Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal

Aniruddha Ghosh
University College Dublin, Ireland.
Aniruddha.Ghosh@ucdconnect.ie

Tony Veale
University College Dublin, Ireland.
Tony.Veale@ucd.ie

Abstract

Sarcasm is a pervasive phenomenon in social media, permitting the concise communication of meaning, affect *and* attitude. Concision requires wit to produce and wit to understand, which demands from each party knowledge of norms, context and a speaker's mindset. Insight into a speaker's psychological profile at the time of production is a valuable source of context for sarcasm detection. Using a neural architecture, we show significant gains in detection accuracy when knowledge of the speaker's mood at the time of production can be inferred. Our focus is on sarcasm detection on Twitter, and show that the mood exhibited by a speaker over tweets leading up to a new post is as useful a cue for sarcasm as the topical context of the post itself. The work opens the door to an empirical exploration not just of sarcasm in text but of the sarcastic state of mind.

1 Introduction

Oscar Wilde memorably described sarcasm as "the lowest form of wit but the highest form of intelligence." Though sarcasm lacks the sophistication of irony, and does little to conceal the speaker's disdain for a target, it is a figurative device that requires as much intelligence from its consumers as its producers. The concision with which sarcasm and irony allow speakers to conflate propositional content and affective stance makes it a pervasive mode of communication in the 140-character *tweets* of Twitter. By combining an overtly positive attitude with a meaning that is more deserving of scorn, sarcasm allows speakers to communicate disappointment about a state of affairs that bites (or etymologically "cuts the flesh") of an addressee. It conveys the feeling the speaker would wish to experience ("I love it when ...") with the state of affairs that up-ends this feeling ("... my friends forget my birthday"). It often combines politeness with mockery to disguise the appearance of hostility while heightening its effect on a listener (Brown and Levinson, 1978; Dews and Winner, 1995). It establishes a wry environment (Dews and Winner, 1999) that has its roots in social norms *and* the speaker's state of mind.

Psychological theories of irony, such as *echoic reminder theory* (Kreuz and Glucksberg, 1989) and *implicit display theory* (Utsumi, 2000b) have yet to fully translate into text-analytic methods. Neuropsychology researchers who have sought patterns of brain activity to identify the neural correlates of sarcasm note that an understanding of sarcasm is highly dependent not just on the context of an utterance but on the state-of-mind and personality of the speaker, as well as on facial expressions and prosody (Shamay-Tsoory et al., 2005). Without the latter markers, purely textual detection must depend largely on the content and context of an utterance, though speaker personality and state-of-mind can also be approximated via text-analytic means. Probabilistic classification models that exploit textual cues – such as the juxtaposition of positive sentiment and negative situations (Riloff et al., 2013), discriminative words and punctuation marks (Davidov et al., 2010), and emoticon usage (González-Ibánez et al., 2011) – have achieved good performance across domains, yet these models typically suffer from an absence of psychological insight into a speaker and topical insight into the context of utterance production. Kreuz and Link (2002) argue that the likelihood of sarcasm is proportional to the amount of knowledge shared by speaker and audience, which includes knowledge of the world *and* knowledge of the speaker and audience. Personality is defined

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 482–491
Copenhagen, Denmark, September 7–11, 2017.

by Olver and Mooradian (2003) as the "enduring characteristics of the individual" though *mood* – which is changeable – is perhaps just as useful if sampled in a timely fashion. The difference between personality and mood can be likened to that between climate and weather. Tausczik & Pennebaker (2010) have developed a Twitter-based mood analysis web service at *AnalyzeWords.com* which uses a variety of psycholinguistic criteria and the LIWC (*Linguistic Inquiry and Word Count*) resource[1] to quantify the recent mood – i.e. the *recent weather* – of a user along 11 dimensions ranging from Arrogance/Remoteness to Anger and Analyticity. To exploit the stable personality of an online user, Celli et al. (2016) sought a correlation between *Big Five* personality traits (Costa and McCrae, 2008) and the LIWC-quantifiable dimensions found in re-tweets amongst Twitter users. (Rajadesingan et al., 2015) have also shown how relevant aspects of personality can be acquired from a speaker's past tweets. Since personality and mood can each influence the detection process, they underpin our first research question: To what extent can the quantifiable dimensions of either lead to a better understanding of sarcasm? Reliable detection depends as much on the context of an utterance – which provides the motivation for sarcasm – as its content. Consider e.g.:

Speaker Utterance: @MSNBC of course all of those jobs will be in China
In reply to *@realDonaldTrump*: I will be the greatest jobs-producing president that God ever created.

The speaker's sarcastic intent cannot be grasped without knowledge of the larger context. This issue provides our second research question: How can we usefully incorporate utterance context into a neural network model of sarcasm detection? Sarcasm is ubiquitous but always in flux, relying on a changing swirl of socially relevant viewpoints. The following tweet is sarcastic by virtue of its echoic mockery of a widely ventilated opinion:

Time to get my Sunday dose of #fakenews from the failing @nytimes.

This begs the third research question that we explore in the following sections: How can we train our sarcasm detection model to exploit evolving social norms and public opinions?

[1] https://liwc.wpengine.com/

2 Related Work and Ideas

Sarcasm has been extensively researched by linguists and psychologists (Gibbs and Clark, 1992; Gibbs and Colston, 2007; Kreuz and Glucksberg, 1989; Utsumi, 2000a), yet due to the limited availability of stimuli, sarcasm detection in text has relied chiefly on the recognition of stock patterns and lexical cues. Sarcasm often highlights failed expectations by engaging in a pragmatic pretense that is designed to be seen through (Campbell and Katz, 2012), so cues such as interjections, intensifiers, punctuation and markers of non-veridicality and hyperbole play a crucial role in recognizing sarcastic intent. Likewise, stock plaudits such as "yay!" or "great!" are common in sarcastic product reviews (Tsur et al., 2010), while hashtags such as *#sarcasm*, as compressed vehicles for user intent, are often used to self-annotate sarcastic texts (Davidov et al., 2010). Liebrecht et al. (2013) used topic-specific information and n-grams as discriminative features, while (Lukin and Walker, 2013) showed that phrases such as "no way", "Oh really?" and "not so much" serve to flag a sarcastic intent when used with specific linguistic patterns.

Capelli et al. (1990); Woodland and Voyer (2011) suggest that contextual awareness is a necessary precursor to identifying sarcasm. Sarcasm is a response to a motivating context that appears to force a rueful incongruity between a text and its context. Exploiting the *principle of inferability* (see Kreuz (1996)), Bamman and Smith (2015) modeled shared common knowledge by extracting features from context, the author, and the audience. Khattri et al. (2015) identified sarcasm by seeking a strong contrast in affect toward named entities in current vs. historical tweets, while Rajadesingan et al. (2015) also exploited a contrast in statistically-derived author traits across current and historical tweets. Zhang et al. (2016) use similar sources of contextual information to show the effectiveness of a neural network over more traditional approaches involving manually-selected, discrete features, claiming that automatic feature induction can uncover more subtle markers of sarcasm. Amir et al. (2016) argue that sarcasm detection hinges on speaker modeling, and exploited *user embedding* to quantify incongruity between utterances and the behavioral traits of their authors. These methods measure the disparity between an utterance and expectations arising from knowledge of context or speaker or both together.

We build on this double-grounding for sarcasm to improve detection in a neural network model of sarcasm and thereby address our first two research questions. We model the speaker *at the time of utterance production* using mood indicators derived from the most recent prior tweets, and model context using features derived from the proximate cause of the new utterance, the tweet to which an utterance is a response. For our third research question, we present a novel feedback-based annotation scheme that engages authors of training/test tweets in a process of explicit annotation, feeding new examples back into the model. Section 3 outlines the kind and source of features exploited in the model. Section 4 outlines our methods of data collection and annotation. Section 5 presents the neural network model, while section 6 & 7 present our experimental set-up and analysis of results. Finally, section 8 offers some closing remarks.

3 Psychological dimensions and Sarcasm

We cannot perceive a user's state-of-mind directly on Twitter, but we might infer one's current disposition from an analysis of recent tweets, as linguistic expressions tend to be congruent with an author's state-of-mind (Campbell and Katz, 2012). An informative if *low-res* psychological portrait is sketched by web services such as *AnalyzeWords* (Tausczik and Pennebaker, 2010), which analyzes the most recent 1000-words or so of a Twitter user using LIWC to score the user on 11-dimensions: *Upbeat, Worried, Angry, Depressed, Plugged in, Personable, Arrogant, Spacy, Analytic, Sensory* and *In-the-moment*. Sarcasm is often perceptible in the incongruity between utterance and context (Joshi et al., 2015) but it can also be conveyed by an incongruity between text and recent mood.

To understand the relationship between these 11 dimensions (each scored 0..100) and a propensity for sarcasm, we performed a k-Nearest Neighbors (KNN) clustering of the Twitter users that provide the tweets of our sarcastic data set. The *AnalyzeWords* snapshot of each user was taken at the time of that user's tweet in the dataset. A value of 30 for k was chosen empirically to ensure a decent size for the clusters. By calculating Spearman correlations between each group and the 11 *AnalyzeWords* dimensions, we estimated the affinity for sarcasm of different dimensions. Unsurprisingly, we observed that clusters showing a high correlation with negative dimensions, such as *Angry*, also tend to use positive expressions such as 'funny" and 'wow" to mark sarcasm. Here is an example:

@realDonaldTrump They can all fit in your head? Wow! Have you seen someone about this?

Unless one knows that *@realDonaldTrump* often elicits anger, or that the author scored 83 (of 100) for *Angry*, this tweet might seem quite positive. Valence shifters such as "not" might also suggest literal positivity if not for the implicit anger of the author. At the time of the following tweet *AnalyzeWords* scored its author as *Angry=98*.

@realdonaldtrump funny the founder of the birther movement is saying that he's not racist #trumpbirther

Polarizing figures such as *@realDonaldTrump* are magnets for sarcasm on Twitter. By identifying these magnets, we can better detect the sarcasm of a tweet that offers plaudits for negative qualities. We use *AnalyzeWords* to obtain the popular affective feelings for common addressees by averaging the affective dimensions of the users that tweet at them. The top 5 magnets for sarcasm in our data-set of 18K sarcastic tweets are *@hillaryClinton, @realDonaldTrump, @bernieSanders, @AP* and *@megynKelly*. Of these, *@hillaryClinton* is the biggest target for *Angry* tweets while *@megynKelly* is the biggest target for *Analytic* tweets.

Addresses in the political domain score high for both angry tweets and analytic tweets: people analyze the news *and* shoot the messenger. We see much less analyticity – a tendency to use complex expressions linked with logical connectives – in tweets about popular entertainers. To mock such targets, users tend to use affective words that contrast with overall public opinion. The magnet with the highest mean *Angry* score for the tweets that target him is *@realDonaldTrump*, yet 63% of the affective words in the tweets that target him in the data-set are positive. Knowing that *@realDonaldTrump* is a magnet for anger can help a sarcasm detector overcome this positive bias.

4 Dataset Construction

Tweets with sarcastic intent are often misclassified due to a lack of shared context or knowledge between speaker and annotator. Opposing social beliefs and a dearth of topical or personal knowledge can lead to serious misjudgments. Relevant tweet sets can be harvested by searching sarcasm specific hashtags (e.g. #sarcasm). This approach

overlooks tweets that are not explicitly tagged as sarcastic by their authors. Thus we have devised a feedback-based system that contacts tweet authors directly *after-the fact* to ask for their authoritative self-annotations for a potentially sarcastic tweet.

4.1 Data collection

To collect annotations from authors for *their own* tweets, we used a Twitterbot named *@onlinesarcasm* to exploit the "retweet with comment" function in Twitter. The bot chooses randomly from tweets that are addressed to any of 700 top Twitter users (as listed by *TwitterCounter.com*), as we expect high-profile figures to be magnets for sarcasm from others. The bot retweets a chosen tweet (s_i) to its author, appending a yes/no question (q_i) as a comment to elicit a reply.

At the time of retweeting (s_i), the 11 *AnalyzeWords.com* dimensions (aw_i) of the tweet's author (u_i) are saved, along with the context tweet (s_j) by author (u_j) that provoked (s_i). Authors respond to the bot by favoriting/retweeting the bot's request or via a reply (re_i) containing #Yes or #No. Author responses often contain more than a simple #Yes or #No response, and so, after observing a series of responses the following linguistic rules were used to extract the training annotations:

- If the number of *retweets* (r_i) or *likes* (l_i) for q_i is non-zero or re_i contains #Yes, then s_i is deemed positive for sarcasm.
- If re_i contains #No or an explicit mention of 'not sarcastic' or 'no sarcasm' or 'truth', then s_i is deemed negative for sarcasm.

We discarded any s_i lacking a context tweet s_j. Using author feedback, a data set of 40K tweets was collected, comprising 18K tweets acknowledged as sarcastic and 22K deemed non-sarcastic. For another test set, we collected 1200 tweets: 550 tweets acknowledged as sarcastic by their authors and 650 acknowledged to be non-sarcastic.

4.2 External datasets

In addition to our own training and test sets, whose annotations come directly from tweet authors, we also used 5 Twitter datasets where tweet information, fetched by tweet identifier, contains identifier of context tweet (Ptáček et al., 2014; Bamman and Smith, 2015; Rajadesingan et al., 2015; Cliche, 2014), from which motivating contexts can be discerned for each. (This contextual requirement prevents us from considering even more of the available sarcasm datasets.) For the context tweets s_j for each s_i in these sets we collected the most recent linked tweets of s_i. To obtain the 11 *AnalyzeWords.com* dimensions for tweet authors, we collect the 50 tweets of u_i posted just prior to s_i, and use the LIWC to estimate the 11 dimensions (Anger, Arrogance, etc.) from those tweets. As *AnalyzeWords.com* does not provide retrospective analyses, and as its code is not public, we reverse-engineered a substitute using the LIWC by following the creators' guidelines in (Tausczik and Pennebaker, 2010). For subsequent evaluations, the 5 external datasets were split into 3 parts each: 80% for training, 10% for development/tuning, and 10% for testing.

5 The Neural Network Model

Ghosh and Veale (2016) described an Artificial Neural Network (ANN) model built around layers of CNNs (Convolutional Neural Networks) and LSTMs (Long Short Term Memory) for sarcasm detection to efficiently capture contrasting text signals of sarcasm within a tweet. We build here on this model as shown in Fig.1, adding input features for the psychological profile of the author and the context of the tweet to those for the tweet itself. The LSTM layer (Hochreiter and Schmidhuber, 1997) captures dependencies amongst non-adjacent contrasting signals for sarcasm within each s_i. We extend this architecture to include a context tweet s_j for each s_i, but instead of concatenating s_j and s_i at the input layer, we stitch them together after the LSTM layer. The text input layer is initialized with embeddings from Google's *Word2Vec* model (Mikolov et al., 2013) with a dimension setting of 300. To further integrate features reflecting the state of mind of the speaker at utterance-time, the values aw_i ($i = 1...11$) for each s_i are concatenated with the feature vector of s_j & s_i in the merge layer. We use a bi-directional LSTM (BLSTM) and forego a *maxpooling* layer to increase throughput to the BLSTM. We prevent overfitting using a dropout layer with a dropout rate of 0.25 after the BLSTM layers. The concatenation layer combines the feature maps of the source and context tweets (s_i & s_j) along with a vector of $aw_{1...11}$ for the author u_i. The concatenation yields a merge layer of size $\Re^{f(2(|s|m+1)+l)}$ where f, s, m and l are, respectively, the number of BLSTM units, the length of the input se-

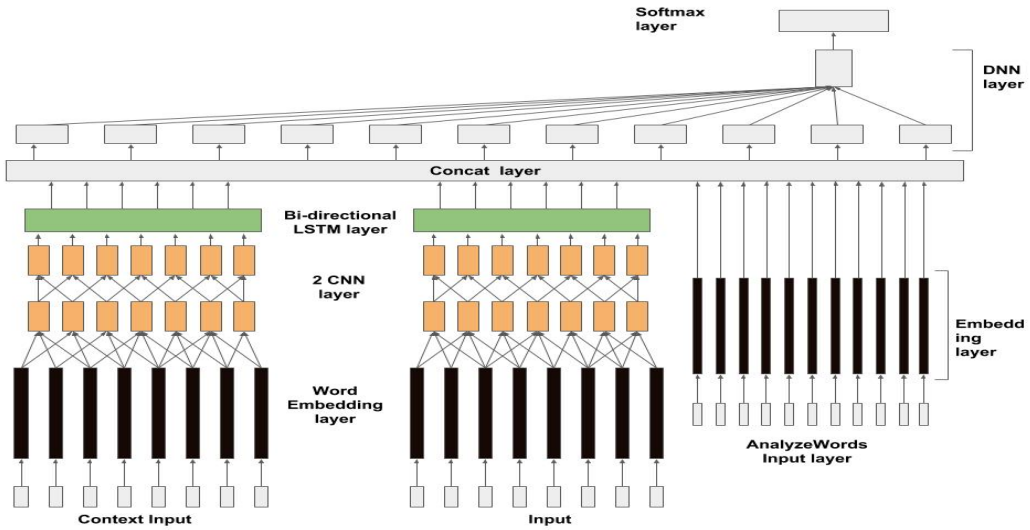

Figure 1: A Neural Architecture for Detecting Sarcasm in Contextualized Utterances

quence, the width of the CNN filter and the length of aw. Notice that the features for a tweet s_i and its immediate context s_j – which we consider the proximate cause of the sarcasm (if any) in s_i – are concatenated only after they have passed through separate sets of CNN and LSTM layers (CNN1 + BLSTM1 and CNN2 + BLSTM2). It is important to keep a tweet and its context separate for as long as possible, as the model is designed to recognize an inherent incongruity between each. This incongruity becomes diffuse if the inputs are combined too soon. EAW is the embedding layer for the 11 *AnalyzeWords* dimensions; it combines the vectors of s_j, s_i and aw, and passes the concatenated features to a Deep Neural Network (DNN) to discriminate both classes (sarcasm vs. non-sarcasm). The code[2] is developed using Keras[3].

6 Evaluation and Experimental Setup

Success with a neural architecture requires apt input features and an equally apt selection of hyper-parameters. After performing a grid search over hyper-parameters, the best configuration of the CNN, LSTM and DNN layers places 1280 hidden memory units into each layer and uses a CNN filter width of 3. A simple baseline will use only the textual content of a tweet s_i without a context s_j or an affective profile aw of the author u_i. To appreciate the contribution of different input sources of information we trained the network on different combinations of these sources.

[2]https://github.com/AniSkywalker/SarcasmDetection
[3]https://keras.io/

6.1 Addressee information

If s_i is addressed to u_j, this information can provide additional insights into s_i's tone. In the TTIA *(Target Tweet Including Addressee)* setting the name of the addressee (but not an estimation of the public opinion of the addressee, as so few addresses are actually famous) is added to the baseline along with s_i. If the addressee is a magnet for sarcasm, aspects of this magnetism should still impress themselves on the network during training.

6.2 Contextual Information

In a variant of the baseline called CT *(Context Tweet)*, the features of the tweet s_j to which the target s_i is a response are also added as inputs to the model, to be stitched together with the features of s_i at the concatenation layer. Changes in performance with and without CT will allow us to estimate the value of context in sarcasm detection.

6.3 Author Profile Information

The 11-dimensional *AnalyzeWords* snapshot aw for author u_i at the time s_i is posted offers valuable insights into the intent of u_i. In the PD *(Psychological Dimensions)* configuration, the 11 affective dimensions aw_i are added to the model. They pass through an embeddings layer to be combined with utterance (and possibly context) features at the concatenation layer. To determine the relative contribution of each dimension aw_i to detection competence, we trained the model in two extra modes. In the first, we fed the model with false values for each aw_i, varying values from 0 to

100, to observe the effects on accuracy when e.g. *Angry* is over- or under-estimated for u_i. In the second extra mode, we excised each aw_i, one at a time in different training runs, to quantify its lack on the model.

6.4 Automatic adaptation

Online sarcasm is often used to comment on the vagaries of politics and current affairs. As topicality is of the essence, we expect a model that regularly acquires new author-annotated training data to bootstrap itself will adapt better to the times and yield better results. To estimate the benefits of bootstrapping we tested the model on an evolving version of the data-set that acquired new training data each week for a month in August 2016.

7 Results & Analysis

Table 3 shows the recall (R), precision (P) and f-score (F1) for our model, called *Sarcasm Magnet*, with alternate configurations on different datasets. The configuration for each setup is given in the second row (e.g. the addition of context tweets requires the use of two LSTMs and two CNNs). Setup TTEA is the baseline which uses only the text of a target tweet; it excludes addressee handles, context tweets (CT) and the psychological dimensions (PD) of authors. Setup TTIA adds addressee handles to the baseline to give our model a small boost, mostly in recall. Setup TTEA+CT adds context tweets to the baseline, yielding a significant boost since a good deal of sarcasm is conversational in nature. In this setup the most significant improvement in recall was observed with the Bamman dataset (Bamman and Smith, 2015). In setup TTIA+CT, which uses context *and* addressee handles, no significant improvement over TTEA+CT is observed, except for precision on Bamman's dataset. In setup TTEA+PD, the affective profile of each author at tweet-time is added to the baseline to yield a significant boost in performance almost as large as that for TTIA+CT. Setup TTIA+CT+PD includes all available information sources (addressee, content, and psychological profile). This column reports (in parentheses) the performance on each dataset by the dataset creator's own system, which is either publicly available or re-implemented from their paper. The results show that *Sarcasm Magnet* beats the state of the art for these data-sets.

Table 1 shows the effect on the model's performance in the absence of specific aw_i values. A boost in recall and a drop in precision shows the bias of the model shifting towards sarcasm when the space of non-sarcastic tweets overlaps with that of sarcastic tweets in the absence of an aw_i that confirms literal intent. So political tweets may be mis-classified as sarcasm in the absence of values for *Angry*, *Depressed*, and *Worried*, suggesting that sarcastic authors often seem less angry, depressed or worried. A drop in precision and recall when *Arrogant/Remote*, *Analytic*, *Plugged in* and *In-the-moment* dimensions are absent suggests sarcastic people to be more socially active and aware, and smarter but more arrogant.

Dimension omitted	Precision	Recall	F-score
None omitted	0.9	0.89	0.9
Sensory	0.85	0.93	0.89
Plugged In	0.84	0.84	0.84
Depressed	0.78	0.96	0.86
Angry	0.81	0.95	0.87
Spacy/Valley girl	0.78	0.97	0.87
Worried	0.79	0.97	0.87
Arrogant/Distant	0.84	0.85	0.84
Analytic	0.86	0.83	0.84
In-the-moment	0.84	0.86	0.85
Upbeat	0.86	0.91	0.88
Personable	0.87	0.88	0.88

Table 1: Performance of the model when a specific dimension aw_i is omitted from training.

7.1 A Tale of Two Contexts

The CT and PD additions each bring significant improvements in F-score, yet when added jointly they bring no significant increases over either used individually. For each is a form of context drawn from different sources that reflects different intuitions but which ultimately offers much the same insights. The impact of the 11 aw dimensions is lower on the 5 external datasets than for the new feedback-based dataset, no doubt because the *AnalyzeWords.com* snapshot of authors in the latter could be taken directly at tweet-time, whilst for the former it was retrospectively approximated using our own jerry-rigged version based on the LIWC. If the official web service were to allow retrospective analyses of Twitter users at specific times we are confident the improvements on the external datasets would mirror those on our own dataset. For now it is interesting to note the effectiveness of the *AnalyzeWords.com* service at affectively profiling Twitter users at specific times,

which is to say, at specific contexts in their Twitter time-lines.The service boils down the most recent tweets (approx. 1000 words in total) to 11 dimensions that are more than simple functions of the lexical scores in the LIWC. Rather, it analyzes the selected text as a coherent product of a coherent mind-set, to measure a local propensity for hostility, optimism, depression, emotional detachment and preference for reason. To use our earlier analogy, *AnalyzeWords.com* forecasts the psychological weather around an author, not the user's stable climate. Though we may often speak of a "sarcastic personality" as a stable aspect of some speakers, most users of sarcasm will not fall into this category. As such, insight into the recent mind-set of an author is more valuable to a detector than knowledge of one's personality overall.

7.2 Rolling With The Punches

Our feedback-annotated dataset was collected during a fertile period for sarcasm online: the heights of the 2016 US presidential campaign. The main body of the new dataset was collected and annotated (as described earlier) in the early summer of 2016. During the month of August we acquired additional annotated training data in four weekly tranches, to incrementally retrain the model to an evolving political and social context. As shown in

Week	Precision	Recall	F-score
Week 1	0.751	0.752	0.752
week 2	0.790	0.752	0.771
Week 3	0.798	0.775	0.786
Week 4	0.839	0.869	0.85

Table 2: Bootstrapping gains (August, 2016)

Table 2, each weekly tranche of extra training data yielded increased dividends in terms of F-score and precision or recall when evaluating the model on the same test set (of 1,200 tweets, 550 sarcastic and 650 non-sarcastic). The new annotated data harvested in the final week of August yielded the biggest dividends, especially in Recall, perhaps in a reflection of the growing bitterness of the campaign and of frantic campaigning in (and online commentary about) the pivotal *swing states*. As the candidates were revealing more of themselves to voters, the voters were revealing more of themselves to our model. Specifically, in week 1, 4719 sarcastic and 5361 non-sarcastic tweets were added for training; in week 2, an additional 3179 and 6901 were added; in week 3, 3571 and 6509; and in a week 4 reversal, 6504 and 3574 tweets.

8 Conclusions & Future work

Context is vital to the understanding of the fruits of any figurative device, whether metaphor, irony or sarcasm. We have explored two sources of contextual information in this work: the linguistic context of the utterance itself – which we take to be another utterance that is the proximate cause of the text under consideration – and the psychological context of the utterance's author – which we take to be the mind-set that is apparent in the author's most recent writings on Twitter. Each source of context is ultimately grounded in a text and understood in text-analytic terms. It is perhaps not so surprising then that each kind of context yields similarly large improvements to a neural model of sarcasm detection when added in isolation, but no large improvements over either alone when both are combined in a single model.

This work makes three principle contributions to the computational analysis of sarcasm. First, as outlined above, it shows how different kinds of context – from the linguistic to the psychological – can be usefully incorporated to yield improved detection. Second, it shows how accurate annotation of training data can be automated on Twitter by going directly to the source of each training text, to obtain a definitive answer as to its figurative status. So the resulting neural model does not learn to approximate the reasoning of independent human annotators but the mind-set and intent of the authors themselves. Thirdly, and perhaps most usefully for future work by others, this feedback-based dataset will be made available for use by other researchers and in other evaluations. Importantly, this dataset is not merely a collection of yes/no annotated texts, even if the *yes*es and *no*s come from authoritative sources. For each text in the dataset, we can provide the linguistic context to which it is a response, and furthermore, we can provide a psychological snapshot of the author *at the time* the tweet was posted on Twitter. In the end we believe this is the most valuable contribution of the work, as it will allow others to incorporate an understanding of personality and mind-set into their own models of that most personal and moody of figurative devices, sarcasm.

Acknowledgements: This work was funded by Science Foundation Ireland (SFI) via the ADAPT centre for Digital Content Technology.

Dataset		Alternate Configurations of the *Sarcasm Magnet** model						
		TTEA	TTIA	TTEA + CT	TTIA + CT	TTEA + PD	TTIA + PD	TTIA + CT + PD
		CNN1 + LSTM1 + DNN	CNN1 + LSTM1 + DNN	CNN1 + CNN2 + LSTM1 + LSTM2 + CL + DNN	CNN1 + CNN2 + LSTM1 + LSTM2 + CL + DNN	CNN1 + LSTM1 + EAW + CL + DNN	CNN1 + LSTM1 + EAW + CL + DNN	CNN1 + CNN2 + LSTM1 + LSTM2 + EAW + CL + DNN
(Ptáček et al., 2014) (balanced dataset) (S:15K, NS:17K)	P	0.821	0.832	0.908	0.92	0.857	0.86	0.947
	R	0.821	0.832	0.908	0.92	0.857	0.86	0.947
	F1	0.821	0.832	0.908	0.92	0.857	0.86	**0.9472** (0.9466)
(Ptáček et al., 2014) (unbalanced dataset) (S:15K, NS:39K)	P	0.814	0.813	0.926	0.937	0.851	0.843	0.946
	R	0.832	0.833	0.93	0.93	0.833	0.838	0.933
	F1	0.823	0.823	0.928	0.933	0.842	0.84	**0.94** (0.924)
(Bamman and Smith, 2015) (S:8K, NS:7K)	P	0.896	0.90	0.886	0.919	0.825	0.835	**0.9** (0.857)
	R	0.651	0.672	0.819	0.817	0.803	0.827	0.858 (**0.872**)
	F1	0.754	0.77	0.851	0.865	0.814	0.831	**0.878** (0.864)
(Cliche, 2014) (S:50K, NS:100K)	P	0.788	0.8	0.874	0.893	0.884	0.883	0.896
	R	0.751	0.769	0.842	0.843	0.812	0.817	0.862
	F1	0.769	0.784	0.858	0.867	0.846	0.849	**0.879** (0.6)
(Rajadesingan et al., 2015) (S:6K, NS:6K)	P	0.957	0.957	0.957	0.957	0.958	0.958	0.956
	R	0.807	0.807	0.807	0.807	0.861	0.861	0.905
	F1	0.875	0.875	0.875	0.875	0.907	0.907	**0.93** (0.903)
Sarcasm Magnet* (this paper)	P	0.733	0.731	0.84	0.841	0.846	0.853	0.90
	R	0.717	0.732	0.833	0.858	0.852	0.861	0.89
	F1	0.725	0.732	0.836	0.849	0.848	0.856	0.90

Table 3: Evaluation of Sarcasm Magnet (P - precision, R - recall, F1 -f-score, TTEA - target tweet excluding addressee; TTIA - target tweet including addressee; CT - Context Tweet; PD - Psychological dimensions; S - sarcastic; NS - non-sarcastic). All results are for the Sarcasm Magnet model; when available, results obtained by other authors on their own datasets are in parentheses. **Sarcasm Magnet* is the name of the current system and its associated dataset.

References

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976* .

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*. http://homes.cs.washington.edu/ nasmith/papers/bamman+smith.icwsm15.pdf.

Penelope Brown and Stephen C Levinson. 1978. *Universals in language usage: Politeness phenomena*. Cambridge University Press. http://www.mpi.nl/publications/escidoc-66660/.

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. http://dx.doi.org/10.1080/0163853X.2012.687863.

Carol A Capelli, Noreen Nakagawa, and Cary M Madden. 1990. How children understand sarcasm: The role of context and intonation. *Child Development* 61(6):1824–1841. https://www.jstor.org/stable/1130840.

Fabio Celli, Arindam Ghosh, Firoj Alam, and Giuseppe Riccardi. 2016. In the mood for sharing contents: Emotions, personality and interaction styles in the diffusion of news. *Information Processing & Management* 52(1):93–98.

Mathieu Cliche. 2014. The sarcasm detector. `http://www.thesarcasmdetector.com/`.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment* 2:179–198.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Uppsala, Sweden, pages 107–116. http://www.aclweb.org/anthology/W10-2914.

Shelly Dews and Ellen Winner. 1995. Muting the meaning a social function of irony. *Metaphor and Symbol* 10(1):3–19. http://dx.doi.org/10.1207/s15327868ms1001_2.

Shelly Dews and Ellen Winner. 1999. Obligatory processing of literal and nonliteral meanings in verbal irony. *Journal of pragmatics* 31(12):1579–1599. http://psycnet.apa.org/psycinfo/1999-01341-003.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of NAACL-HLT*. pages 161–169.

Deanna W. Gibbs and Herbert H. Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language* 31(2):183–194. doi:10.1016/0749-596X(92)90010-U.

Raymond W. Gibbs and Herbert L. Colston. 2007. *Irony in language and thought: A cognitive science reader*. Psychology Press. https://books.google.ie/books?isbn=0805860622.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 581–586. http://www.aclweb.org/anthology/P11-2102.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780. http://dl.acm.org/citation.cfm?id=1246450.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Association for Computational Linguistics Volume 2: Short Papers*. pages 757–762. https://www.aclweb.org/anthology/P/P15/P15-2124.pdf.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2015. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Lisboa, Portugal, pages 25–30. http://aclweb.org/anthology/W15-2905.

Roger J Kreuz. 1996. The use of verbal irony: Cues and constraints. *Metaphor: Implications and applications* pages 23–38.

Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General* 118(4):374. http://dx.doi.org/10.1037/0096-3445.118.4.374.

Roger J Kreuz and Kristen E Link. 2002. Asymmetries in the use of verbal irony. *Journal of Language and Social Psychology* 21(2):127–143. http://dx.doi.org/10.1177/02627X02021002002.

Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. 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*. Association for Computational Linguistics, Atlanta, Georgia, pages 29–37. http://www.aclweb.org/anthology/W13-1605.

Stephanie Lukin and Marilyn Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*. Association for Computational Linguistics, Atlanta, Georgia, pages 30–40. http://www.aclweb.org/anthology/W13-1104.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

James M. Olver and Todd A. Mooradian. 2003. Personality traits and personal values: a conceptual and empirical integration. *Personality and Individual Differences* 35(1):109 – 125. https://doi.org/http://dx.doi.org/10.1016/S0191-8869(02)00145-9.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 213–223. http://www.aclweb.org/anthology/C14-1022.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, pages 97–106. http://dl.acm.org/citation.cfm?id=2685316.

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 *Empirical Methods on Natural Language Processing*. volume 13, pages 704–714. http://www.anthology.aclweb.org/D/D13/D13-1066.pdf.

SG Shamay-Tsoory, Rachel Tomer, and Judith Aharon-Peretz. 2005. The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology* 19(3):288. http://www.apa.org/pubs/journals/releases/neu-193288.pdf.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54. http://dx.doi.org/10.1177/0261927X09351676.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*.

A. Utsumi. 2000a. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics* http://www.utm.se.uec.ac.jp/ utsumi/paper/jop2000-utsumi.pdf.

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

Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol* 26(3):227–239. http://dx.doi.org/10.1080/10926488.2011.583197.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, pages 2449–2460. http://aclweb.org/anthology/C16-1231.