Detecting Sarcasm is Extremely Easy ;-)

Natalie Parde and **Rodney D. Nielsen**
Human Intelligence and Language Technologies (HiLT) Laboratory
Department of Computer Science and Engineering
University of North Texas
{natalie.parde, rodney.nielsen}@unt.edu

Abstract

Detecting sarcasm in text is a particularly challenging problem in computational semantics, and its solution may vary across different types of text. We analyze the performance of a domain-general sarcasm detection system on datasets from two very different domains: Twitter, and Amazon product reviews. We categorize the errors that we identify with each, and make recommendations for addressing these issues in NLP systems in the future.

1 Introduction

Sarcasm detection is a tricky problem, even for humans. The definition of sarcasm is hazy, sarcasm can be heavily context-dependent, and it is often marked more by prosodic cues than syntactic characteristics, all of which make its computational detection particularly complex. Nonetheless, some researchers have achieved success in predicting whether or not instances of text contain sarcasm based on domain-specific features (Maynard and Greenwood, 2014; Rajadesingan et al., 2015), sentiment (Riloff et al., 2013), text patterns (Davidov et al., 2010), and other semantic features (Ghosh et al., 2015; Amir et al., 2016).

Since most prior work in this area has been domain-specific, the findings resulting from these models may not be broadly applicable. For example, Twitter, a popular domain for sarcasm researchers, constrains posts to 140 (or as of very recently, 280) characters; this means that the type of sarcasm found in tweets may be quite different from that found in a domain that allows lengthy posts, such as Amazon product reviews. Previously, we explored this phenomenon by experimenting with various models to identify an approach better capable of learning domain-general sarcasm detection (Parde and Nielsen, 2017) . In this paper, we build upon that work by conducting a performance analysis of our best-performing approach on two different text domains, and identifying common types of errors made by the system. We follow this with recommendations for improvement in future sarcasm detection systems.

2 Background

Research on automatic sarcasm detection to date has taken place on a variety of domains, including news articles (Burfoot and Baldwin, 2009), web forums (Justo et al., 2014), product reviews (Buschmeier et al., 2014), and tweets (Maynard and Greenwood, 2014; Rajadesingan et al., 2015; Liebrecht et al., 2013; Riloff et al., 2013; Bamman and Smith, 2015; González-Ibáñez et al., 2011; Reyes et al., 2013; Ghosh et al., 2015; Amir et al., 2016). The last of these, Twitter-based sarcasm detection, has dominated the research arena.

Twitter is a popular domain choice for sarcasm researchers because tweets are readily-available and may be freely downloaded, and moreover many tweets are self-labeled by Twitter users for various attributes using *hashtags*, or keywords prefaced with the "#" symbol. However, tweets are not necessarily representative of text in general. Their strict length requirement causes users to adopt sometimes-confusing acronyms and shorthand spellings. Hashtags often consist of smashed-together words without any token markers, and may convey critical content not otherwise detectable in the tweet text. Finally, tweets may refer to external context that renders them confusing to later readers. For example, tweeting "Great." minutes after an election is called may be easily understandable to readers at that moment, but ambiguous to readers who see the tweet several days later, and much too vague for today's computational sarcasm detector to decipher.

Researchers who have focused on detecting sarcasm in tweets have taken several approaches.

Maynard and Greenwood (2014) learned hashtags that commonly correspond with sarcastic tweets, and checked for those in subsequent tweets to determine whether or not the tweets were sarcastic. Other researchers utilized Twitter histories, developing behavioral models of sarcasm usage specific to individual users (Rajadesingan et al., 2015), or features based on the users, their audiences, and the author-audience relationship of the tweet in question (Bamman and Smith, 2015). Some researchers considered the sentiment (Riloff et al., 2013) or emotional scenario (Reyes et al., 2013) of a tweet when deciding whether or not it contained sarcasm, and finally others experimented with n-grams (Liebrecht et al., 2013) and word embeddings (Ghosh et al., 2015; Ghosh and Veale, 2016; Amir et al., 2016).

Amazon product reviews, which have also interested sarcasm researchers, differ from tweets in several key ways: they are of variable (and often much longer) length, they do not utilize hashtags, and they generally contain more context. The primary domain-specific feature employed by sarcasm detection researchers using Amazon product reviews has been a product's "star rating" (the number of stars assigned to the product by the review writer) (Buschmeier et al., 2014; Parde and Nielsen, 2017). Other characteristics that researchers have considered in this domain include syntactic features (Buschmeier et al., 2014; Davidov et al., 2010) and the presence of interjections or laughter terms (Buschmeier et al., 2014).

Finally, we learned a general sarcasm detection model from many tweets and fewer Amazon product reviews (Parde and Nielsen, 2017). We found that by applying a domain adaptation step prior to training the model, we were able to achieve higher performance in predicting sarcasm in Amazon product reviews over models that trained on reviews alone or on a simple combination of reviews and tweets. Our prior work was notable in that it was the first approach that specifically sought domain-generality. We analyze its performance on different datasets in this work.

3 Sarcasm Detection Methods

We train our sarcasm detection approach on the same training data used in our previous work (3998 tweets and 1003 Amazon product reviews), and apply it to two test datasets: AMAZON, a 251-instance set of sarcastic (87) and non-

sarcastic (164) Amazon product reviews originally collected by Filatova (2012), and TWITTER, a 1000-instance set of sarcastic (391) and non-sarcastic (609) tweets containing the hashtags *#sarcasm* (the sarcastic class) or *#happiness*, *#sadness*, *#anger*, *#surprise*, *#fear*, or *#disgust* (the negative class).[1] The approach utilizes features that seek to convey informative characteristics from the domains considered as well as general characteristics expected to remain indicative of sarcasm across many domains. We briefly describe each in Table 1; for additional information, the reader is referred to our earlier paper.

3.1 Classification Algorithm

All features were extracted from each instance, regardless of its domain (feature values were left empty when it was impossible to fill them, e.g., star rating for tweets). Then, the feature space was transformed using the domain adaptation approach originally outlined by Daumé III (2007). Daumé's approach works by modifying the feature space such that it contains three mappings of the original features: a source version, a target version, and a general version. More formally, letting $\breve{\mathcal{X}} = \mathbb{R}^{3F}$ be the augmented version of a feature space $\mathcal{X} = \mathbb{R}^F$, and $\Phi^s, \Phi^t : \mathcal{X} \to \breve{\mathcal{X}}$ be mappings for the source and target data, respectively,

$$\Phi^s(\mathbf{x}) = \langle \mathbf{x}, \mathbf{0}, \mathbf{x} \rangle, \quad \Phi^t(\mathbf{x}) = \langle \mathbf{0}, \mathbf{x}, \mathbf{x} \rangle \quad (1)$$

where $\mathbf{0} = \langle 0, 0, ..., 0 \rangle \in \mathbb{R}^F$ is the zero vector. It is then left to the classification algorithm to decide how to best take advantage of this supplemental information. We use Naïve Bayes, following our earlier work.

4 Model Performance

We compute precision (P), recall (R), and f-measure (F_1) on the positive (sarcastic) class for both TWITTER and AMAZON, and report results relative to the performance of other systems on the same data (Table 2). Our results on AMAZON are identical to those reported originally (Parde and Nielsen, 2017). Our previous paper reported results on TWITTER when training only on Twitter data; here we instead apply the same model as applied to AMAZON and achieve slightly higher results. Thus, the approach outperforms other sar-

---

[1]These hashtags were removed prior to using the data.

Feature Type	Description
CONTAINS TWITTER INDICATOR	Multiple binary features indicating whether the instance contains one of the sarcasm-related hashtags, emoticons, and/or indicator phrases learned by Maynard and Greenwood (2014).
TWITTER-BASED PREDICATES AND SITUATIONS	Multiple binary features indicating whether the instance contains a positive predicate, positive sentiment, and/or negative situation phrase learned by Riloff et al. (2013) from a corpus of tweets. Includes an additional binary feature that indicates whether one of those positive predicates or sentiments precedes one of those negative situation phrases by ≤ 5 tokens.
STAR RATING	The number of stars (1-5) associated with the review.
LAUGHTER AND INTERJECTIONS	Multiple binary features indicating whether the instance contains: *hahaha*, *haha*, *hehehe*, *hehe*, *jajaja*, *jaja*, *lol*, *lmao*, *rofl*, *wow*, *ugh*, and/or *huh*.
SPECIFIC CHARACTERS	Multiple binary features indicating whether the instance contains an ellipsis, an exclamation mark, and/or a question mark.
POLARITY	Multiple features indicating the most polar (positive or negative) unigram in the instance, the polarity score (-5 to +5) associated with that unigram, the average polarity of the instance, the overall (sum) polarity for the instance, the largest difference in polarity between any two words in the instance, and the percentages of positive and negative words in the instance.
SUBJECTIVITY	The percentages of strongly subjective positive words, strongly subjective negative words, weakly subjective positive words, and weakly subjective negative words in the instance.
PMI	Multiple features indicating the pointwise mutual information (PMI) between the most polar unigram and the 1, 2, 3, and 4 words that immediately follow it.
CONSECUTIVE CHARACTERS	Multiple features indicating the highest number of consecutive repeated characters in the instance (e.g., "Sooooo" $\Rightarrow 5$) and the highest number of consecutive punctuation characters in the instance.
ALL-CAPS	Multiple features indicating the number and percentage of all-caps words in the instance.
BAG OF WORDS	Two types of bag-of-words features: one in which the words included in the "bag" are those most closely associated with four groups of training instances (Sarcastic $\times$ Non-Sarcastic) $\times$ (Amazon $\times$ Twitter), and one in which the words in the "bag" were the most common words in those groups (any duplicates across groups were removed).

Table 1: Features included in the sarcasm detection system.

		P	R	F_1
TWITTER	Parde and Nielsen (2017)	0.55	0.62	0.58
	Our Results	0.53	**0.68**	**0.59**
AMAZON	Buschmeier et al. (2014)	0.82	0.69	0.74
	Our Results	0.75	**0.82**	**0.78**

Table 2: Performance of our sarcasm detection model relative to prior work on the same datasets.

	Amazon	Twitter
Predicted Sarcastic	24	235
Predicted Non-Sarcastic	16	127

Table 3: Errors included in the analysis.

casm detection methods on both AMAZON and TWITTER.

5 Error Analysis

5.1 Methodology

We conduct our error analysis on all misclassified instances (402 total) in both AMAZON and TWITTER. The errors were distributed as shown in Table 3. For both datasets, there were more false positives (instances predicted to be sarcastic when they really weren't) than false negatives.

We analyzed each misclassified instance, making notes regarding characteristics that may have led to the misclassification. We then compiled these notes into more general error categories, identified (with examples from our data) in Tables 4 and 5. Some instances were assigned to multiple error categories.

5.2 Results

There were several leading trends in the misclassifications. Among false negatives in both datasets, in many cases the sarcasm expressed could only be inferred using world knowledge (an example tweet from this category, noted in Table 4, is *When my 10 yr old niece texts me to let me know she is taller than me. #thanks #sarcasm #hateyoubut-loveyou*). Within tweets specifically, some (23) did not convey sarcasm once the sarcasm hashtag was removed. Some (8) also contained sarcastic content only in other hashtags associated with the tweet. Other tweets (13) were found upon manual inspection to not be sarcastic, despite containing the sarcasm hashtag; instead, these tweets discussed sarcasm in some way.

Nine false negatives contained words typically associated with sarcasm; developing better ways of identifying these words could eliminate such errors. For product reviews, a common trait of misclassified instances was that they developed sarcastic stories about the product (for instance, one review describes the magical qualities of a pair of

Type	Amazon	Twitter	Example
Requires World Knowledge	7	63	*When my 10 yr old niece texts me to let me know she is taller than me. #thanks #sarcasm #hateyoubutloveyou*
Formatted as Story	4	0	*I thought that I had shrunk everything in the dryer, so I gave up and put on these new Hanes cushioned crews. Immediately, I felt a sense of joy. It was strange, like the first time you kiss someone, or how you felt as a kid when waking up on Christmas morning. I kept wondering if it was the socks that made me feel that way, or if I was just somehow subconsciously triggered to reminisce. That day, in it's entirety, was wonderful. At least 50 people I had never spoken to before somehow knew my name. These were people on the street, even. At the coffee shop, the girl who normally had the demeanor of a disgruntled, middle-age cafeteria worker actually gave me a free coffee and tried to flirt with me. Not just to flirt, but a stumbly sort of flirting that only comes about when desire has made you lose your grasp of language structure. At the university, I was excused from an upcoming midterm for a reason I don't even remember. I think it involved being "an attentive enough listener at lectures."*
Positive Sentiment + Negative Situation	0	9	*#HappyBirthdayTwitter Thanks for providing a platform where people can troll and abuse each other #sarcasm*
Negative Sentiment + Positive Situation	0	3	*No one's awake at home, should've gone to the gym. Life is tough doing nothing all day #messyhouse #nodinner #sarcasm*
Highly Negative	1	3	*Don't waste your money on this convoluted and unfriendly piece of over-priced junk. ... If you find out too slowly how lousy this item is, you are stuck with it. And don't give it as a gift at Xmas - your recipients can't return it either. You have given them an expensive paperweight unless all the stars are in alignment for them, andthen they'll probably find it useless anyway.*
Many All-Caps Words	0	6	*My brain at 3am = ALWAYS A GREAT TIME. #sarcasm*
Requires #sarcasm	0	23	*Some people know how to really make you feel valued #sarcasm*
Sarcasm in Hashtags	0	8	*Oh hi LA! Long time no see! #sarcasm #yesterday #IneedALLTHENAPS*
Contains Sarcastic Word or Phrase	0	9	*Not jealous at all of anyone who could afford a pair of the #Irregular-Choice #AliceInWonderland shoes today. Ohh no, not at all. #sarcasm*
Mostly Non-Sarcastic with Some Sarcastic Phrases	4	1	*I drive a Toyota Sienna minivan with JBL stuff on my speakers. Apparently that was important. Now it works great. Reception in Houston has been great. It plays through the line-in Aux port great (I use it with my ipod and creative zen) and USB keys work. I'm not sure it ever shows the file names it's playing off the USB, which is weird but not worth $100 to upgrade to a better stereo. So, it works but had quite a bit of fiddling to make it go. It's great for the $. I have fairly low standards...I only listen to audiobooks, podcasts, NPR, etc. So I have no idea what the audiophiles would think. (and, for the snarky, YES, there was a sale on the word "great" today.)*
Non-Sarcastic	1	13	*I was being sarcastic with that tweet by the way incase people thought I was serious.... #sarcasm*

Table 4: Errors: Instances incorrectly predicted as non-sarcastic.

socks at length); in such stories there tend to be particularly few linguistic indicators of sarcasm.

False positives were typified by different characteristics. Many tweets (109) in this category included excessive punctuation, a trait commonly associated with sarcastic text. Other instances (29 tweets and 5 product reviews) contained a mix of positive and negative sentiment, which the model mistook for sarcasm. Some misclassified instances contained many technical or "niche" words, for which few of the polarity-based features could have been computed, and others included ambiguous phrases often found in sarcastic text (e.g., **Jeez**, *how am I supposed to react to meeting someone who identifies her spirit animal as Claire Underwood? #HouseOfCards #Fear*).

Some tweets contained misspellings that may have confused the model, and some product reviews were non-sarcastic reviews of "silly" products. In the case of these latter reviews, the model may have simply learned to mark any reviews associated with those products as sarcastic. Finally, upon manual inspection we found that four of the Amazon product reviews marked as non-sarcastic actually contained at least some sarcastic text, and 27 of the tweets that did not contain the sarcasm hashtag were in fact sarcastic.

5.3 Recommendations

Based on our analysis, we recommend that the following factors be taken into account in future systems. Beyond their anticipated direct bene-

Type	Amazon	Twitter	Example
Odd Product/Product that Seems Sarcastic	5	0	*I haven't had the chance to use it yet as the whip is broken. I'm hoping I can either get a replacement whip or just get my money back.*
Mix of Positive and Negative Sentiment	5	29	*Good morning. Coffee. Portfolio. Torment. School. #school #sadness*
Very Negative	6	5	*This book is so terrible that I couldn't even make it past the first 1/4 of it - the characters were horrible, shallow people, and the plot is so see-through. Clearly, this book is one of Sophie's earlier works - the "plot" is terrible. Don't waste your money - don't take a chance in case you crack the spine - you won't be able to return it!*
Very Positive	0	9	*Be happy. Not because everything is good, but because u can see the good side of everything #happiness*
Ambiguous Phrases	2	26	*Jeez, how am I supposed to react to meeting someone who identifies her spirit animal as Claire Underwood? #HouseOfCards #Fear*
Contains Technical Terminology	3	31	*But it's not much louder than the two-stage oillube compressor it replaced. I needed something that could be moved in a pinch, something that could run off 110V 15A service I have in the garage, and something with enough capacity to run my air ratchets, cut off tools, etc.*
Lots of Punctuation	3	109	*#SongToday WORK by @rihanna heavyyyyyyyyy!!!!! #fancy #happiness*
Short	3	11	*Oh exams coming up #sadness*
Many All-Caps Words	1	14	*Episode 42 of @TTGpodcast is outstanding! I was like "yeah good qustion Rocket–OMG THAT WAS ME I ASKED THAT!!" #surprise*
Contains Misspellings	0	13	*Thank u Spring for this beautiful snow #Spring #snow #Surprise*
Sarcastic	4	27	*I'm truly thrilled to find out which of my bodily fluids will start leaking next. Is there a bingo card for the third trimester? #surprise*

Table 5: Errors: Instances incorrectly predicted as sarcastic.

fits, adopting these recommendations should decrease reliance on syntactic features (e.g., excessive punctuation and all-caps words).

World Knowledge: For many false negatives, the sarcasm expressed was detectable only through knowledge of the world. Frame-semantic resources could be used to detect some sarcasm instantiated through script-based inconsistencies. Furthermore, features could be derived from commonsense knowledge bases such as that of the Never-Ending Language Learner (Mitchell et al., 2015) to better detect contradictory expressions.

Text Normalization: When detecting sarcasm in user-generated content (e.g., Twitter), word splitting algorithms should be applied in the future to disambiguate compound hashtags into their constituent words, and spelling correction algorithms can be applied to normalize text. The latter should be done with caution, as in some cases, spelling normalization may not be desirable—for instance, "sooooo" may convey something different from "so," while "mihgt" likely conveys the same information as "might."

Enhanced Lexicon of Sentiment and Situation Phrases: Some of the errors we identified could have been easily addressed had the system understood that they described negative situations in positive terms, or vice versa. We attempted to capture this phenomenon by employing features based on the work of Riloff et al. (2013). However, we found that the phrases identified by Riloff et al. were virtually non-existent in our Twitter dataset. To properly employ these types of features, new events and sentiment phrases should be continually mined from Twitter to account for evolving linguistic patterns and trends in public opinion.

6 Conclusion

In this work, we analyze the performance of a domain-general sarcasm detection approach on two datasets: TWITTER and AMAZON. We verify that the approach outperforms others on the same data, and conduct an analysis of the misclassified instances to identify common error types. Finally, we make recommendations for addressing these errors. It is our hope that these insights will enable researchers to build high-performing sarcasm detection systems suited to many text domains.

Acknowledgements

References

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 *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 574–577.

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

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*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal. Association for Computational Linguistics.

Roberto González-Ibáñez, 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*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ins Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124 – 133.

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*, pages 29–37, Atlanta, Georgia. Association for Computational Linguistics.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland. European Language Resources Association.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Justin Betteridge, Andrew Carlson, Bhavana D. Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapa Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 2302–2310.

Natalie Parde and Rodney Nielsen. 2017. #sarcasmdetection is soooo general! towards a domain-independent approach for detecting sarcasm. In *Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference*, pages 276–281.

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*, pages 97–106. ACM.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268.

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 *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.