilize information from both the text and the context of its disclosure, such as author information. There is a limited amount of work in this direction. 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 an embedding of that document. Building on this, Hazarika et al. (2018) extract additional personality features from the merged historical tweets with a model pre-trained on a personality detection corpus. Using the same strategy, Oprea and Magdy (2019) build separate embeddings for each historical tweet and identify author context with their the weighted average.

Despite reporting encouraging results, all previous 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 sarcasm in texts as intended by the authors of the texts without noise. We now discuss how noise can occur.

3 Limitations of Current Labelling Methods

In this section, we discuss limitations of current labelling methods that make them sub-optimal for capturing intended sarcasm. We demonstrate them 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. 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 false positives. The assumption that some accounts always generate sarcasm (Barbieri et al., 2014a) is similarly problematic. In addition, the intended sarcasm that distant supervision does capture might be of a specific flavor, such that, for instance, the inclusion of a tag would be essential to ensure inferability. Building a model trained on such a dataset

	with tag	without tag
annot. sarcastic	345	26
annot. non-sarcastic	486	975

Table 2: The agreement between manual annotation and the presence of sarcasm tags in the Riloff dataset, as discussed in Section 3.2.

might, therefore, be biased to a specific flavour of sarcasm, being unable to capture other flavours, increasing the risk of false negatives and limiting the ability of trained models to generalise. 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.

3.2 Limitations of Manual labelling

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, 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 sub-optimal for capturing intended sarcasm. To investigate this further, we looked at the Riloff dataset, which 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 that contained the predefined hashtags 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. We could not perform the same analysis on the SemEval-2018 dataset because only the text of the tweets is provided, hashtags are filtered out, and tweet IDs are not available.

As we have shown, both labelling methods use a proxy for labelling sarcasm, in the form of predefined tags, predefined sources, or third-party annotators. As such, they are unable to capture the sarcastic intention of the authors of the texts they label, generating both false positives and false negatives. Our objective is to create a noise-free dataset of texts labelled for sarcasm, where labels reflect the sarcastic intention of the authors.

4 Data Collection

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 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 implemented the following quality control steps:

- 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 [3].

- Tweets should contain at least 5 words, excluding any hashtags and URLs;
- Links to tweets should not have been submitted in a previous response;
- Responses submitted in less than three minutes are discarded.

Each contributor agreed on a consent form before entering the survey, which informed them that only the IDs of the tweets they provide will be made public, to allow them to delete a tweet anytime and thus be in control of their own privacy in the future. They have agreed that we may collect public information from their profile, which is accessible via the Twitter API as long as the tweets pointed to by the provided IDs are not removed.

We published our survey on multiple crowdsourcing platforms, including Figure-Eight (F8), Amazon Mechanical Turk (AMT) and Prolific Academic (PA)[4]. 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

We then asked a trained linguist to inspect each collected sarcastic tweet, along with the explanation provided by the author and the non-sarcastic rephrase, in order to validate the quality of the response and further assign the tweet to one of the following categories of *ironic speech* defined by Leggitt and Gibbs Jr. (2000):

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.

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 part of our dataset, which we consider later as the test set, and compare the resulting *perceived sarcasm* labels to the *intended sarcasm* labels collected from the authors of the tweets. Our aim was 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 to collect three annotations for each tweet in the iSarcasm dataset, and considered the dominant one as the label, which follows the same procedure as with building the Riloff dataset (Riloff et al., 2013).

5 Data Statistics and Analysis

5.1 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 the quality control steps described in Section 4, and further 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 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 1 shows the tweet length distribution across iSarcasm. The average length of

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

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

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.
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: Examples of sarcastic tweets from our datasets along with the explanations that authors gave to what makes their tweets sarcastic (explanation) and how they can rephrase them to be non-sarcastic (rephrased).

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 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 2 shows the age and gender distributions across contributors.

In iSarcasm, we investigated the presence of the hashtags #sarcasm, #sarcastic, and others often used to mark sarcasm in previous distant supervision datasets. 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, and that these tags might capture only one flavor of sarcasm, not sarcasm present on social media in general.

Regarding the categories of sarcasm, assigned by the linguist to the sarcastic tweets, Table 3 shows the distribution of the tweets into these categories. As shown, sarcasm and irony are the largest two categories (73%), while understatement is the smallest one (with only 12 tweets). Table 4 shows examples of the sarcastic tweets, along with the explanations and rephrases provided by the authors.

iSarcasm is published 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

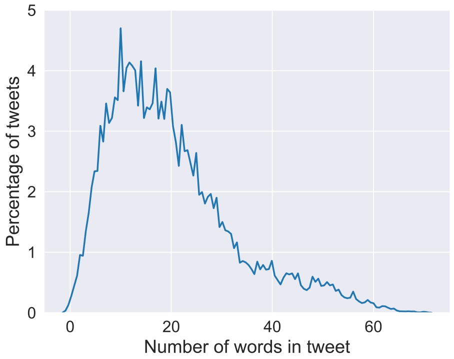

Figure 1: Tweet length distribution across iSarcasm.

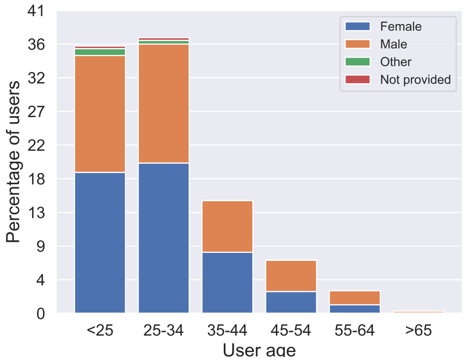

Figure 2: Age and gender distributions across the Twitter users who provided tweets in iSarcasm.

intended sarcasm labels. For sarcastic tweets we also provide the category of ironic speech they belong to. This is in accordance with the consent form that the contributors have agreed to, whose privacy we take seriously. Nonetheless, we still offer the tweets text along with the explanations and rephrases of the sarcastic tweets provided by the authors for free for research purposes, under an agreement that protects the privacy of our contributors.

5.2 Third-Party 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 $\kappa_{23} = 0.36$, which highlights the high subjectivity of the task. We used majority voting to select the final perceived sarcasm label for each tweet. Table 5 shows the disagreement 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 actually not intended to be sarcastic by their authors. This supports our argument that third-party annotation for sarcasm should not be trusted.

	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 iSarcasm test set.

6 Detecting Intended Sarcasm

In the following, we examine the effectiveness of state-of-the-art sarcasm detection models on iSarcasm. We aim to investigate their ability to detect intended sarcasm rather than sarcasm labeled using distant supervision or manual annotation. As we have shown, these labelling methods could produce noisy labels. We experiment with those models that have achieved state-of-the-art results on previous benchmark datasets for sarcasm detection.

6.1 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 and manual annotation for actual labelling. The other two datasets, Ptacek (Ptáček et al., 2014) and SARC (Khodak et al., 2018), are labeled using distant supervision. As mentioned earlier, we managed to collect 1,832 tweets from the Riloff dataset. SemEval-2018 is a balanced dataset consisting of 4,792 tweets. For the Ptacket dataset, we collected 27,177 tweets out of the 50K published tweet IDs. Finally, The SARC datasets 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 sarcastic and non-sarcastic comments.

6.2 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 as well as contextual information available about the authors. In this paper we only focus on content modelling, using a convolutional network (**3CNN**) in a setting similar to what they describe. **3CNN** uses three filter types of sizes 3, 4, and 5, with 100 filters for each size.

SemEval-2018 dataset The SemEval 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 type of sarcasm present[5]. Wu et al. (2018) report the best results on 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 dataset (Van Hee et al., 2018) to predict three components: the binary sarcasm label, one of the four types 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. Wu et al. (2018) report an F-score of 0.674 using a fixed dropout rate of 0.3 in all layers. They further report an F-score of 0.705 by averaging the performance of 10 Dense-LSTM models, varying the dropout rate to random values between 0.2 and 0.4. We implement and train it 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 6, providing confidence in the correctness of our implementations.

6.3 Results and Analysis

Table 7 reports precision, recall and f-score results on the test set of iSarcasm using the detection models discussed, alongside third-party annotator performance. As shown, all the models perform significantly worse than humans, who achieve an F-score of only 0.616. MIARN is the best performing model with a considerably low F-score of 0.364, compared to its performance on the Riloff and Ptacek datasets (0.741 and 0.874 F-scores 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 imbalanced, respectively. Similarly, Dense-LSTM

[5] We use “sarcasm” to mean what they refer to as “verbal irony”.

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 6: F-score yielded by our implementations of state-of-the-art models on previous datasets, compared to published results on those datasets.

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

achieves 0.318, compared to 0.666 on SemEval-2018.

Previous models that achieved high performance in detecting sarcasm on datasets sampling perceived sarcasm (third-party labels) or hash-tagged sarcasm (distant supervision) have failed dramatically to detect sarcasm as meant by its author. This motivates the need to develop more effective methods for detecting intended sarcasm. Potentially, building models that account for sociocultural traits of the authors (available on, or inferred from, their Twitter profiles), or consider other contextual elements to judge the sarcasm in our dataset (Rockwell and Theriot, 2001). Previous research has considered certain contextual elements (Bamman and Smith, 2015b; Amir et al., 2016; Hazarika et al., 2018; Oprea and Magdy, 2019), but only on sarcasm captured by previous labelling methods.

We believe the iSarcasm dataset, with its novel method of sampling sarcasm as intended by its authors, shall revolutionise research in sarcasm detection in the future; and open the direction for new sub-tasks, such as sarcasm category prediction, and sarcasm decoding/encoding, using information found both in the tweets themselves, and in the explanations and rephrases provided by the authors, available with each sarcastic tweet in the dataset.

7 Conclusion and Future Work

In this paper, we presented iSarcasm, a dataset of intended sarcasm consisting of 4,484 tweets labeled and explained by their authors, and further revised and categorised by an expert linguistic. 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.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1); the University of Edinburgh; and The Financial Times.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

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.

Farah Benamara, Cyril Grouin, Jihen Karoui, Véronique Moriceau, and Isabelle Robba. 2017. Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du défi fouille de textes deft2017.

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.

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

Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69:86 – 99.

Reynier Ortega Bueno, Francisco M. Rangel Pardo, Delia Irazú Hernández Farías, Paolo Rosso, Manuel Montes y Gómez, and José Medina-Pagola. 2019. Overview of the task on irony detection in spanish variants. In *IberLEF@SEPLN*.

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.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, 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.

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.

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.

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.

John S Leggitt and Raymond W Gibbs Jr. 2000. Emotional reactions to verbal irony.

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.

Saif Mohammad and Felipe Bravo-Marquez. 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. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *ACL*, pages 2854–2859. ACL.

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

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

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.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*. AAAI Press.

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

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

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.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multitask learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56, New Orleans, Louisiana. Association for Computational Linguistics.

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.