Proceedings of the Ninth International AAAI Conference on Web and Social Media

Contextualized Sarcasm Detection on Twitter

David Bamman and Noah A. Smith
School of Computer Science
Carnegie Mellon University
{dbamman,nasmith}@cs.cmu.edu

Abstract

Sarcasm requires some shared knowledge between speaker and audience; it is a profoundly *contextual* phenomenon. Most computational approaches to sarcasm detection, however, treat it as a purely linguistic matter, using information such as lexical cues and their corresponding sentiment as predictive features. We show that by including extra-linguistic information from the context of an utterance on Twitter – such as properties of the author, the audience and the immediate communicative environment – we are able to achieve gains in accuracy compared to purely linguistic features in the detection of this complex phenomenon, while also shedding light on features of interpersonal interaction that enable sarcasm in conversation.

Introduction

Most approaches to sarcasm detection to date have treated the task primarily as a text categorization problem, relying on the insights of Kreuz and Caucci (2007) that sarcastic utterances often contain lexical indicators (such as interjections and intensifiers) and other linguistic markers (such as nonveridicality and hyperbole) that signal their irony. These purely text-based approaches can be surprisingly accurate across different domains (Carvalho et al. 2009; Davidov, Tsur, and Rappoport 2010; González-Ibáñez, Muresan, and Wacholder 2011; Riloff et al. 2013; Lukin and Walker 2013; Reyes, Rosso, and Veale 2013), but are divorced from any notion of their potentially useful *context*. Yet context seems to matter. For example, humans require access to the surrounding context in which a Reddit post was written (such as the thread it appears in) in order to judge its tone (Wallace et al. 2014). On Twitter, modeling the relationship between a tweet and an author's past tweets can improve accuracy on this task (Rajadesingan, Zafarani, and Liu 2015).

This kind of contextual information is only one small part of the shared common ground that must be present between a speaker and their audience in order for sarcasm to be available for use between them. Kreuz (1996) calls this the "principle of inferability" – speakers only use sarcasm if they can be sure it will be understood by the audience – and finds in surveys that sarcasm is more likely to be used between two people who know each other well than between those who do not.

In all of these cases, the relationship between author and audience is central for understanding the sarcasm phenomenon. While the notion of an "audience" is relatively well defined for face-to-face conversations between two people, it becomes more complex when multiple people are present (Bell 1984), and especially so on social media, when a user's "audience" is often unknown, underspecified or "collapsed" (boyd 2008; Marwick and boyd 2011), making it difficult to fully establish the shared ground required for sarcasm to be detected, and understood, by its intended (or imagined) audience.

We present here a series of experiments to discern the effect of extra-linguistic information on the detection of sarcasm, reasoning about features derived not only from the local context of the message itself (as in past work), but also using information about the author, their relationship to their audience and the immediate communicative context they both share. Our main findings are:

- Including any aspect of the environment (features derived from the communicative context, the author, or the audience) leads to improvements in prediction accuracy.
- Users are more likely to tag their message with the explicit hashtag `#sarcasm` when they are *less* familiar with their audience. Following Kreuz (1996), we argue that this is a means of ensuring inferability in the face of uncertainty.

In the course of this work, we also present a rigorous analysis of new and previous features used in sarcasm detection so that future work in this area can choose to focus its efforts.

Data

Prior work on sarcasm detection on Twitter (González-Ibáñez, Muresan, and Wacholder 2011) found low agreement rates between human annotators at the task of judging the sarcasm of *others'* tweets; consequently, recent research exploits users' self-declarations of sarcasm in the form of `#sarcasm` or `#sarcastic` tags of their own tweets. This design choice does not capture the likely more common varieties of sarcasm expressed without an explicit hashtag, but does yield positive examples with high precision. Figure 1 gives one such example.

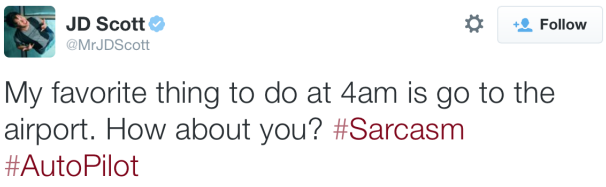

Figure 1: User self-reporting of sarcasm.

We follow the same methodology here as well, identifying the authors of all tweets mentioning `#sarcasm` or `#sarcastic` in the Gardenhose sample of tweets from August 2013–July 2014, and crawling up to the most recent 3,200 tweets of those authors. As in past work, we label a tweet as SARCASTIC if it contains the hashtag `#sarcasm` or `#sarcastic` as its final term, is written in English, is not a retweet, and contains at least three words. To explore the influence of the communicative context on our perception of sarcasm, we further subsample this set to include only tweets that are responses to another tweet. This yields a positive training set of 9,767 tweets; for negative data, we select an equal number of tweets from users over the same time period who have *not* mentioned `#sarcasm` or `#sarcastic` in their messages. The total dataset is evenly balanced at 19,534 tweets. Since the hashtags `#sarcasm` and `#sarcastic` are used to define the positive examples, we remove those tags from all tweets for the prediction task.

Experimental Setup

For the classification task of deciding whether a tweet is SARCASTIC or NOT SARCASTIC, we adopt binary logistic regression with ℓ_2 regularization using tenfold cross-validation, split on authors (so that tweets by the same author do not appear in multiple splits). We tune the ℓ_2 regularization parameter on development data (train on $^8/_{10}$, tune on $^1/_{10}$, test on the remaining held-out $^1/_{10}$) and repeat across ten folds. We perform this cross-validation and parameter tuning for *every* feature combination reported below, since different feature sets (with different cardinalities) will result in different optimal parameter settings.

Features

We can divide the features used in our models into four classes: those scoped only over the immediate tweet being predicted (§Tweet Features); those that reason over the author of that tweet, including historical data by that author (§Author Features); those that reason over the addressee of the tweet (the person to whom the target tweet under consideration is responding), including historical data for that individual and the author's history of interaction with them (§Audience Features); and features that consider the interaction between the tweet being predicted and the tweet that it is responding to (§Response Features).

The baseline accuracy, using only the majority class in each training fold, is 47.4% (this is lower than an even 50% since the folds are split by author and contain varying numbers of tweets). In describing each feature below, we also report in parentheses the tenfold cross-validated accuracy of a model trained *only* on that feature type.

Tweet Features

- **Word unigrams (72.4%) and bigrams (69.5%).** We create binary indicators of lowercased word unigrams and bigrams. The most indicative unigrams include *dare, shocked, clearly, #lol* and *gasp*, and the most indicative bigrams include *you mean*, *how dare*, *i'm shocked*, *i'm sure* and *at all*.
- **Brown cluster unigrams (72.0%) bigrams (69.1%).** For dimensionality reduction, we map each word in our vocabulary to one of 1000 non-overlapping clusters using the Brown clusters of Owoputi et al. (2013), which group words used in similar contexts into the same cluster. We compute unigrams and bigrams over terms in this reduced space.
- **Unlabeled dependency bigrams, lexicalized (70.3%) and Brown clusters (70.2%).** We create binary features from unlabeled dependency arcs between a.) two words and b.) their corresponding Brown clusters after parsing the tweet with TweeboParser (Kong et al. 2014).
- **Part of speech features (66.0%).** Past work has shown that part of speech information (such as the density of hashtags and emoticons) is among the most informative for this task. We apply the POS tagger of Owoputi et al. (2013) and include features based on the absolute count and ratio of each of the 25 tags, along with the "lexical density" of the tweet, which models the ratio of nouns, verbs, adjectives and adverbs to all words (Rajadesingan, Zafarani, and Liu 2015).
- **Pronunciation features (57.5%)** To model the use of Twitter-specific writing style (as in Rajadesingan et al., 2015), we include the number of words with only alphabetic characters but no vowels (e.g., *btw*) and the number of words with more than three syllables.
- **Capitalization features (57.5%).** We include the number of words with initial caps and all caps and the number of POS tags with at least initial caps.
- **Tweet whole sentiment (55.0%).** We include several types of tweet-level sentiment features. The first is a feature containing the numeric value of the entire tweet's sentiment as determined by the Stanford Sentiment Analyzer (Socher et al. 2013); since this phrase-based analyzer also determines the sentiment value of each non-terminal node in its syntactic parse tree, we also include the fraction of nonterminals with each sentiment score as a feature (which allows us to capture differences in sentiments across the tree).
- **Tweet word sentiment (53.7–54.7%).** As in much past work, we also include word-level sentiment features, modeling the maximum word sentiment score, minimum word sentiment score, and distance between the max and min. As in Rajadesingan et al. (2015), we use the dictionaries of Warriner et al. (2013) (54.7%) and the emotion scores of Thelwall et al. (2010) (53.7%).
- **Intensifiers (50.1%).** Since prior theoretical work has stressed the importance of hyperbole for sarcasm (Kreuz and Roberts 1995), we include a binary indicator for whether the tweet contains a word in a list of 50 intensifiers (*so, too, very,*

really) drawn from Wikipedia (http://en.wikipedia.org/wiki/Intensifier).

Author Features

- **Author historical salient terms (81.2%).** For each author, we identify the 100 terms in their historical tweets (excluding the test dataset) with the highest TF-IDF score and create binary indicators for each of those 100 terms. This is the single most informative feature of all those we evaluated.
- **Author historical topics (77.4%).** We create broader topic-based features by inferring a user's topic proportions under LDA (Blei, Ng, and Jordan 2003) with 100 topics over all tweets, where each document consists of up to 1,000 words of a user's tweets (excluding all messages in the test dataset). The topics most indicative of sarcasm include those relating to art and television shows.
- **Profile information (73.7%).** We create features for the author of the tweet drawn from their user profile information, including gender (as inferred by their first name, compared to trends in U.S. Social Security records), number of friends, followers and statuses, their duration on Twitter, the average number of posts per day, their timezone, and whether or not they are verified by Twitter (designating a kind of celebrity status). Being unverified, male, and from time zones in the United States are all strong markers of sarcasm.
- **Author historical sentiment (70.8%).** As in Rajadesingan et al. (2015), we model the distribution over sentiment in the user's historical tweets (excluding the test dataset), using the same word-level dictionaries applied to tweet-level sentiment described above. Users with historically negative sentiments have higher likelihoods of sarcasm.
- **Profile unigrams (66.2%).** We create binary indicators for all unigrams in the author's profile. The most indicative terms include *sarcasm, chemistry, #atheist* and *humor*.

Audience Features

- **Author historical topics (71.2%), Author historical salient terms (70.0%), Profile unigrams (68.6%), Profile information (66.3%).** As above, but for the author of the tweet to which the target tweet being predicted responds.
- **Author/Addressee interactional topics (73.9%).** To capture the similarity in interests between the author and addressee, we include features defined by the elementwise product of the author and addressee's historical topic distribution (resulting in a feature that it high if the two have both together tweeted about the same topics).
- **Historical communication between author and addressee (61.7%).** To model Kreuz's finding that sarcasm is more likely to take place between two people who are more familiar with each other, we include features that model that the degree of interaction between two users, including the number of previous messages sent from the author to the addressee, the rank of the addressee among the user's @-mention recipients and whether or not there have been at least one (and two) mutual @-messages exchanged between the author and the addressee (i.e., not simply unrequited messages sent from the author).

Environment Features

- **Pairwise Brown features between the original message and the response (71.7%).** To model the interaction between a target tweet and the tweet to which it is responding, we include binary indicators of pairwise Brown features between all terms in the two tweets.
- **Unigram features of the original message (68.8%).** To capture the original linguistic context a tweet is responding to, we include binary indicators of all unigrams in the original tweet as features. The most indicative terms in the original tweet include clear markers that already define a sarcastic environment, including *#sarcasm, sarcastic* and *sarcasm* as well as *worry, defense, advice, vote* and *kidding*.

Results

To compare performance across different features, we consider five feature combinations: those with access only to tweet-level information (defined in §TWEET FEATURES above); §TWEET FEATURES + §RESPONSE FEATURES; §TWEET FEATURES + §AUDIENCE FEATURES; §TWEET FEATURES + §AUTHOR FEATURES, and adding all features together in one model.

Figure 2 illustrates the relative gains in accuracy that result from including contextual information outside the immediate scope of the tweet being predicted: while tweet-only information yields an average accuracy of 75.4% across all ten folds, adding response features pushes this to 77.3%, audience features to 79.0% and author features to 84.9%. Including all features together yields the best performance at 85.1%, but most of these gains come simply from the addition of author information.

While the individual features above all report the accuracy of a model trained *only* on that feature, an ablation test (training the model on the full feature set excluding one feature) reveals that no feature is crucial for model performance: the most critical features are AUTHOR HISTORICAL SALIENT TERMS (–0.011), AUTHOR PROFILE FEATURES (–0.008), PAIRWISE BROWN FEATURES between the original message and the response (–0.008), PART OF SPEECH FEATURES (–0.002) and RESPONSE UNIGRAMS (–0.001). Training a model on these five features alone yields an accuracy of 84.3%, less than a point behind the full feature set.

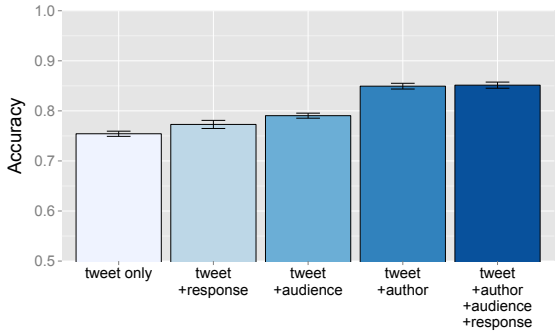

Figure 2: Accuracy across different feature sets, with 95% confidence intervals on the mean across 10 folds.

Analysis

While features derived from the author yield the greatest improvements in accuracy over the tweet alone, all feature classes (response, audience and author) display statistically significant improvements over the tweet-only features that ignore the communicative context. This confirms an effect on the interaction of the author and audience in the recognition of sarcasm, which can lead us to ask: who is this audience, and what about them is predictive of sarcasm across users? While Kreuz (1996) shows that sarcasm is primarily available between people who know each other well, we find that the strongest audience-based features that act as markers of sarcasm in this dataset are not those that suggest intimacy between the author and audience; the strongest audience predictors of sarcasm are the *absence* of mutual mentions (at least one mutual mention is a contraindicator, and at least two is more so); living in different time zones (i.e., not being geographically proximate) and features of celebrity (being verified and having many followers). In total, these features suggest that the `#sarcasm` hashtag is not a natural indicator of sarcasm expressed between friends, but rather serves an important communicative function of signaling the author's intent to an audience who may not otherwise be able to draw the correct inference about their message (as distinct from close friends who may be able to infer sarcasm without such labels). This has important consequences for the study of sarcasm and other speech acts on social media sites with complex audiences: in the absence of shared common ground required for their interpretation, explicit illocutionary markers are often necessary to communicate intent. Studying sarcasm that *does* rely on common ground (and does not require such explicit markers) will likely need to rely on other forms of supervision.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was made possible through the use of computing resources made available by the Open Science Data Cloud (OSDC), an Open Cloud Consortium (OCC)-sponsored project. The research was supported by NSF CAREER grant IIS-1054319 and by IARPA via DoI/NBC contract number D12PC00337. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the sponsors.

References

Bell, A. 1984. Language style as audience design. *Language in Society* 13:145–204.

Blei, D. M.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *JMLR* 3:993–1022.

boyd, d. 2008. *Taken Out of Context: American Teen Sociality in Networked Publics*. Ph.D. Dissertation, University of California-Berkeley, School of Information.

Carvalho, P.; Sarmento, L.; Silva, M. J.; and de Oliveira, E. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, 53–56.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *CoNLL*, 107–116.

González-Ibáñez, R.; Muresan, S.; and Wacholder, N. 2011. Identifying sarcasm in Twitter: A closer look. In *ACL*.

Kong, L.; Schneider, N.; Swayamdipta, S.; Bhatia, A.; Dyer, C.; and Smith, N. A. 2014. A dependency parser for tweets. In *EMNLP*, 1001–1012.

Kreuz, R. J., and Caucci, G. M. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, 1–4.

Kreuz, R. J., and Roberts, R. M. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbol* 10(1):21–31.

Kreuz, R. J. 1996. The use of verbal irony: Cues and constraints. In Mio, J. S., and Katz, A. N., eds., *Metaphor: Implications and Applications*, 23–38.

Lukin, S., and Walker, M. 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*.

Marwick, A. E., and boyd, d. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1):114–133.

Owoputi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*.

Rajadesingan, A.; Zafarani, R.; and Liu, H. 2015. Sarcasm detection on Twitter: A behavioral modeling approach. In *WSDM*.

Reyes, A.; Rosso, P.; and Veale, T. 2013. A multidimensional approach for detecting irony in Twitter. *Lang. Resour. Eval.* 47(1):239–268.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, 704–714.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; and Kappas, A. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12):2544–2558.

Wallace, C. B.; Choe, K. D.; Kertz, L.; and Charniak, E. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *ACL*, 512–516.

Warriner, A.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4):1191–1207.