Identifying Sarcasm in Twitter: A Closer Look

Roberto González-Ibáñez **Smaranda Muresan** **Nina Wacholder**

School of Communication & Information
Rutgers, The State University of New Jersey
4 Huntington St, New Brunswick, NJ 08901
{rgonzal, smuresan, ninwac}@rutgers.edu

Abstract

Sarcasm transforms the polarity of an apparently positive or negative utterance into its opposite. We report on a method for constructing a corpus of sarcastic Twitter messages in which determination of the sarcasm of each message has been made by its author. We use this reliable corpus to compare *sarcastic* utterances in Twitter to utterances that express *positive* or *negative* attitudes without sarcasm. We investigate the impact of lexical and pragmatic factors on machine learning effectiveness for identifying sarcastic utterances and we compare the performance of machine learning techniques and human judges on this task. Perhaps unsurprisingly, neither the human judges nor the machine learning techniques perform very well.

1 Introduction

Automatic detection of sarcasm is still in its infancy. One reason for the lack of computational models has been the absence of accurately-labeled naturally occurring utterances that can be used to train machine learning systems. Microblogging platforms such as Twitter, which allow users to communicate feelings, opinions and ideas in short messages and to assign labels to their own messages, have been recently exploited in sentiment and opinion analysis (Pak and Paroubek, 2010; Davidov et al., 2010). In Twitter, messages can be annotated with hashtags such as #bicycling, #happy and #sarcasm. We use these hashtags to build a labeled corpus of naturally occurring sarcastic, positive and negative tweets.

In this paper, we report on an empirical study on the use of lexical and pragmatic factors to distinguish *sarcasm* from *positive* and *negative* sentiments expressed in Twitter messages. The contributions of this paper include i) creation of a corpus that includes only sarcastic utterances that have been explicitly identified as such by the composer of the message; ii) a report on the difficulty of distinguishing *sarcastic* tweets from tweets that are straight-forwardly *positive* or *negative*. Our results suggest that lexical features alone are not sufficient for identifying sarcasm and that pragmatic and contextual features merit further study.

2 Related Work

Sarcasm and irony are well-studied phenomena in linguistics, psychology and cognitive science (Gibbs, 1986; Gibbs and Colston 2007; Kreuz and Glucksberg, 1989; Utsumi, 2002). But in the text mining literature, automatic detection of sarcasm is considered a difficult problem (Nigam & Hurst, 2006 and Pang & Lee, 2008 for an overview) and has been addressed in only a few studies. In the context of spoken dialogues, automatic detection of sarcasm has relied primarily on speech-related cues such as laughter and prosody (Tepperman et al., 2006). The work most closely related to ours is that of Davidov et al. (2010), whose objective was to identify sarcastic and non-sarcastic utterances in Twitter and in Amazon product reviews. In this paper, we consider the somewhat harder problem

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 581–586,
Portland, Oregon, June 19-24, 2011.

of distinguishing sarcastic tweets from non-sarcastic tweets that directly convey positive and negative attitudes (we do not consider neutral utterances at all).

Our approach of looking at lexical features for identification of sarcasm was inspired by the work of Kreuz and Caucci (2007). In addition, we also look at pragmatic features, such as establishing common ground between speaker and hearer (Clark and Gerring, 1984), and emoticons.

3 Data

In Twitter, people (tweeters) post messages of up to 140 characters (tweets). Apart from plain text, a tweet can contain references to other users (@<user>), URLs, and hashtags (#hashtag) which are tags assigned by the user to identify topic (#teaparty, #worldcup) or sentiment (#angry, #happy, #sarcasm). An example of a tweet is: *"@UserName1 check out the twitter feed on @UserName2 for a few ideas :) http://xxxxxx.com #happy #hour"*.

To build our corpus of sarcastic (S), positive (P) and negative (N) tweets, we relied on the annotations that tweeters assign to their own tweets using hashtags. Our assumption is that the best judge of whether a tweet is intended to be sarcastic is the author of the tweet. As shown in the following sections, human judges other than the tweets' authors, achieve low levels of accuracy when trying to classify sarcastic tweets; we therefore argue that using the tweets labeled by their authors using hashtag produces a better quality gold standard. We used a Twitter API to collect tweets that include hashtags that express sarcasm (#sarcasm, #sarcastic), direct positive sentiment (e.g., #happy, #joy, #lucky), and direct negative sentiment (e.g., #sadness, #angry, #frustrated), respectively. We applied automatic filtering to remove retweets, duplicates, quotes, spam, tweets written in languages other than English, and tweets with URLs.

To address the concern of Davidov et al. (2010) that tweets with #hashtags are noisy, we automatically filtered all tweets where the hashtags of interest were not located at the very end of the message. We then performed a manual review of the filtered tweets to double check that the remaining end hashtags were not part of the message. We thus eliminated messages *about* sarcasm such as "I really love #sarcasm" and kept only messages that *express* sarcasm, such as "lol thanks. I can always count on you for comfort :) #sarcasm".

Our final corpus consists of 900 tweets in each of the three categories, sarcastic, positive and negative. Examples of tweets in our corpus that are labeled with the #sarcasm hashtag include the following:

1) *@UserName That must suck.*
2) *I can't express how much I love shopping on black Friday.*
3) *@UserName that's what I love about Miami. Attention to detail in preserving historic landmarks of the past.*
4) *@UserName im just loving the positive vibes out of that!*

The sarcastic tweets are primarily negative (i.e., messages that sound positive but are intended to convey a negative attitude) as in Examples 2-4, but there are also some positive messages (messages that sound negative but are apparently intended to be understood as positive), as in Example 1.

4 Lexical and Pragmatic Features

In this section we address the question of whether it is possible to empirically identify lexical and pragmatic factors that distinguish sarcastic, positive and negative utterances.

Lexical Factors. We used two kinds of lexical features – unigrams and dictionary-based. The dictionary-based features were derived from i) Pennebaker et al.'s LIWC (2007) dictionary, which consists of a set of 64 word categories grouped into four general classes: Linguistic Processes (LP) (e.g., adverbs, pronouns), Psychological Processes (PP) (e.g., positive and negative emotions), Personal Concerns (PC) (e.g, work, achievement), and Spoken Categories (SC) (e.g., assent, non-fluencies); ii) WordNet Affect (WNA) (Strapparava and Valitutti, 2004); and iii) list of interjections (e.g., ah, oh, yeah)[1], and punctuations (e.g., !, ?). The latter are inspired by results from Kreuz and Caucci (2007). We merged all of the lists into a single dictionary. The token overlap between the words in combined dictionary and the words in the tweets was 85%. This demonstrates that lexical coverage is good, even though tweets are well

[1] *http://www.vidarholen.net/contents/interjections/*

known to contain many words that do not appear in standard dictionaries.

Pragmatic Factors. We used three pragmatic features: i) positive emoticons such as smileys; ii) negative emoticons such as frowning faces; and iii) *ToUser*, which marks if a tweets is a reply to another tweet (signaled by <@user>).

Feature Ranking. To measure the impact of features on discriminating among the three categories, we used two standard measures: presence and frequency of the factors in each tweet. We did a 3-way comparison of Sarcastic (S), Positive (P), and Negative (N) messages (S-P-N); as well as 2-way comparisons of i) Sarcastic and Non-Sarcastic (S-NS); ii) Sarcastic and Positive (S-P) and Sarcastic and Negative (S-N). The NS tweets were obtained by merging 450 randomly selected positive and 450 negative tweets from our corpus.

We ran a χ^2 test to identify the features that were most useful in discriminating categories. Table 1 shows the top 10 features based on *presence* of all dictionary-based lexical factors plus the pragmatic factors. We refer to this set of features as LIWC$^+$.

S-P-N	S-NS	S-N	S-P
Negemo(PP)	Posemo(PP)	Posemo(PP)	Question
Posemo(PP)	Present(LP)	Negemo(PP)	Present(LP)
Smiley(Pr)	Question	Joy(WNA)	ToUser(Pr)
Question	ToUser(Pr)	Affect(PP)	Smiley(Pr)
Negate(LP)	Affect(PP)	Anger(PP)	AuxVb(LP)
Anger(PP)	Verbs(LP)	Sad(PP)	Ipron(LP)
Present(LP)	AuxVb(LP)	Swear(PP)	Negate(LP)
Joy(WNA)	Quotation	Smiley(Pr)	Verbs(LP)
Swear(PP)	Social(PP)	Body(PP)	Time(PP)
AuxVb(LP)	Ingest(PP)	Frown(Pr)	Negemo(PP)

Table 1: 10 most discriminating features in LIWC$^+$ for each task

In all of the tasks, negative emotion (*Negemo*), positive emotion (*Posemo*), negation (*Negate*), emoticons (*Smiley*, *Frown*), auxiliary verbs (*AuxVb*), and punctuation marks are in the top 10 features. We also observe indications of a possible dependence among factors that could differentiate sarcasm from both positive and negative tweets: sarcastic tweets tend to have positive emotion words like positive tweets do (*Posemo* is a significant feature in S-N but not in S-P), while they use more negation words like negative tweets do (*Negate* is an important feature for S-P). Table 1 also shows that the pragmatic factor *ToUser* is important in sarcasm detection. This is an indication of the possible importance of features that indicate *common ground* in sarcasm identification.

5 Classification Experiments

In this section we investigate the usefulness of lexical and pragmatic features in machine learning to classify sarcastic, positive and negative Tweets.

We used two standard classifiers often employed in sentiment classification: support vector machine with sequential minimal optimization (SMO) and logistic regression (LogR). For features we used: 1) unigrams; 2) *presence* of dictionary-based lexical and pragmatic factors (LIWC$^+$_P); and 3) *frequency* of dictionary-based lexical and pragmatic factors (LIWC$^+$_F). We also trained our models with bigrams and trigrams; however, results using these features did not report better results than unigrams and LICW$^+$. The classifiers were trained on balanced datasets (900 instances per class) and tested through five-fold cross-validation.

In Table 2, shaded cells indicate the best accuracies for each class, while bolded values indicate the best accuracies per row. In the three-way classification (S-P-N), SMO with unigrams as features outperformed SMO with LIWC$^+$_P and LIWC$^+$_F as features. Overall SMO outperformed LogR. The best accuracy of 57% is an indication of the difficulty of the task.

Class	Features	SMO	LogR
S-P-N	*Unigrams*	***57.22***	49.00
	LIWC$^+$_F	**55.59**	55.56
	LIWC$^+$_P	**55.67**	55.59
S-NS	*Unigrams*	***65.44***	60.72
	LIWC$^+$_F	**61.22**	59.83
	LIWC$^+$_P	62.78	**63.17**
S-P	*Unigrams*	***70.94***	64.83
	LIWC$^+$_F	66.39	**67.44**
	LIWC$^+$_P	67.22	**67.83**
S-N	*Unigrams*	***69.17***	64.61
	LIWC$^+$_F	**68.56**	67.83
	LIWC$^+$_P	68.33	**68.67**
P-N	*Unigrams*	**74.67**	72.39
	LIWC$^+$_F	74.94	***75.89***
	LIWC$^+$_P	**75.78**	**75.78**

Table 2: Classifiers accuracies using 5-fold cross-validation, in percent.

We also performed several two-way classification experiments. For the S-NS classification the best results were again obtained using SMO with

Task	S – N – P (10% dataset)			S – NS (10% dataset)		S – NS (100 tweets + emoticons)	
HBI	[43.33%-62.59%]			[59.44% - 66.85%]		[70% - 73%]	
Test	Features	SMO	LogR	SMO	LogR	SMO	LogR
1	*Unigrams*	**55.92**	46.66	***68.33***	57.78	***71.00***	66.00
2	LIWC$^+$_F	54.07	**54.81**	**62.78**	61.11	**60.00**	58.00
3	LIWC$^+$_P	***57.41***	57.04	**67.78**	67.22	51.00	**53.00**

Table 3: Classifiers accuracies against humans' accuracies in three classification tasks.

unigrams as features (65.44%). For S-P and S-N the best accuracies were close to 70%. Overall, our best result (75.89%) was achieved in the polarity-based classification P-N. It is intriguing that the machine learning systems have roughly equal difficulty in separating sarcastic tweets from positive tweets and from negative tweets.

These results indicate that the lexical and pragmatic features considered in this paper do not provide sufficient information to accurately differentiate sarcastic from positive and negative tweets. This may be due to the inherent difficulty of distinguishing short utterances in isolation, without use of contextual evidence.

In the next section we explore the inherent difficulty of identifying sarcastic utterances by comparing human performance and classifier performance.

6 Comparison against Human Performance

To get a better sense of how difficult the task of sarcasm identification really is, we conducted three studies with human judges (not the authors of this paper). In the first study, we asked three judges to classify 10% of our S-P-N dataset (90 randomly selected tweets per category) into sarcastic, positive and negative. In addition, they were able to indicate if they were unsure to which category tweets belonged and to add comments about the difficulty of the task.

In this study, overall agreement of 50% was achieved among the three judges, with a Fleiss' Kappa value of 0.4788 ($p<.05$). The mean accuracy was 62.59% (7.7) with 13.58% (13.44) uncertainty. When we considered only the 135 of 270 tweets on which all three judges agreed, the accuracy, computed over to the entire gold standard test set, fell to 43.33%[2]. We used the accuracy when the judges agree (43.33%) and the average accuracy (62.59%) as a human baseline interval (HBI).

We trained our SMO and LogR classifiers on the other 90% of the S-P-N. The models were then evaluated on 10% of the S-P-N dataset that was also labeled by humans. Classification accuracy was similar to results obtained in the previous section. Our best result -- an accuracy of 57.41%-- was achieved using SMO and LIWC$^+$_P (Table 3: S-P-N). The highest value in the established HBI achieved a slightly higher accuracy; however, when compared to the bottom value of the same interval, our best result significantly outperformed it. It is intriguing that the difficulty of distinguishing sarcastic utterances from positive ones and from negative ones was quite similar.

In the second study, we investigated how well human judges performed on the two-way classification task of labeling sarcastic and non-sarcastic tweets. We asked three other judges to classify 10% of our S-NS dataset (i.e, 180 tweets) into sarcastic and non-sarcastic. Results showed an agreement of 71.67% among the three judges with a Fleiss' Kappa value of 0.5861 ($p<.05$). The average accuracy rate was 66.85% (3.9) with 0.37% uncertainty (0.64). When we considered only cases where all three judges agreed, the accuracy, again computed over the entire gold standard test set, fell to 59.44%[3]. As shown in Table 3 (S-NS: 10% tweets), the HBI was outperformed by the automatic classification using unigrams (68.33%) and LIWC$^+$_P (67.78%) as features.

Based on recent results which show that non-linguistic cues such as emoticons are helpful in interpreting non-literal meaning such as sarcasm and irony in user generated content (Derks et al., 2008; Carvalho et al., 2009), we explored how much emoticons help humans to distinguish sarcastic from positive and negative tweets. For this test, we created a new dataset using only tweets with emoticons. This dataset consisted of 50 sarcastic

[2] The accuracy on the set they agreed on (135 out of 270 tweets) was 86.67%.

[3] The accuracy on the set they agreed on (129 out of 180 tweets) was 82.95%.

tweets and 50 non-sarcastic tweets (25 P and 25 N). Two human judges classified the tweets using the same procedure as above. For this task judges achieved an overall agreement of 89% with Cohen's Kappa value of 0.74 ($p<.001$). The results show that emoticons play an important role in helping people distinguish sarcastic from non-sarcastic tweets. The overall accuracy for both judges was 73% (1.41) with uncertainty of 10% (1.4). When all judges agreed, the accuracy was 70% when computed relative the entire gold standard set[4]

Using our trained model for S-NS from the previous section, we also tested our classifiers on this new dataset. Table 3 (S-NS: 100 tweets) shows that our best result (71%) was achieved by SMO using unigrams as features. This value is located between the extreme values of the established HBI.

These three studies show that humans do not perform significantly better than the simple automatic classification methods discussed in this paper. Some judges reported that the classification task was hard. The main issues judges identified were the lack of context and the brevity of the messages. As one judge explained, sometimes it was necessary to call on world knowledge such as recent events in order to make judgments about sarcasm. This suggests that accurate automatic identification of sarcasm on Twitter requires information about interaction between the tweeters such as common ground and world knowledge.

7 Conclusion

In this paper we have taken a closer look at the problem of automatically detecting sarcasm in Twitter messages. We used a corpus annotated by the tweeters themselves as our gold standard; we relied on the judgments of tweeters because of the relatively poor performance of human coders at this task. We semi-automatically cleaned the corpus to address concerns about corpus noisiness raised in previous work. We explored the contribution of linguistic and pragmatic features of tweets to the automatic separation of sarcastic messages from positive and negative ones; we found that the three pragmatic features – *ToUser*, *smiley* and *frown* – were among the ten most discriminating features in the classification tasks (Table 1).

We also compared the performance of automatic and human classification in three different studies. We found that automatic classification can be as good as human classification; however, the accuracy is still low. Our results demonstrate the difficulty of sarcasm classification for both humans and machine learning methods.

The length of tweets as well as the lack of explicit context makes this classification task quite difficult. In future work, we plan to investigate the impact of contextual features such as common ground.

Finally, the low performance of human coders in the classification task of sarcastic tweets suggests that gold standards built by using labels given by human coders other than tweets' authors may not be reliable. In this sense we believe that our approach to create the gold standard of sarcastic tweets is more suitable in the context of Twitter messages.

Acknowledgments

We thank all those who participated as coders in our human classification task. We also thank the anonymous reviewers for their insightful comments.

References

Carvalho, P., Sarmento, S., Silva, M. J., and de Oliveira, E. 2009. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). *In Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (TSA '09)*. ACM, New York, NY, USA, 53-56.

Clark, H. and Gerrig, R. 1984. On the pretence theory of irony. *Journal of Experimental Psychology: General*, 113:121–126. D.C.

Davidov, D., Tsur, O., and Rappoport, A. 2010. Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon, Dmitry Proceeding of Computational Natural Language Learning (ACL-CoNLL).

Derks, D., Bos, A. E. R., and Grumbkow, J. V. 2008. Emoticons and Online Message Interpretation. *Soc. Sci. Comput. Rev.*, *26*(3), 379-388.

Gibbs, R. 1986. On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General*, 105:3–15.

Gibbs, R. W. and Colston H. L. eds. 2007. *Irony in Language and Thought*. Routledge (Taylor and Francis), New York.

[4] The accuracy on the set they agreed on (83 out of 100 tweets) was 83.13%.

Kreuz, R. J. and Glucksberg, S. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118:374-386.

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* (pp. 1-4). Rochester, New York: Association for Computational.

LIWC Inc. 2007. The LIWC application. Retrieved May 10, 2010, from http://www.liwc.net/liwcdescription.php.

Nigam, K. and Hurst, M. 2006. Towards a Robust Metric of Polarity. In *Computing Attitude and Affect in Text: Theory and Applications* (pp. 265-279). Retrieved February 22, 2010, from http://dx.doi.org/10.1007/1-4020-4102-0_20.

Pak, A. and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, in 'Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)' , European Language Resources Association (ELRA), Valletta, Malta

Pang, B. and Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, July.

Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001).

Linguistic Inquiry and Word Count (LIWC): LIWC2001 (this includes the manual only). Mahwah, NJ: Erlbaum Publishers

Strapparava, C. and Valitutti, A. 2004. *Wordnet-affect: an affective extension of wordnet.* In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon.

Tepperman, J., Traum, D., and Narayanan, S. 2006. Yeah right: Sarcasm recognition for spoken dialogue systems. In *InterSpeech ICSLP*, Pittsburgh, PA.

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