Identification of Nonliteral Language in Social Media: A Case Study on Sarcasm

Smaranda Muresan
Center for Computational Learning Systems, Columbia University, New York, New York, USA. E-mail: smara@ccls.columbia.edu

Roberto Gonzalez-Ibanez
Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Santiago, Chile. E-mail: rgonzal@gmail.com

Debanjan Ghosh, Nina Wacholder
Department of Library and Information Science, Rutgers University, New Brunswick, New Jersey, USA. E-mail: debanjan.ghosh@rutgers.edu; ninwac@rutgers.edu

With the rapid development of social media, spontaneously user-generated content such as tweets and forum posts have become important materials for tracking people's opinions and sentiments online. A major hurdle for current state-of-the-art automatic methods for sentiment analysis is the fact that human communication often involves the use of sarcasm or irony, where the author means the opposite of what she/he says. Sarcasm transforms the polarity of an apparently positive or negative utterance into its opposite. Lack of naturally occurring utterances labeled for sarcasm is one of the key problems for the development of machine-learning methods for sarcasm detection. 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.

Introduction

In recent years, informal texts in discussion forums and microblogging platforms have become a major form of online communication, enabling the sharing of information by both individuals and organizations. This information can be factual (for example, about events or entities) or it can be about the people producing the content, such as their personal states (for example, sentiments, opinions, and beliefs). Being able to automatically track users' sentiments towards a product or a presidential candidate can have a major impact in many areas, from economics to political science. In recent years social media have been exploited as a source for sentiment and opinion analysis (e.g., Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Ramnath et al., 2011; Barbosa & Feng, 2010; Bermingham & Smeaton, 2010; Bollen, Pepe, & Mao, 2011; Go, Bhayani, & Huang, 2009; Hao et al., 2012; Pak & Paroubek, 2010).

One of the key problems in sentiment analysis is that human communication often employs *sarcasm* or *verbal irony*. For example, in the Internet Argumentation Corpus (Walker, Anand, Fox Tree, Abbott, & King, 2012) collected from 4forums.com, 12% of utterances contain sarcasm (as labeled by humans). In this article, we consider sarcasm to be "speech or writing which actually means the opposite of what it seems to say.[. . .]" (Collins COBUILD Advanced Learner's Dictionary). In the context of sentiment and opinion mining, detection of sarcasm is crucial, since what might look like a positive utterance is, in fact, a negative one, due to the use of sarcasm. For example, the utterances "Oh man, do I love doing sample returns" and "yeah, I really

This article is an extension of our short paper: R. González-Ibáñez, S. Muresan, and N. Wacholder (2011). Identifying Sarcasm in Twitter: A Closer Look. *Proceedings of ACL-HLT 2011*, pp. 581–586.

Received December 13, 2013; revised July 21, 2015; accepted July 22, 2015

© 2015 ASIS&T • Published online 23 December 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23624

wanna be on public transportation ALL DAY. sounds GREAT!" are sarcastic utterances, which use positive words and phrases ("love," "wanna," "great") but are in fact sarcastic and thus express a negative sentiment.

Automatic detection of sarcasm is in its infancy. One reason for this is that the ambiguity of sarcasm makes it hard even for people to identify it without the explicit indication that the sentence is sarcastic, as in the following sentence "that's what I love about Miami. Attention to detail in preserving historic landmarks of the past." Unlike the previous two examples, where one can use lexical factors to identify sarcasm (e.g., interjections such as "oh man" and "yeah") and other factors such as capitalization of words, in this utterance the recognition of sarcasm might depend on pragmatic factors, such as the establishment of common ground between the speaker and hearer, that is, their mutual knowledge, beliefs, and suppositions (Clark & Gerrig, 1984).

Another reason for the lack of computational models for detection of sarcasm has been the absence of naturally occurring utterances labeled as sarcastic to be used by supervised machine-learning methods. Microblogging platforms such as Twitter allow users not only to communicate their feelings, opinions, and ideas, but also to assign labels to their own messages. In Twitter, for example, messages can be annotated with hashtags by the author of the message, with the goal of allowing users to retrieve tweets on related themes (e.g., #bicycling, #teaparty, #joy, #sad, #sarcastic). Of course, the fact that some people feel that they need to make the sarcasm explicit does not mean that all sarcastic tweets are labeled with the #sarcastic or #sarcasm hashtags, but in this study we rely on tweets labeled with #sarcastic or #sarcasm as a corpus of labeled, naturally occurring data.

We conducted an empirical study of the use of lexical and pragmatic factors to distinguish sarcasm from straightforwardly positive and negative sentiments expressed in Twitter messages. Recent work on sentiment analysis on Twitter has focused primarily on identifying positive, negative, and neutral tweets (Go et al., 2009; Pak & Paroubek, 2010). A notable exception is the work of Davidov, Tsur, and Rappoport (2010), which presents a method for classifying sarcastic and nonsarcastic tweets. In this article, we focus on distinguishing among sarcastic, positive, and negative tweets, as one of our goals is to see whether we can identify specific lexical or pragmatic features that distinguish sarcastic from nonsarcastic sentiment utterances, with the final goal of improving sentiment classification of tweets. The contributions of this article include (a) the creation of a corpus that includes sarcastic utterances that have been explicitly identified as such by the writer of the message, and nonsarcastic utterances that express positive or negative attitudes without sarcasm; (b) a report on the performance of distinguishing sarcastic tweets from tweets that are straightforwardly positive or negative both by automatic methods (supervised machine-learning methods) and by human judges. Our results suggest that lexical features alone are not sufficient for identifying sarcasm and that pragmatic and contextual features merit further study. In addition, the results show that using more training data helps classifier performance.

In the next section we discuss Related Work on sarcasm detection. In Data Collection and Computational Framework we present our approach and an overview. In the following section we discuss the Lexical and Pragmatic Features used in our approach, and an analysis of their significance in distinguishing among sarcastic, positive, and negative utterances. Then the results of our experiments using three supervised machine-learning approaches: support vector machines, naïve bayes and logistic regression are presented; we also assess the impact of the size of the training data on classifier performance. In Comparison of Machine-Learning Methods Against Human Performance we establish an upper bound and assess the difficulty of our task by comparing human performance to that of machine-learning methods. Following that is a Discussion of our findings and plans for future work.

Related Work

Sarcasm and irony are well-studied phenomena in linguistics, psychology, and cognitive science (Gibbs, 1986; Gibbs & Colston, 2007; Kreuz & Glucksberg, 1989; Utsumi, 2000). There is a fine line between sarcasm and irony: for Colston (2007), sarcasm is a term commonly used to describe an expression of verbal irony; whereas for Gibbs (2007), sarcasm, along with jocularity, hyperbole, rhetorical questions, and understatement, are types of irony. Attardo (2007) considers sarcasm to be an overtly aggressive type of irony. In our work, we use a relatively straightforward definition of sarcasm as "speech or writing which actually means the opposite of what it seems to say. [Colling Cobuild English Dictionary for Advanced Learners 4th Edition, 2003. HarperCollins Publishers.]"

In the text-mining literature, automatic detection of sarcasm is considered a difficult problem. Its presence is a major obstacle to accurate sentiment analysis (Councill et al., 2010; Nigam & Hurst, 2006; Li et al., 2012; Pang & Lee, 2008) and has been addressed in only a few studies. In the context of spoken dialogs, automatic detection of sarcasm has relied primarily on speech-related cues such as laughter and prosody (Tepperman, Traum, & Narayanan, 2006). The works most closely related to ours are Davidov et al. (2010) and Reyes and Rosso (2011). The former aims to identify sarcastic and nonsarcastic utterances in Twitter and in Amazon product reviews. In this article, we consider the somewhat harder problem of distinguishing sarcastic tweets from nonsarcastic tweets that directly convey positive and negative attitudes (we do not consider neutral utterances at all). Our work aims at investigating what features distinguished sarcastic utterances from positive and negative ones. Reyes and Rosso (2011) tackle the more general problem of irony detection in customer reviews on Amazon, comparing the ironic reviews against plain negative reviews from Amazon and Slashdot.

Our approach of looking at lexical features for identification of sarcasm was inspired by the work of Kreuz and

Caucci (2007), which studied the role of lexical factors in sarcasm identification. In their study, Kreuz and Caucci collected statements containing the phrase “said sarcastically,” and removed this phrase from the statement. They then presented the abbreviated statements to human subjects and asked them to code each one as sarcastic or nonsarcastic. They looked at a handful of lexical factors (presence of adjective and adverbs, interjections, and punctuation). Their results showed that interjections are useful in distinguishing sarcastic and nonsarcastic utterances. In our work we used an empirical approach to identifying lexical factors based on lexicons such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001) and WordNet Affect (Strapparava & Valitutti, 2004), as well as punctuation and interjections. In addition, our work explores the use of pragmatic factors such as emoticons (explored also by Carvalho, Sarmento, Silva, & de Oliveira, 2009) and common ground.

An important line of work is building corpora for sarcasm and irony. Most work has relied on crowd sourcing (Filatova, 2012; Reyes & Rosso, 2011; Walker et al., 2012) to make judgments of whether a text is sarcastic or not; the limitation of this approach is that the original intention of the author/speaker is unknown. In this study, we treat the #sarcastic and #sarcasm hashtags assigned by the authors of the tweets as evidence of the author’s intention.

Since our work is focused on Twitter data, it is worth mentioning that Twitter has become a major resource for work in natural language processing, from sentiment analysis (Agarwal et al., 2011; Bollen et al., 2011; Go et al., 2009; O’Connor, Balasubramanyan, Routledge, & Smith, 2010; Pak & Paroubek, 2010), to event detection (Becker, Naaman, & Gravano, 2011; Petrovic, Osborne, & Lavrenko, 2010), topic modeling (Ramage, Dumais, & Liebling, 2010), and dialog acts identification (Ritter, Cherry, & Dolan, 2010). Twitter has also been a major source of information for different applications such as earthquake detection (Sakaki, Okazaki, & Matsuo, 2010) and disease surveillance (Lamb, Paul, & Dredze, 2013; Signorini, Segre, & Polgreen, 2011; Sadilek, Kautz, & Silenzio, 2012).

Data Collection and Computational Framework

Twitter is a very popular microblogging service, which allows users to post short messages up to 140 characters, called tweets. The users who post the messages are called tweeters. A tweet can contain references to other users (@<user>), URLs, and hashtags (#hashtag), which are tags assigned by the tweeter to mark content/topic (#teaparty, #worldcup), sentiment (#angry, #sad, #happy, #sarcasm), and/or location (#Paris, #NYC), among other uses. These hastags are later used for indexing purposes. An example of a tweet is: “@UserName1 check out the twitter feed on @UserName2 for a few ideas :) http://xxxxxx.com #happy #hour.” The characters in the URLs and hashtags count toward the 140-character limit. This length restriction forces the user to be very concise. For natural language processing techniques this is both an advantage (lexical factors may be more prominent than syntactic factors) and a challenge (abbreviations and symbols with special interpretation in Twitter may decrease the effectiveness of natural language processing tools optimized for more standard language).

In terms of knowledge needed to interpret the meaning of a tweet we can classify tweets into three categories: (a) **external-context tweets:** tweets that include a URL as reference; (b) **conversational-context tweets:** tweets that are part of a conversation thread among users, marked usually by @<user>, where the interpretation of a particular tweet is most likely dependent on the entire conversation; and (c) **noncontext tweets:** tweets that do not contain external references such as URLs, and are not part of a conversation thread among users. In this study, as in most work on sentiment analysis on Twitter, we treat all tweets as noncontext tweets, that is, we do not use the URL as external context, and we do not consider the entire conversation thread, even if the tweet is part of the conversation thread as indicated by the @<user> markup. We treat each tweet individually, both in the machine-learning experiments (Classification Experiments) and in the human judges studies (Comparison of Machine-Learning Methods Against Human Performance).

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 Comparison of Machine-Learning Methods Against Human Performance, 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, a point we come back to in the final section, where we discuss our findings. 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) (see Table 1 for a full list of hashtags used). For the straightforward expression of positive and negative sentiments we selected seed words expressing positive and negative emotions listed in resources such as LIWC (Pennebaker et al., 2001) and WordNet-Affect (Strapparava & Valitutti, 2004). After collecting thousands of tweets for each of the three categories we applied a set of automatic filtering steps to clean the corpus. We removed retweets,

TABLE 1. Tweets in corpus after filtering process.

Category	Hashtags	#Instances
S	#sarcasm, #sarcastic	2,151
P	#happy, #joy, #happiness, #love, #grateful, #optimistic, #loved, #excited, #positive, #wonderful, #positivity, #lucky	900
N	#angry, #frustrated, #sad, #scared, #awful, #frustration, #disappointed, #fear, #sadness, #hate, #stressed	1,276

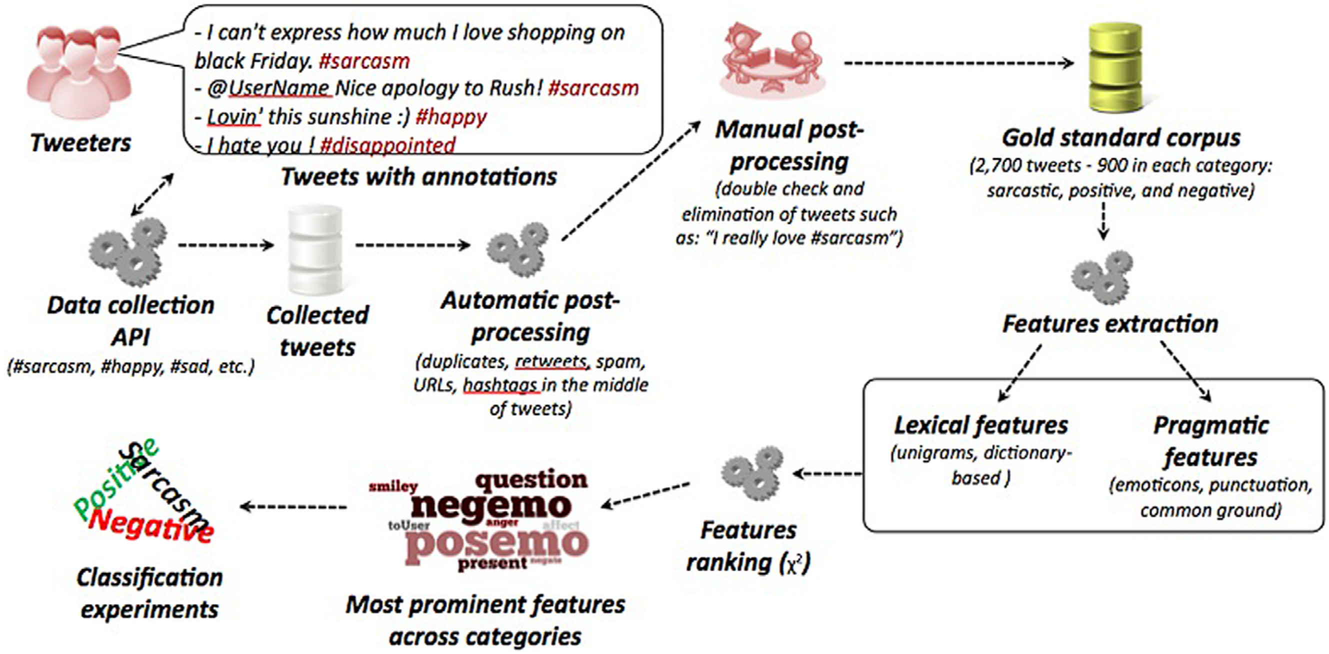
- I can't express how much I love shopping on black Friday. #sarcasm
- @UserName Nice apology to Rush! #sarcasm
- Lovin' this sunshine :) #happy
- I hate you ! #disappointed
Tweeters
Tweets with annotations
Manual post-processing
(double check and elimination of tweets such as: "I really love #sarcasm")
Gold standard corpus
(2,700 tweets - 900 in each category: sarcastic, positive, and negative)
Data collection API
(#sarcasm, #happy, #sad, etc.)
Collected tweets
Automatic post-processing
(duplicates, retweets, spam, URLs, hashtags in the middle of tweets)
Features extraction
Lexical features
(unigrams, dictionary-based)
Pragmatic features
(emoticons, punctuation, common ground)
Positive
Sarcasm
Negative
Classification experiments
smiley
question
negemo
toUser
anger
affect
posemo
present
negate
Most prominent features across categories
Features ranking (χ²)

FIG. 1. Overview of the computational framework for sarcasm detection. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

duplicates, quotes, spam, and tweets written in languages other than English. In addition, we removed all tweets where the hashtags of interest were not located at the very end of the message. Then we performed a manual review of the filtered tweets in order to double check that hashtags were not part of the message. As a result, messages such as "lol thanks. I can always count on you for comfort :) #sarcasm" were kept, while messages like "I really love #sarcasm" were eliminated. However, for the nonsarcastic tweets (positive or negative), the absence of contact with the original tweeters meant that we could not verify that the tweets were definitely nonsarcastic.

Table 1 shows the number of tweets remaining after filtering with manual validation was completed.

Our final corpus consists of 900 tweets in each of the three categories, sarcastic, positive, and negative. This corpus is used in all experiments in the following three sections. However, in Classification Experiments we report on the impact of using a larger data set (collected using the same methods described earlier) on classifier performance. 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 to 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.

Figure 1 presents an overview of our computational framework, including the steps used to build the gold standard corpus.

Our task can be seen as a text classification problem, where the "text" to be classified is the tweet, and the "categories" are sarcastic, positive, and negative. We use three standard machine-learning algorithms for text classification: support vector machines, naïve Bayes, and logistic regression. To implement these machine-learning algorithms on our gold-standard data, we used the standard bag-of-features framework. We use both lexical features (n-grams and lexicon-based) and pragmatic features (emoticons, common ground) (see next section for details). Using a χ^2 test, we then analyze whether there are discriminative features useful to distinguish among the three categories (i.e., sarcastic, positive, and negative). All tweets are represented using a feature vector representation and three standard classification algorithms are used. We compare the accuracy of these automatic methods on the task of classifying sarcastic, positive, and negative tweets (see Classification Experiments for more detail).

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 among sarcastic, positive, and negative utterances (tweets).

Lexical Features

We used two kinds of lexical features—n-grams (unigrams and bigrams) and lexicon-based. The lexicon-based features were derived from Pennebaker et al.'s LIWC (Pennebaker et al., 2007) dictionary, emotion words from WordNet-Affect (Strapparava & Valitutti, 2004), and a list of interjections and punctuation.

LIWC dictionary[1] has been used widely in computational approaches to sentiment, emotion, and opinion analysis. It consists of a set of 64 word categories grouped into four general classes organized hierarchically: (a) Linguistic Processes (LP) (e.g., Adverbs, Pronouns, Past Tense, Negation); (b) Psychological Processes (PP) (e.g., Affective Processes [Positive Emotions, Negative Emotions (Anxiety, Anger, Sadness)], Perceptual Processes [See, Hear, Feel], Social Processes, etc); (c) Personal Concerns (PC) (e.g., Work, Achievement, Leisure); and (d) Spoken Categories (SC) (Assent, Nonfluencies, Fillers). The LIWC dictionary contains around 4,500 words and word stems.

WordNet-Affect (Strapparava & Valitutti, 2004) is an affective lexical resource of words referring to emotional states. WordNet-Affect extends WordNet (Miller, 1990) by assigning a variety of affect labels to a subset of synsets representing affective concepts in WordNet. In our study we used the words annotated for associations with six emotions considered to be the most basic—joy, sadness, fear, disgust, anger, and surprise (Ekman, 1992), a total of 1,536 words.

In addition to LIWC and the emotion words from WordNet-Affect we use a list of interjections (e.g., ah, oh, yeah),[2] and punctuations (e.g., !, ?). The latter are inspired by results from Kreuz and Caucci (2007). We merged all of these lexicons into a single combined lexicon. The token overlap between the words in the combined lexicon and the words in our corpus of tweets was 85%. This demonstrates that lexical coverage is good, even though tweets are well known to contain many words that do not appear in standard lexicons. In addition, some of the tokens that are not part of the lexicon can be #hashtags that appear in other tweets (besides the #hashtags used to define the class). For example, the tweet "How convenient. A crucial call in michigan's favor when they are trailing . . . that NEVER happens. #eyeroll #sarcasm," contains an additional #hashtag #eyeroll, which could be an additional feature to indicate sarcasm in addition to the adverb NEVER in all capital letters. That is why, as lexical features, we use both n-grams and lexicon-based features.

Pragmatic Features

Regarding pragmatic features, we take into account the conversation among users. If a sarcastic/positive/negative tweet is in reply to another user (identified by the @UserName as seen in examples 1, 2, and 3 given earlier), it can be considered an indication of common ground. We denote this feature by *ToUser*. In addition, we used positive emoticons such as smileys and negative emoticons such as frowning faces (Carvalho et al., 2009).

Feature Ranking

To measure the impact of features on discriminating among the categories under study, first each tweet is represented as a feature vector. Let $\{f_1, \ldots, f_m\}$ be a predefined set of m features (e.g., lexicon-based features and pragmatic features we discussed in the previous two subsections). Each tweet t is represented by a feature vector: $\vec{t} = (n_1(t), \ldots, n_m(t))$, where $n_i(t)$ can be defined either by the presence of the feature f_i in the tweet/document t (1 if f_i is present or 0 if it is not present) or by frequency (i.e., the number of time f_i appears in t).

We performed four studies: (a) three-way classification of sarcastic (S) versus positive (P) versus negative (N) messages (S-P-N); (b) a two-way classification of sarcastic versus nonsarcastic (S-NS), where the NS set is built by merging 450 randomly selected positive and 450 randomly selected negative tweets from our corpus (so that we obtain balanced data sets of 900 tweets in each of S and NS); (c) a two-way classification sarcastic versus positive (S-P); and (d) a two-way classification of sarcastic versus negative (S-N).

We ran a χ^2 test to identify the features that were most useful in discriminating among these categories. Table 2 shows the top 10 features based on the presence of all the lexicon-based features plus the pragmatic features discussed earlier, a total of 80 features. We refer to this set of features as LexPrag-P. Similar findings are obtained using the frequency-based representation.

In all tasks, negative emotion (*Negemo*), positive emotion (*Posemo*), negation (*Negate*), emoticons (*Smiley, Frown*), auxiliary verbs (*AuxVb*), and punctuation marks (Question) are in the top 10 features. For example, question marks appear more in sarcastic tweets than in positive and negative tweets (Figure 2). Kreuz and Caucci (2007), who looked at

TABLE 2. Ten most discriminating features in LexPrag-P for each task.

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)

[1] For a list of all LIWC categories and examples of words see http://www.liwc.net/descriptiontable1.php

[2] http://www.vidarholen.net/contents/interjections/

DOI: 10.1002/asi

sarcastic utterances in published works from Google Books, found that punctuations, such as question marks and exclamation, were not predictive of a significant amount of the variance in the participants' ratings of sarcastic intent. Our study could be an indication that the expression of sarcasm varies by genre and medium of communication (tweets are quite different from narrative books). However, more work will need to be performed, since our task is different from previous studies by looking at sarcastic messages compared to positive and negative messages. With regard to emoticons, 3.1% of sarcastic tweets included smiley faces and 1.6% of the sarcastic tweets included frowning faces. On the other hand, 11.1% of positive tweets included smiley faces and 0.5% frowning faces. Finally, only 0.2% of negative tweets have smiley faces and 5.2% have frowning faces. Thus, it seems that sarcastic tweets are similar to positive tweets in terms of the use of more smiley faces and closer to negative tweets in terms of the use of more frowning faces. However, due to low coverage of emoticons in our corpus (only 7.5% of the annotated tweets contained emoticons), we cannot draw a significant conclusion.

We also observe indications of a possible dependence among factors that could differentiate sarcasm from both positive and negative tweets: sarcastic tweets tend to contain positive emotion words, just as positive tweets do (Posemo is a significant feature in S-N but not in S-P; also see Figures 2, 4, and 5), while they use more negation words like negative tweets do (Negate is an important feature for S-P; also see Figures 2 and 5).

Table 2 and Figures 2 and 3 also show that the pragmatic feature ToUser is important in sarcasm detection. In our corpus, 28.7% of sarcastic tweets included a user as recipient of the message (ToUser category), while only 9.9% of the positive and 17.9% of negative tweets were addressed to other users. This is an indication of the possible importance of features that indicate common ground in sarcasm identification.

Figures 2, 3, 4, and 5 show the presence of the top 20 features in tweets of different categories (S-P-N, S-NS, S-P, S-N).

Classification Experiments

In this section we address the question "Are lexical and pragmatic factors useful features in machine-learning algorithms for automatic classification of utterances (tweets) into sarcastic, positive, and negative?"

We used three standard classifiers often employed in sentiment classification: naïve Bayes (NB), support vector machine with sequential minimal optimization (SVM), and logistic regression (LogR). For features we used: (a) unigrams; (b) presence of lexicon-based features and pragmatic features (LexPrag-P); (c) frequency of lexicon-based features and pragmatic features (LexPrag-F); and (d) combination of unigrams and lexicon-based and pragmatic features (unigrams + LexPrag-P). The classifiers were trained on balanced data sets (900 instances per class) and tested through 5-fold cross-validation. We run both the three-way

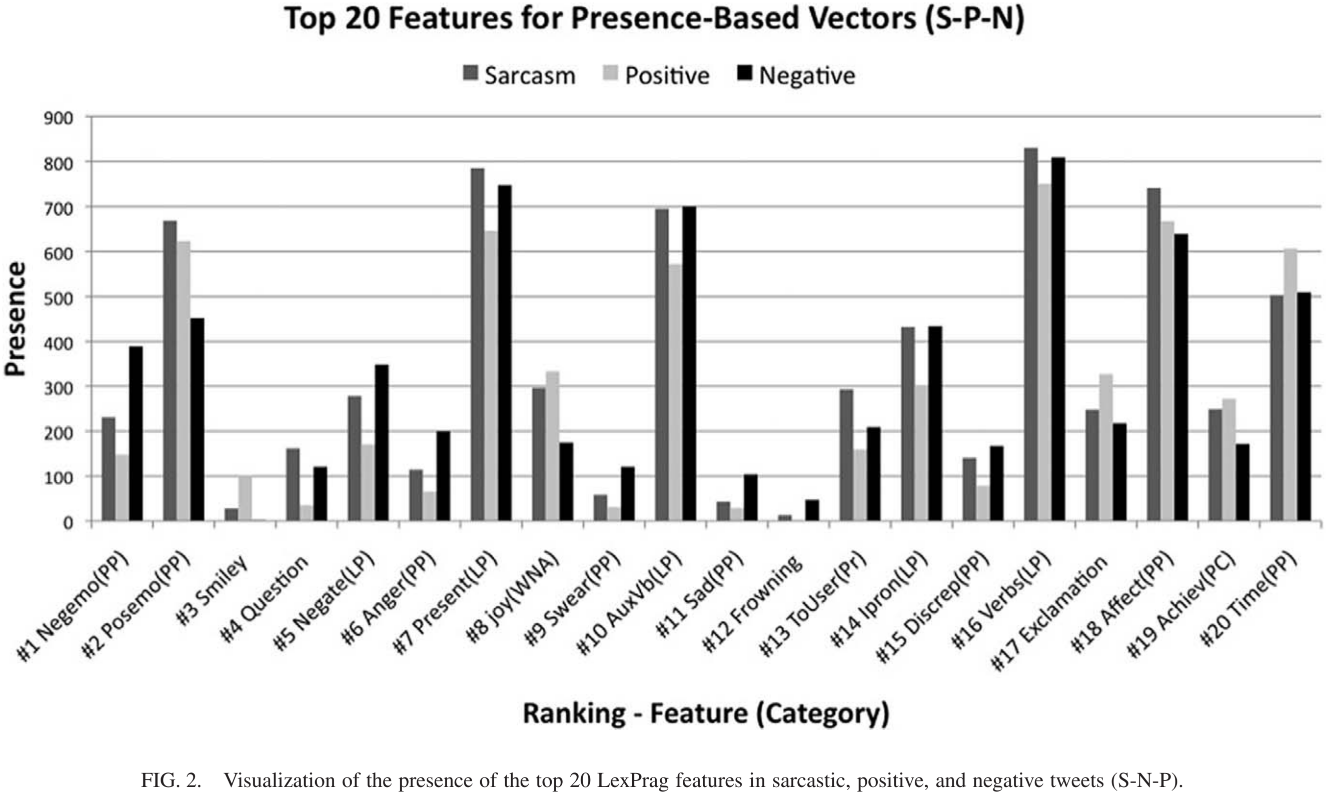

FIG. 2. Visualization of the presence of the top 20 LexPrag features in sarcastic, positive, and negative tweets (S-N-P).

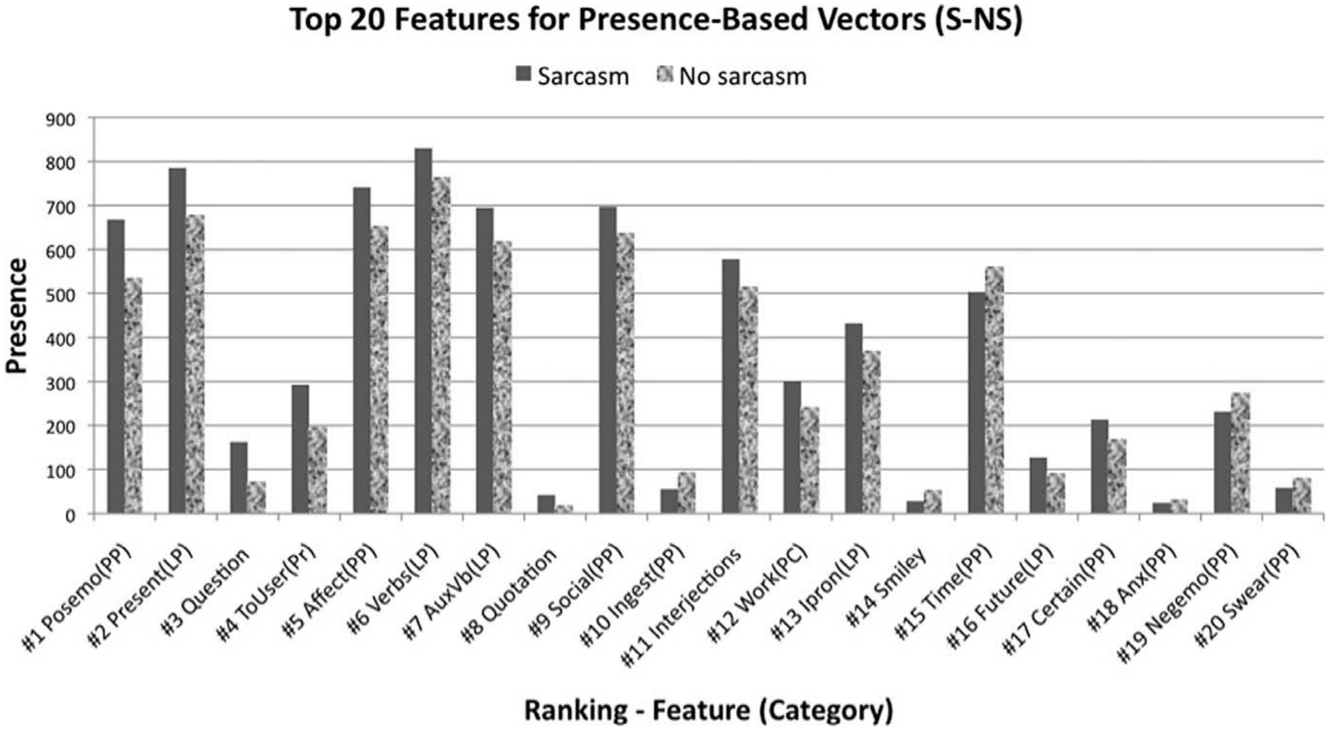

FIG. 3. Visualization of the presence of the top 20 LexPrag features in sarcastic and nonsarcastic tweets (S-NS).

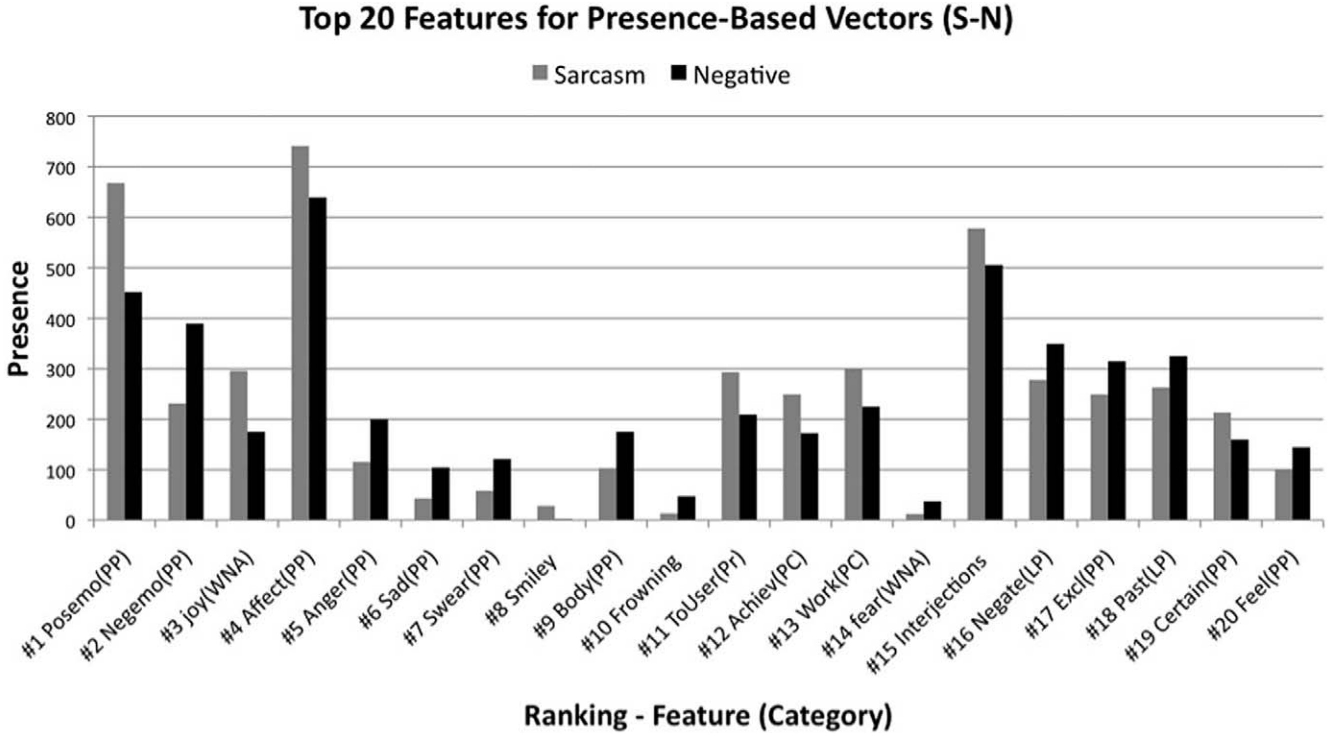

FIG. 4. Visualization of the presence of the top 20 LexPrag features in sarcastic an negative tweets (S-N).

classification (S-P-N) and several binary classifications: sarcastic versus nonsarcastic (S-NS), sarcastic versus positive (S-P), sarcastic versus negative (S-N), and positive versus negative (P-N). For the three-way classification (S-P-N), the SVM with sequential minimal optimization performs a one-against-all multiclass classification.

In Table 3, bolder values indicate the best accuracies for each task. In the three-way classification (S-P-N), SVM with

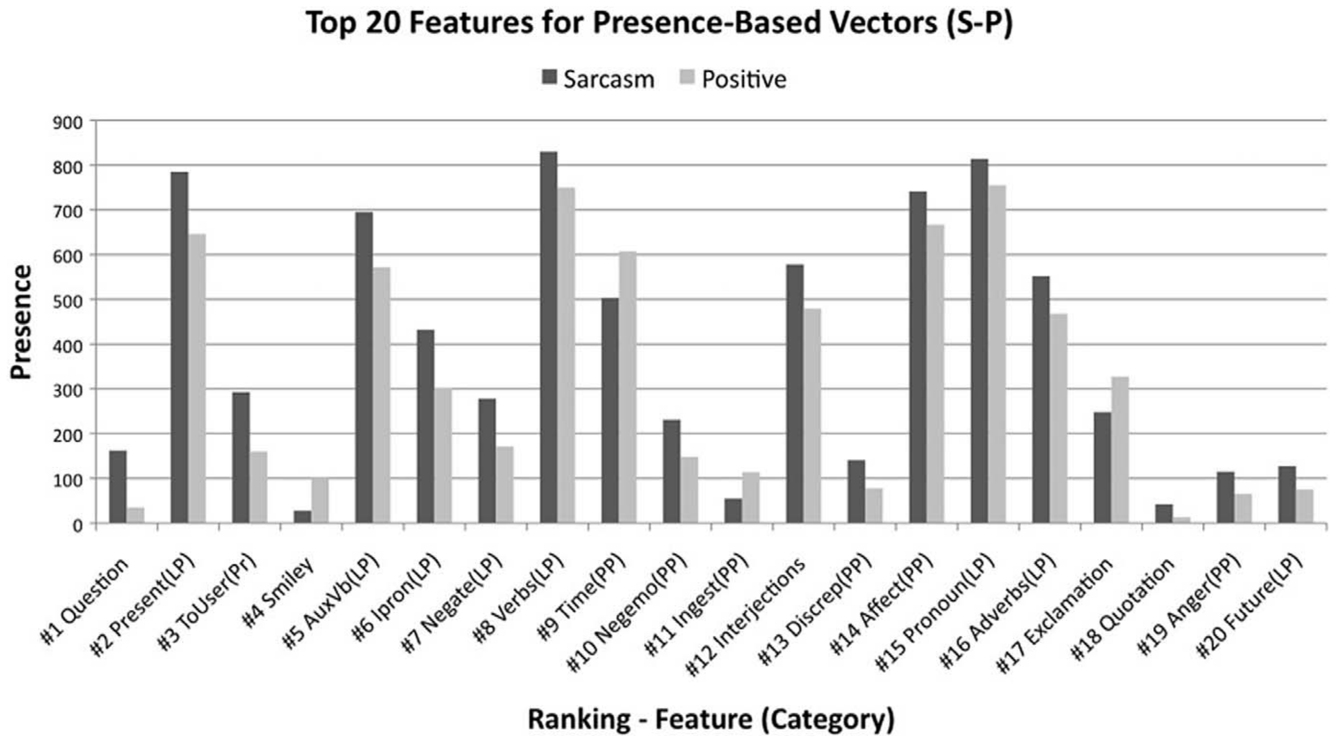

FIG. 5. Visualization of the presence of the top 20 LexPrag features in sarcastic and positive tweets (S-P).

TABLE 3. Classifiers accuracies using 5-fold cross-validation, in percentage.

Class	Test	Features	NB	SMO	LogR
S-P-N	1	Unigrams	55.77	*58.55*	50.11
	2	LexPrag-F	50.11	55.81	55.48
	3	LexPrag-P	52.67	55.77	56.22
	4	Unigrams + LexPrag-P	57.30	**60.52**	59.07
S-NS	5	Unigrams	64.66	*66.44*	60.61
	6	LexPrag-F	57.28	61.33	60.16
	7	LexPrag-P	62.22	62.77	62.88
	8	Unigrams + LexPrag-P	66.17	**68.10**	63.94
S-P	9	Unigrams	70.72	*71.05*	64.5
	10	LexPrag-F	66.83	66.66	67.50
	11	LexPrag-P	65.83	67.44	67.55
	12	Unigrams + LexPrag-P	70.00	**72.10**	70.10
S-N	13	Unigrams	66.44	69.5	63.55
	14	LexPrag-F	60.67	68.44	68.00
	15	LexPrag-P	66.78	68.33	*68.72*
	16	Unigrams + LexPrag-P	69.38	**73.00**	*69.11*
P-N	17	Unigrams	71.77	74.44	71.94
	18	LexPrag-F	68.33	75.83	75.72
	19	LexPrag-P	72.61	*75.88*	75.72
	20	Unigrams + LexPrag-P	75.77	***78.30***	75.89

unigrams as features outperformed SVM with LexPrag-P and LexPrag-F as features. Overall, SVM outperformed LogR and NB methods. The best result, 60.52%, was obtained using SVM with the combination of unigram and LexPrag-P features (a baseline classifier that chooses the majority class that would give 33%, since we are using balanced data sets). The best result is an indication of the difficulty of the identification of sarcasm by machine-learning systems. In the following section we present a study showing how humans perform on this task, thus establishing an upper bound.

The performance of the three classifiers improved in the evaluations of two-way classifications. First, after merging 450 randomly chosen positive and 450 randomly chosen negative tweets to create the nonsarcastic (NS) class, the best results were again obtained using SVM with the combination of unigram and LexPrag-P features (68.10%). A baseline classifier that chooses the majority class would give 50%, since we are using balance data sets. The presence-based features (LexPrag-P) outperformed the frequency-based features (LexPrag-F) using all three classifiers. For the S-P and S-N task the best accuracies were 72.10% and 73%, respectively. Overall, our best result (78.3%) was achieved in the polarity-based classification P-N, which is to be expected. The machine-learning systems have roughly equal difficulty in separating sarcastic tweets from positive tweets as from negative tweets.

These results indicate that the lexical and pragmatic features considered in this article do not provide sufficient information to accurately differentiate sarcastic from positive and negative tweets. One reason might be the size of the data set, which is relatively small (900 data points per category). To assess the impact of the size of training data on classifier performance, we collected a larger data set using the exact same procedure as discussed, except that we did not do the manual review at the end due to the large size

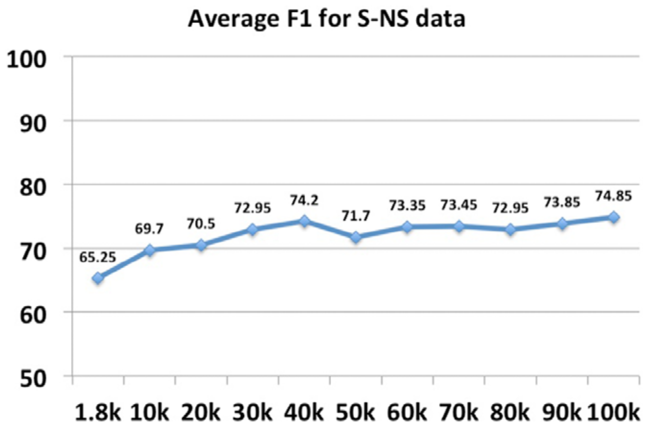

FIG. 6. Impact of the size of training data on classification accuracy for the S-NS task. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

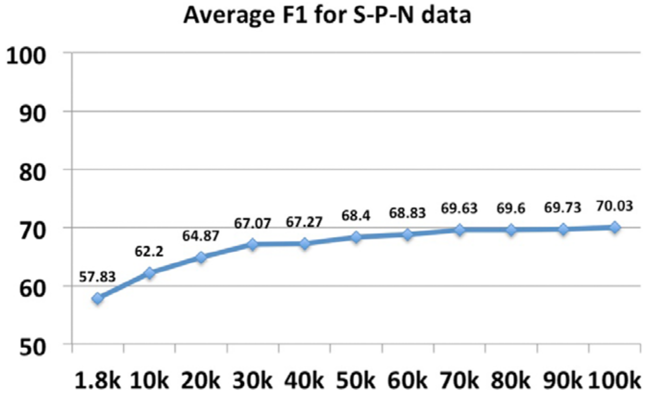

FIG. 7. Impact of the size of training data on classification accuracy for the S-P-N task. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

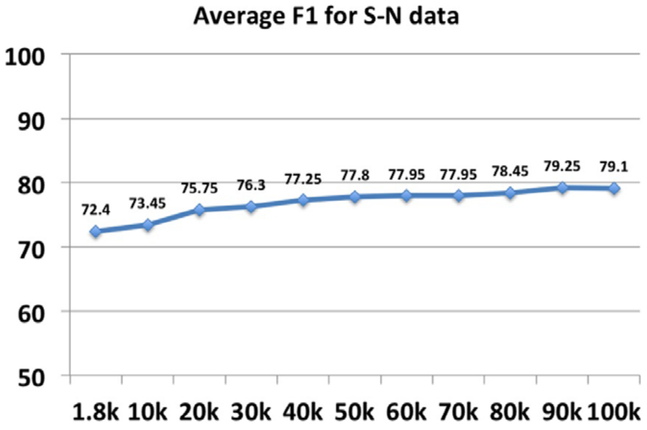

FIG. 8. Impact of the size of training data on classification accuracy for the S-N task. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

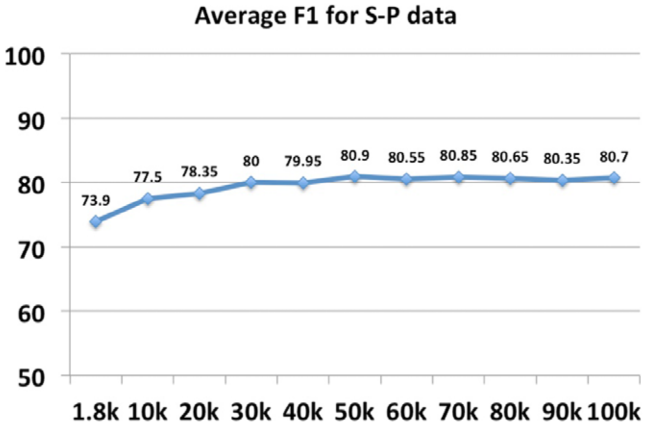

FIG. 9. Impact of the size of training data on classification accuracy for the S-P task. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of our data set. We created a test set of 1,000 tweets in each of the three categories: S, P, and N. For the two-way classification tasks (S-NS, S-P, S-N) the training data varied in size from 1.8 K[3] (as in the previous experiments, 900 data points per category) to 100 K (50 K per category). Similarly for the three-way classification S-P-N, the training data varied in size from 2.7 K (900 data points per category) to 100 K (~33 K per category). In these experiments, we used the best feature combination (Unigrams + LexPrag-P) and the best classifier (SVM). The results in Figures 6, 7, 8, and 9 show that for all classification tasks, adding more data leads to a significant improvement in performance. The figures show the F1 measure on the Y axis, and the size of training data on the X axis (increments of 10 K). For example, for the S-NS task (Figure 6) when training on a data set of size 1.8 K (similar to our original data set of 900 data points per category) the F1 is 65.25%, while training one data set of size 100 K the performance increases to 74.85% F1 measure. However, for this task the performance when using a training set of 40 K instances is similar to performance when using 100 K training instances (74.2% vs. 74.85%, respectively). For the S-P-N task, adding more data showed consistent improvement, with the best performance achieved when using 100 K instances in the training set (70.3% F1 compared to 57.83% F1 when using only 2.7 K, which is the same size as our original data set).

These experiments show that using more data helps improve the classifier performance. However, the results show that there is still room for improvement. This may be due to the inherent difficulty of distinguishing short utterances in isolation, without the use of contextual evidence. We present a more detailed discussion and avenues for future work in the final section.

[3] This data set of 1.8 K is different from the data set used in the previous experiments, since we collected tweets from a different timeframe than the ones we used in the previous study. But we keep the same size for this first data point, as our goal for this experiment is to test the impact of training set size.

TABLE 4. SVM and LogR accuracies against human performance for the S-P-N task on the T1 and T2 test sets.

Test set	T1 test set (270 tweets)				T2 test set (135 tweets)			
Human	HBI = [43.33%–62.59%]				86.67%			
ML	Features	NB	SVM	LogR	Features	NB	SVM	LogR
1	*Unigrams*	51.11	58.51	50.04	*Unigrams*	53.33	57.80	51.20
2	LexPrag-F	50.74	52.22	54.90	LexPrag-F	52.80	54.81	55.30
3	LexPrag-P	51.90	52.40	51.11	LexPrag-P	52.90	56.29	**59.90**
4	*Unigrams* + LexPrag-P	55.93	**59.30**	56.67	*Unigrams* + LexPrag-P	59.60	58.50	58.20

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

Comparison of Machine-Learning Methods 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, native speakers of English, and not ourselves.

In the first study, we asked three judges to classify 10% of our original S-P-N data set (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. The annotators were not informed about the distribution of the S, P, and N classes.

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). In 13.58% (13.44) of cases, judges were unsure about the correct category. When we considered only the 135 of 270 tweets on which all three judges agreed, the accuracy computed over the entire gold standard test set fell to 43.33%. The accuracy of the human judges on the 135 tweets about which all agreed was 86.67%. One issue when using #hashtags as automatic labels for sarcastic and nonsarcastic messages is that the nonsarcastic class(es) (i.e., positive and negative classes) might in fact contain sarcastic messages, where the user just did not label the tweet as #sarcastic. To address this issue, we looked to see how many of the 180 nonsarcastic tweets (P and N) were considered as sarcastic by all three coders or by two coders (majority voting). None of the 180 nonsarcastic tweets were considered sarcastic when all three coders agree, and only one out of 180 was considered sarcastic by two coders ("Like what's wrong with u can't get into a REAL sorority huh!"). This finding shows the promise of our method of selecting nonsarcastic tweets using hashtags based on sentiment-words such as #happy and #sad. In future work we plan to do a larger study regarding this issue.

We trained our NB, SVM, and LogR classifiers on the other 90% of the original S-P-N data set. The models were then evaluated on two test sets: T1 includes the entire 10% of the S-P-N data set that was also labeled by humans (270 tweets); and T2 includes a subset of T1 which represents only the tweets on which all the judges agreed (135 tweets). This unbalanced set consists of 29 sarcastic tweets, 60 positive tweets, and 46 negative tweets. For T1, we calculated a human baseline interval (HBI) to compare the machine-learning results against the human performance: we used the accuracy when the judges agree computed over the entire gold standard (43.33%) and the average accuracy (62.59%). For the second set, we used the accuracy of the human judges on that set (86.67%).

On the first test set T1, the automatic classification accuracies were similar to results obtained in the previous section. Our best result—an accuracy of 59.3%—was achieved using SVM with a combination of Unigrams and LexPrag_P features (Table 4: S-P-N). The highest value in the established HBI had a slightly higher accuracy; however, when compared to the bottom value of the same interval, our best result significantly outperformed it.

On the second test set T2, the automatic classification accuracies were comparable with the accuracies on the T1 test (best result is 59.90% using LogReg with LexPrag_P features). This result is somewhat unexpected, as T2 represents the tweets on which all human subjects agree, being thus arguably easier to label.

In the second study, we investigated how well human judges performed on the two-way classification task of labeling sarcastic and nonsarcastic tweets. We asked three other judges to classify 10% of our original S-NS data set (i.e., 180 tweets) into sarcastic and nonsarcastic. The results showed an agreement of 71.67% among the three judges, with a Fleiss's 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%. The accuracy on the set they agreed on (129 out of 180 tweets) was 82.95%.

As with the previous study, we trained our NB, SVM, and LogR classifiers on the other 90% of the original S-NS data set. The models were then evaluated on two test sets: T3—the entire 10% of the S-NS data set that was also labeled by humans (180 tweets); and T4—a subset of T3 which represents only the tweets on which all the judges agreed (129 tweets). For T3, we calculated the HBI, as in our

TABLE 5. SVM and LogR accuracies against human performance for the S-NS task on the T3 and T4 test sets (tweets without emoticons).

Test set	T3 test set (180 tweets)				T4 test set (129 tweets)			
Human	HBI = [59.44%–66.85%]				82.95%			
ML	Features	NB	SVM	LogR	Features	NB	SVM	LogR
1	*Unigrams*	64.44	*68.33*	58.88	*Unigrams*	*62.02*	66.67	59.91
2	LexPrag-F	58.89	62.22	61.67	LexPrag-F	58.14	61.24	59.69
3	LexPrag-P	63.33	67.22	67.22	LexPrag-P	62.02	62.02	60.46
4	*Unigrams* + LexPrag-P	67.78	**71.67**	68.90	*Unigrams* + LexPrag-P	63.40	**68.10**	66.80

TABLE 6. SVM and LogR accuracies against human performance for the S-NS task on the T5 and T6 test sets (tweets with emoticons).

Test set	T5 test set (100 tweets)				T6 test set (83 tweets)			
Human	HBI = [70%–73%]				83.13%			
ML	Features	NB	SVM	LogR	Features	NB	SVM	LogR
1	*Unigrams*	55	72.00	71.00	*Unigrams*	*65.06*	77.11	72.29
2	LexPrag-F	53	61.00	60.00	LexPrag-F	50.60	62.65	61.45
3	LexPrag-P	55	52.00	53.00	LexPrag-P	59.04	54.22	56.63
4	*Unigrams* + LexPrag-P	59.60	66.67	55.8	*Unigrams* + LexPrag-P	68.5	70.90	62.20

previous study: we used the accuracy when the judges agree computed over the entire gold standard as lower end (59.44%) and the average accuracy as higher end (66.85%). For T4, we used the accuracy of the human judges on that set (82.95%).

In this second study, all classifiers using the combination of unigrams and LexPrag_P features outperformed the HBI accuracies on the T3 set (67.78% for NB, 71.67% for SVM, 68.90% for LogReg, compared to the HBI of 59.44%–66.85%) (see Table 5). On the T4 set, the automatic classification accuracies were slightly worse than for the T3 set (best accuracy was 68.80%) and significantly worse than human performance for this set (82.95%).

Based on recent results which show that nonlinguistic cues such as emoticons are helpful in interpreting nonliteral meaning, such as sarcasm and irony in user-generated content (Carvalho et al., 2009; Derks, Bos, & Grumbkow, 2008), we explored how much emoticons help humans to distinguish sarcastic from positive and negative tweets. In this third study, we created a new data set using only tweets with emoticons. This data set consisted of 50 sarcastic tweets and 50 nonsarcastic tweets (25 positive and 25 negative). Two human judges classified the tweets using the same procedure as described earlier. For this task, judges achieved an overall agreement of 89% with Cohen's kappa value of 0.74 ($p < .001$). This result shows that emoticons play an important role in helping people distinguish sarcastic from nonsarcastic 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 to the entire gold standard set. The accuracy on the set they agreed on (83 out of 100 tweets) was 83.13%.

In this third study, we used our trained models (NB, SVM, and LogR) for S-NS from the second study. Again, we tested the automatic methods on two sets: T5 consists of the full 100 tweets set, and T6 consists of the subset of T5 where all the judges agree (83 tweets). Table 6 shows that on T5, SVM using unigram features achieved the best results for the automatic methods (71%). This value is located between the extreme values of the established HBI (70%–73%). The lower performance of the automatic classification methods can be explained by the fact that our gold standard data contained a relatively small number of tweets containing emoticons. For the T6 test set (where all human judges agree), the accuracies were significantly higher than for the T5 test set (highest accuracy was again obtained using SVM with unigram features). However, the results of the automatic methods were significantly lower than the accuracy of human judges, which was 83.13%.

These three studies show that humans do not always perform significantly better than the simple automatic classification methods discussed in this article. The humans significantly outperformed the automatic methods only when we considered the part of the test sets where all the judges agreed. In the next section we present a discussion of these results, limitation of our current studies, and pointers for future work.

Discussion, Limitations, and Future Work

When asking the human subjects to label the sarcastic, positive, and negative utterances, we gave them the opportunity to add comments related to their choice. A qualitative analysis of the comments highlighted that judges considered that the classification task is 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 knowledge of the world 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. Also, the fact that a tweet was explicitly labeled with a #sarcasm hashtag might indicate that the author wanted to make sure that the message is understood as being sarcastic. Without the hashtag the content can be interpreted in its literal (positive) sense ("It was such a great game"). This ambiguity makes distinguishing sarcastic tweets from positive and negative ones difficult for both computational models and humans. In addition, we plan to expand the set of hashtags used for the sarcastic class, such as #irony or #not and test whether systems trained on data using a particular hashtag generalizes to data labeled with different hashtags.

Our results (both computational and human performance) suggest that to improve sarcasm identification, the context of the utterance will need to be taken into account. For example, ToUser was one of the key features in distinguishing sarcastic tweets from both positive and negative tweets. Furthermore, human judges often attributed difficulty in assigning a category to the lack of context. The following example shows that context is indeed used to clarify the intention to convey sarcasm.

> A: What do you do again?
> B: I'm a Concierge specialist. Basically handled escalated issues from the Call center!
> A: Sounds fun
> B: Its ok. Not necessarily fun, but eventful!
> A: I was being very sarcastic

Speaker A responds *Sounds fun*, referring to speaker B's previous utterance about her job as a concierge specialist. Speaker B, assuming that speaker A really means what she said, replies "*Not necessarily fun, but eventful*." Faced with this misunderstanding, speaker A clarifies her previous utterance by saying "*I was being very sarcastic*." Taking into account the entire conversation and detecting the utterance/turn where the speaker who employed sarcasm clarifies what she originally meant, may provide additional cues to identify sarcastic utterances correctly. In addition to conversation context, in future work we will explore user characteristics such as gender (Tepperman et al., 2006) and prior utterances (some users tend to be sarcastic, while others do not). We will also investigate characteristics of the user's social network. Recent work on exploring users' social networks for sentiment analysis obtained mixed results and more work needs to be pursued in this area, but we strongly believe that a user's characteristics and the social context could be very useful information for sarcasm detection.

Conclusions

In this article we have taken a closer look at the problem of automatically detecting sarcasm in Twitter messages. The contributions of this article include (a) the creation of a corpus that includes sarcastic utterances that have been explicitly identified as such by the writer of the message; (b) a report on the performance of distinguishing sarcastic tweets from tweets that are straightforwardly positive or negative both by automatic methods (supervised machine-learning methods) and by human subjects. 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 the three pragmatic features—ToUser, smiley, and frown—were among the 10 most discriminating features in the classification tasks. (Table 1). We also show that using more training data helps increase classifier performance for all the tasks (S-NS, S-P-N, S-P, and S-N).

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 performance is still relatively weak. Our results demonstrate the difficulty of sarcasm classification both for humans and for machine-learning methods.

Acknowledgments

Parts of this paper are based on work supported by the DARPA-DEFT program. The views expressed are ours and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We would like to thank the anonymous reviewers for helpful comments.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment analysis of Twitter data*. In Proceedings of the Workshop on Languages in Social Media (LSM'11).

Attardo, S. (2007). Irony as relevant inappropriateness. In R. Gibbs & H. Colston (Eds.), Irony in language and thought (pp. 135–174). New York: Erlbaum.

Barbosa, L., & Feng, J. (2010). *Robust sentiment detection on twitter from biased and noisy data*. In Proceedings of the 23rd International Conference onComputational Linguistics: Posters, pp. 36–44.

Becker, H., Naaman, M., & Gravano, L. (2011). *Beyond trending topics: Real-world event identification on Twitter*. In Proceedings, International Conference on Weblogs and Social Media. (ICWSM 2011), July 2011, Barcelona, Spain.

Bermingham, A., & Smeaton, A. (2010). *Classifying sentiment in microblogs: Is brevity an advantage is brevity an advantage?* In Proceedings of CIKM. pp. 1833–1836.

Bollen, J., Pepe, A., & Mao, H. (2011). *Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena*. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 450–453.

Carvalho, P., Sarmento, S., Silva, M.J., & de Oliveira, E. (2009). *Clues for detecting irony in user-generated contents: Oh . . .!! It's "so easy";-)*. In Proceedings of the 1st International CIKM workshop on Topic-Sentiment Analysis for Mass Opinion (TSA '09). ACM, New York, NY, USA, pp. 53–56.

Clark, H., & Gerrig, R. (1984). On the pretence theory of irony. Journal of Experimental Psychology. General, 113, 121–126.

Colston, H. (2007). On necessary conditions for verbal irony comprehension. In R. Gibbs & H. Colston (Eds.), Irony in language and thought (pp. 97–134). New York: Erlbaum.

Councill, I.G., McDonald, R., & Velikovich, L. (2010). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP '10.

Davidov, D., Tsur, O., & Rappoport, A. (2010). *Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon, Dmitry Proceeding of Computational Natural Language Learning (ACL-CoNLL)*. Uppsala, Sweden, July 15–16, 2010. pp. 107–116.

Derks, D., Bos, A.E.R., & Grumbkow, J.V. (2008). Emoticons and online message interpretation. Social Science Computer Review, 26(3), 379–388.

Ekman P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3/4):169–200, 1992.

Filatova, E. (2012). *Irony and sarcasm: Corpus generation and analysis using crowdsourcing*. In Proceedings of LREC, Istanbul, Turkey.

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

Gibbs, R. (2007). Irony in talk among friends. In R. Gibbs & H. Colston (Eds.), Irony in language and thought (pp. 339–360). New York: Erlbaum.

Gibbs, R.W., & Colston, H.L. (Eds.). (2007). Irony in language and thought. New York: Routledge, Taylor and Francis.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Technical report, Stanford, CA.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). *Identifying sarcasm in Twitter: A closer look*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers, Portland, Oregon, June 19–24, 2011. pp. 581–586.

Hao, L., Yu, C., Heng, J., & Smaranda, M. (2012). Combining Social Cognitive Theories with Linguistic Features for Multi-genre Sentiment Analysis. Proceedings of 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 26).

Kreuz, R.J., & 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, NY: Association for Computational.

Kreuz, R.J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. Journal of Experimental Psychology. General, 118, 374–386.

Lamb, A., Paul, M.J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 789–795, June 2013.

Miller, G.A. (1995). WordNet: a lexical database for English. Commununications of the ACM 38(11) (November 1995), 39–41.

Nigam, K., & Hurst, M. (2006). Towards a robust metric of polarity. In J.G. Shanahan, Y. Qu, & J. Wiebe (Eds.), Computing attitude and affect in text: Theory and applications (pp. 265–279). Springer Netherlands. Retrieved from http://dx.doi.org/10.1007/1-4020-4102-0_20

O'Connor, B., Balasubramanyan, R., Routledge, B.R., & Smith, N.A. (2010). *From Tweets to polls: Linking text sentiment to public opinion time series*. In Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

Pak, A., & 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., & Lee, L. 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2, 1–2 (January 2008).

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of LIWC2007. [Software manual]. Austin, TX (www.liwc.net)

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.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to Twitter. Los Angeles: NAACL. June 2010.

Ramage, D., Dumais, S., & Liebling, D. (2010). *Characterizing microblogs with topic models*. In Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

Ramnath, B., Cohen, W.W., Pierce, D., & Redlawsk, D.P. (2011). What pushes their buttons?: predicting comment polarity from the content of political blog posts. In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 12–19.

Reyes, A., & Rosso, P. (2011). *Mining subjective knowledge from customer reviews: A specific case of irony detection*. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '11). Association for Computational Linguistics, Stroudsburg, PA, pp. 118–124.

Ritter, A., Cherry, C., & Dolan, B. (2010). *Unsupervised modeling of Twitter conversations*. In Proceedings of HLT-NAACL 2010.

Sadilek, A., Kautz, H., & Silenzio, V. (2012). *Modeling spread of disease from social interactions*. In Sixth AAAI International Conference on Weblogs and Social Media (ICWSM).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: Real-time event detection by social sensors*. In WWW, New York.

Signorini, A., Segre, A.M., & Polgreen, P.M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza a H1N1 pandemic. PLoS ONE, 6(5), e19467.

Strapparava, C., & 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., & 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.

Walker, M.A., Anand, P., Fox Tree, J.E., Abbott, R., & King, J. (2012). *A corpus for research on deliberation and debate*. In Proceedings of Language Resources and Evaluation Conference (LREC) 2012.