Information Processing and Management xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Signaling sarcasm: From hyperbole to hashtag

Florian Kunneman [a], Christine Liebrecht [b], Margot van Mulken [a], Antal van den Bosch [a,*]

[a] Centre for Language Studies, Radboud University Nijmegen, The Netherlands
[b] The Amsterdam School of Communication Research ASCoR, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form 6 March 2014
Accepted 24 July 2014
Available online xxxx

Keywords:
Social media
Automatic sentiment analysis
Opinion mining
Sarcasm
Verbal irony

ABSTRACT

To avoid a sarcastic message being understood in its unintended literal meaning, in micro-texts such as messages on Twitter.com sarcasm is often explicitly marked with a hashtag such as '#sarcasm'. We collected a training corpus of about 406 thousand Dutch tweets with hashtag synonyms denoting sarcasm. Assuming that the human labeling is correct (annotation of a sample indicates that about 90% of these tweets are indeed sarcastic), we train a machine learning classifier on the harvested examples, and apply it to a sample of a day's stream of 2.25 million Dutch tweets. Of the 353 explicitly marked tweets on this day, we detect 309 (87%) with the hashtag removed. We annotate the top of the ranked list of tweets most likely to be sarcastic that do not have the explicit hashtag. 35% of the top-250 ranked tweets are indeed sarcastic. Analysis indicates that the use of hashtags reduces the further use of linguistic markers for signaling sarcasm, such as exclamations and intensifiers. We hypothesize that explicit markers such as hashtags are the digital extralinguistic equivalent of non-verbal expressions that people employ in live interaction when conveying sarcasm. Checking the consistency of our finding in a language from another language family, we observe that in French the hashtag '#sarcasme' has a similar polarity switching function, be it to a lesser extent.

1. Introduction

In the general area of sentiment analysis, sarcasm is a disruptive factor that causes the polarity of a message to flip. Unlike a simple negation, a sarcastic message often conveys a negative opinion using only positive words – or even intensified, hyperbolic positive words. Likewise, but less frequently, sarcasm can flip the polarity of an opinion with negative words to the intended positive meaning. The detection of sarcasm is therefore important, if not crucial, for the development and refinement of sentiment analysis systems, but is at the same time a serious conceptual and technical challenge.

In this article we introduce a sarcasm detection system for tweets, messages on the microblogging service offered by Twitter.[1] In doing this we are helped by the fact that sarcasm appears to be a commonly recognized concept by many Twitter users, who explicitly mark their sarcastic messages by using hashtags such as '#sarcasm' or '#not'. Hashtags in tweets are explicitly marked keywords, and often act as categorical labels or metadata in addition to the body text of the tweet (Chang, 2010). By using the explicit hashtag any remaining doubt a reader may have is taken away: the message is not to be taken literally; it is sarcastic.

* Corresponding author. Address: Centre for Language Studies/CIW, Faculty of Arts, Radboud University Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands. Tel.: +31 24 3611647.
E-mail address: a.vandenbosch@let.ru.nl (A. van den Bosch).

[1] http://www.twitter.com.

http://dx.doi.org/10.1016/j.ipm.2014.07.006
0306-4573/

While such hashtags primarily function as conversational markers of sarcasm, they can be leveraged as annotation labels in order to generate a model of sarcastic tweets from the text co-occurring with these hashtags. A clear advantage of this approach is the easy acquisition of a vast amount of training data. On the other hand, its performance is dependent on the correctness of two assumptions: first that users who include one of the selected hashtags in their tweet actually intended to convey sarcasm and indeed intended to flip the polarity of the message, and second that the pattern of sarcasm in a tweet still holds when the hashtag is excluded from it as a training label. We set out to test these assumptions along with the quality of the resulting sarcasm detection system by applying it on a realistically large and unbiased sample of tweets (of which the vast majority is non-sarcastic) posted on the same day.

The hashtag as a marker of sarcasm has been leveraged in previous research to detect sarcasm in tweets (González-Ibánez, Muresan, & Wacholder, 2011; Reyes, Rosso, & Buscaldi, 2012). One contribution of this paper to the existing body of work is that a sarcasm classifier is trained on several markers of sarcasm in tandem, the most frequent being '#not', and performance is assessed on a realistically large and unbiased sample of tweets. Furthermore, we provide insight into the role of hyperbole in sarcastic tweets, and we perform a cross-lingual comparison of the use of sarcasm in Twitter by annotating French tweets ending with '#sarcasme'.

This paper is structured as follows. In Section 1.1 we discuss the concepts of sarcasm, the broader category of verbal irony, and their communicative function according to the literature. In Section 2 we offer a brief survey of related work on the development of automatic detectors of polarity in social media. Our experimental setup is described in Section 3. We report on the results of our experiments in Section 4 and analyse our results in view of the theoretical work discussed earlier in Section 5. We summarize our results, draw conclusions, and identify points for future research in Section 6.

1.1. Definitions

Twitter members mark their sarcastic messages with different hashtags. As described in more detail in Section 3.1, we find that four words tend to be used as hashmarks in sarcastic posts: '#sarcasm', '#irony', '#cynicism' and '#not'. Although sarcasm, irony and cynicism are not synonymous, they have much in common. This is especially true for sarcasm and irony; many researchers treat those phenomena as strongly related (Attardo, 2007; Brown, 1980; Gibbs & O'Brien, 1991; Kreuz & Roberts, 1993; Mizzau, 1984; Muecke, 1969), and sometimes even equate the terms in their studies in order to work with a usable definition (Grice, 1978; Tsur, Davidov, & Rappoport, 2010). Cynicism is more mocking and tells us more about human beliefs than irony and sarcasm (Eisinger, 2000), but there is a close correlation between these concepts (Yoos, 1985). The hashtag '#not' is not the name of a rhetorical device or trope such as sarcasm, irony and cynicism, but it is a conventionally used meta-communication marker to indicate the message contains a shift in evaluative valence.

In psycholinguistics and cognitive linguistics sarcasm has been widely studied, often in relation with concepts such as cynicism, and with verbal irony as a broader category term. A brief overview of definitions, hypotheses and findings from communication studies regarding sarcasm and related concepts may help clarify what the hashtags convey.

In this study, we are interested in sarcasm as a linguistic phenomenon, and how we can detect it in social media messages. Yet, Brown (1980) warns that sarcasm 'is not a discrete logical or linguistic phenomenon' (p. 111), while verbal irony is. Indeed, Reyes and Rosso (2012) see sarcasm 'as specific extension[s] of a general concept of irony' (p. 755). In line with the extensive use of #sarcasm in tweets to mark verbal irony, we take the liberty of using the term sarcasm while verbal irony would be the more appropriate term. Even then, according to Gibbs and Colston (2007) the definition of verbal irony is still a 'problem that surfaces in the irony literature' (p. 584).

There are many different theoretical approaches to verbal irony. It should (a) be evaluative, (b) be based on incongruence of the ironic utterance with the co-text or context, (c) be based on a reversal of valence between the literal and intended meaning, (d) be aimed at some target, and (e) be relevant to the communicative situation in some way (Burgers, Van Mulken, & Schellens, 2011). Although it is known that irony is always directed at someone or something (the sender himself, the addressee, a third party, or a combination of the three, see Burgers et al. (2011) and Livnat (2004)) and irony is used relatively often in dialogic interaction (Gibbs, 2007), these two elements of irony are hardly examinable in the case of Twitter: the context of the Twitter messages is missing and it is inconvenient to investigate interaction. Therefore, it is hard to interpret the communicative situation and the target of the message. However, it is possible to analyse texts, such as tweets, on their evaluative meaning and a potential valence shift in the same way as Burgers et al. (2011) did. Burgers et al.'s own definition of verbal irony is 'an utterance with a literal evaluation that is implicitly contrary to its intended evaluation.' (p. 190).

Thus, a sarcastic utterance involves a shift in evaluative valence, which can go two ways: it could be a shift from a literally positive to an intended negative meaning, or a shift from a literally negative to an intended positive evaluation. Since Reyes, Rosso, and Veale (2013) also argue that users of social media often use irony in utterances that involve a shift in evaluative valence, we use the definition of verbal irony of Burgers et al. (2011) in this study on sarcasm, and we use both terms synonymously. The definition of irony as saying the opposite of what is meant is commonly used in previous corpus-analytic studies, and is reported to be reliable (Kreuz, Roberts, Johnson, & Bertus, 1996; Leigh, 1994; Srinarawat, 2005).

In order to ensure that the addressees detect the sarcasm in the utterance, senders use markers in their utterances. Attardo (2000) states that those markers are clues a writer can give that 'alert a reader to the fact that a sentence is ironical' (p. 7). The use of markers in written and spoken interaction may be different (Jahandarie, 1999). In spoken interaction, sarcasm is often marked with a special intonation (Attardo, Eisterhold, Hay, & Poggi, 2003; Bryant & Tree, 2005; Rockwell, 2007), air quotes (Attardo, 2000) or an incongruent facial expression (Attardo et al., 2003; Muecke, 1978; Rockwell,

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

2003). In written communication, authors do not have such clues at their disposal. Since sarcasm is more difficult to comprehend than a literal utterance (Burgers, Van Mulken, & Schellens, 2012a; Gibbs, 1986; Giora, 2003), it is likely that addressees do not pick up on the sarcasm and interpret the utterances literally. To avoid misunderstandings, writers use linguistic markers for irony (Burgers, Van Mulken, & Schellens, 2012b): tropes (a metaphor, hyperbole, understatement or rhetorical question), schematic irony markers (repetition, echo, or change of register), morpho-syntactic irony markers (exclamations, interjections, or diminutives), or typographic irony markers (such as capitalization, quotation marks and emoticons). Thus, besides hashtags to mark the opposite valence of a tweet, Twitter members may also use linguistic markers. A machine-learning classifier that learns to detect sarcasm should in theory be able to discover at least some of the features that Burgers et al. (2012b) list, if given sufficient examples of all of them in a training phase. While metaphor and understatement may be too complex to discover, exclamations and typographical markers should be easy to learn. Hyperbole, or the 'speaker overstating the magnitude of something' (Colston, 2007, p. 194), may be discovered by the classifier by the fact that it is often linked to words that signal intensity, as we now analyse in more detail.

1.2. Linguistic markers of hyperbole

Especially in the absence of visual markers, sarcastic utterances need strong linguistic markers to be perceived as sarcastic (Attardo et al., 2003), and hyperbole is often mentioned as a particularly strong marker (Kreuz & Roberts, 1995). It may be that a sarcastic utterance with a hyperbole ('fantastic weather') is identified as sarcastic with more ease than a sarcastic utterance without a hyperbole ('the weather is good'). While both utterances convey a literally positive attitude towards the weather, the utterance with the hyperbolic 'fantastic' may be easier to interpret as sarcastic than the utterance with the non-hyperbolic 'good'. Hyperbolic words carry intensity. Bowers (1964) defines language intensity as 'the quality of language which indicates the degree to which the speaker's attitude toward a concept deviates from neutrality' (p. 416). According to Van Mulken and Schellens (2012), an intensifier is a linguistic element that can be removed or replaced while respecting the linguistic correctness of the sentence and context, but resulting in a weaker evaluation. Intensifiers, thus, strengthen an evaluative utterance and could make an utterance hyperbolic. Typical word classes of intensifiers used for hyperbolic expressions, inter alia, are adverbs ('very', 'absolutely') and adjectives ('fantastic' instead of 'good'), in contrast to words which leave an evaluation unintensified, like 'pretty', 'good' and 'nice'. According to Liebrecht (in preparation), typographical elements such as capitals and exclamation marks are also intensifying elements which can create hyperbolic utterances. So, there is an overlap between linguistic elements to intensify and linguistic elements to overstate utterances. It may be that senders use such elements in their tweets to make the utterance hyperbolic, in order to signal sarcasm.

2. Related research

The automatic classification of communicative constructs in short texts has become a widely researched subject in recent years. Large amounts of opinions, status updates, and personal expressions are posted on social media platforms such as Twitter. The automatic labeling of their polarity (to what extent a text is positive or negative) can reveal, when aggregated or tracked over time, how the public in general thinks about certain things. See Montoyo, Martínez-Barco, and Balahur (2012) for an overview of recent research in sentiment analysis and opinion mining.

A major obstacle for automatically determining the polarity of a (short) text are constructs in which the literal meaning of the text is not the intended meaning of the sender, as many systems for the detection of polarity primarily lean on positive and negative words as markers. The task to identify such constructs can improve polarity classification, and provide new insights into the relatively new genre of short messages and microtexts on social media. Previous works describe the classification of emotions (Davidov, Tsur, & Rappoport, 2010a; Mohammad, 2012), irony (Reyes et al., 2013), sarcasm (Davidov, Tsur, & Rappoport, 2010b; González-Ibánez et al., 2011; Tsur et al., 2010), satire (Burfoot & Baldwin, 2009), and humor (Reyes et al., 2012).

Most common to our research are the works by Reyes et al. (2013), Tsur et al. (2010), Davidov et al. (2010b) and González-Ibánez et al. (2011). Reyes et al. (2013) collect a training corpus of ironic tweets labeled with the hashtag '#irony', and train classifiers on different feature sets representing higher-level concepts such as unexpectedness, style, and emotions. The classifiers are trained to distinguish '#irony'-tweets from tweets containing the hashtags '#education', '#humour', or '#politics', achieving F1-scores of around 70. Tsur et al. (2010) focus on product reviews on the World Wide Web, and try to identify sarcastic sentences from these in a semi-supervised fashion. Training data is collected by manually annotating sarcastic sentences, and retrieving additional training data based on the annotated sentences as queries. Sarcasm is annotated on a scale from 1 to 5. As features, Tsur et al. infer patterns from these sentences consisting of high-frequency words and content words. Their system achieves an F1-score of 79. Davidov et al. (2010b) apply a comparable system on a small set of tweets manually annotated on sarcasm, and achieve an F-score of 83. When testing the system on tweets marked with '#sarcasm', the F-score drops to 55. They state that apart from indicating the tone of a tweet, '#sarcasm' might be used as a search anchor and as a reference to a former sarcastic tweet, adding a fair amount of noise to the data. González-Ibánez et al. (2011) aim to distinguish sarcasm from literally positive and negative sentiments in tweets. Tweets belonging to all three categories were collected based on hashtags describing them ('#sarcasm' and '#sarcastic' for sarcastic tweets) and tested through 5-fold cross-validation on a set comprising 900 tweets for each of the three categories. As features they make use of word unigrams

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

Table 1
Overview of the dataset with sarcastic tweets used for training.

	# Tweets after filtering
#not	353,758
#sarcasm	48,992
#irony	3,285
#cynicism	404
Total	406,439

and higher-level word categories. While the classifier achieves an accuracy of only 57%, it outperforms human judgement. González-Ibánez et al. (2011) conclude that the lack of context makes the detection of sarcasm in tweets difficult, both for humans and for machines.

In the works described above, a system is tested in a controlled setting: Reyes et al. (2013) compare irony to a restricted set of other topics, Tsur et al. (2010) take from the unlabeled test set a sample of product reviews with 50% of the sentences classified as sarcastic, and González-Ibánez et al. (2011) train and test in the context of a small set of positive, negative and sarcastic tweets. In contrast, we apply a trained sarcasm detector to a real-world test set representing a realistically large sample of tweets posted on a random day, the vast majority of which is not sarcastic. Detecting sarcasm in social media is, arguably, a needle-in-a-haystack problem: of the 2.25 million tweets we gathered on a single day, 353 are explicitly marked with the #sarcasm or its pseudo-synonyms. It is therefore only reasonable to test a system in the context of a typical distribution of sarcasm in tweets.

3. Experimental setup

3.1. Data

3.1.1. Hashtag selection

As argued in Section 1.1, while '#sarcasm' ('#sarcasme' in Dutch – we use the English translations of our hashtags throughout this article) is the most obvious hashtag for sarcastic tweets, we base our training set on an expanded set of pseudo-synonymous hashtags. We also found empirical evidence that we need to expand the set of hashtags. Liebrecht, Kunneman, and Van den Bosch (2013) reported that training a classifier solely on '#sarcasm' as a training label resulted in high weights for hashtags that have the same function as '#sarcasm': to switch the evaluative valence or give a description of the type of tweet. While Qadir and Riloff (2013) expand sets of hashtags denoting the emotion of a tweet by bootstrapped learning, this approach does not seem appropriate for the more subtle rhetorical instrument of sarcasm. We decided to extract all hashtags from the ranked list of features from the (Liebrecht et al., 2013) study and manually examine the tweets accompanying them by means of twitter.com. From this examination, we selected the hashtags that almost unambiguously denoted sarcasm in a tweet in addition to '#sarcasm': '#irony', '#cynicism', and '#not'. The former two denote tropes comparable to sarcasm, while the latter is also typically used to switch the evaluative valence of a message. Hashtags that only partly overlap in function such as '#joke' or '#haha' were not included due to their ambiguous usage (either shifting the evaluative valence of a message or simply denoting a funny tweet).

3.1.2. Data collection

For the collection of tweets we made use of a database provided by the Netherlands e-Science Centre consisting of IDs of a substantial portion of all Dutch tweets posted from December 2010 onwards (Tjong Kim Sang & Van den Bosch, 2013).[2] From this database, we collected all tweets that contained the Dutch versions of the selected hashtags '#sarcasm', '#irony', '#cynicism', and '#not' until January 31st 2013. This resulted in a set of 644,057 tweets in total. Following Mohammad (2012) and González-Ibánez et al. (2011), we cleaned up the dataset by only including tweets in which the given hashtag was placed at the end or exclusively followed by other hashtags or a url. Hashtags placed somewhere in the middle of a tweet are more likely to be a grammatical part of the sentence than a label (Davidov et al., 2010b), and may refer to only a part of the tweet. Additionally, we discarded re-tweets (repostings of an earlier tweet by someone else). Applying these filtering steps resulted in 406,439 tweets in total as training data. Table 1 offers more details on the individual hashtags; '#not' occurs a factor more frequently than '#sarcasm', which in turn occurs a factor more frequently than '#irony';'#cynicism' again occurs a factor less frequently.

We trained a classifier on sarcastic tweets by contrasting them against a background corpus. For this, we took a sample of tweets in the period from October 2011 until September 2012 (not containing tweets with any of the sarcastic hashtags). To provide the classifier with an equal number of cases for the sarcasm and background categories and thus produce a training set without class skew, 406,439 tweets were selected randomly, equal to the amount of sarcastic tweets. Again, we did not include re-tweets in the sample.

[2] http://twiqs.nl/.

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

Table 2
Retrieval of sarcastic tweets from the testset of 2/1/2013 (TPR = True Positive Rate, FPR = False Positive Rate, AUC = Area Under the Curve).

Class	# Trainingdata	TPR	FPR	AUC	Samples	Classifications	Correct
Background	406,439	0.83	0.13	0.85	2,246,551	1,870,760	1,870,714
Sarcasm	406,439	0.87	0.17	0.85	353	376,144	307

To test our classifier in a realistic setting, we collected a large sample of tweets posted on a single day outside the time frame from which the training set is collected, namely February 1, 2013. After removal of re-tweets, this set of tweets contains approximately 2.25 million tweets, of which 353 carry one of the sarcasm hashtags at the end.

While the distribution of sarcastic versus other tweets on the test day is highly imbalanced, we chose not to copy this distribution to the training stage. As pointed out by Chawla, Japkowicz, and Kotcz (2004), in an imbalanced learning context classifiers tend to be overwhelmed by the large classes and ignore the small ones. To avoid the influence of class size, we decided to completely balance the tweets with and without '#sarcasm' in the training set. This is likely to drive the classifier to overshoot its classification of tweets as sarcastic in the testset, but we consider only the top of its confidence-based ranking; in our evaluation of the ability of the classifier to detect sarcasm in tweets that lack an explicit hashtag, we evaluate the ranking of the classifier with precision at $n < 250$.

3.2. Winnow classification

All collected tweets were tokenized.[3] Punctuation, emoticons, and capitalization information were kept, as these may be used to signal sarcasm (Burgers et al., 2012b). We made use of word uni-, bi- and trigrams as features (including punctuation and emoticons as separate words). User names and URLs were normalized to 'USER' and 'URL' respectively. We removed features containing one of the hashtags from the training set. Finally, we removed terms that occurred three times or less or in two tweets or less.

As classification algorithm we made use of Balanced Winnow (Littlestone, 1988) as implemented in the Linguistic Classification System.[4] This algorithm is known to offer state-of-the-art results in text classification, and produces interpretable per-class weights that can be used to, for example, inspect the highest-ranking features for one class label. The α and β parameters were set to 1.05 and 0.95 respectively. The major threshold ($\theta+$) and the minor threshold ($\theta-$) were set to 2.5 and 0.5. The number of iterations was bounded to a maximum of three.

3.3. Evaluation

To evaluate the outcome of our machine learning experiment we ran two evaluations. The first evaluation focuses on the 353 tweets in the test set ending with one of the selected sarcasm-hashtags, among 2.25 million other non-sarcastic tweets. We measured how well these tweets were identified using the true positive rate (TPR, also known as recall), false positive rate (FPR) and their joint score, the area under the curve (AUC). AUC is a common evaluation metric that is argued to be more resistant to skew than F-score, due to relying on FPR rather than precision (Fawcett, 2004).

For the second evaluation we manually inspect the test tweets identified by the classifier as sarcastic, but that do not carry any of the sarcastic hashtags. While they would be labeled as false positives in the first evaluation, the absence of one of these hashtags does not necessarily imply the tweet is non-sarcastic. In fact, the proper detection of sarcastic tweets not explicitly marked as such with a hashtag would be the ideal functionality of our classifier. For this evaluation we make use of the classifier's characteristic to assign per-instance scores to each label, which can be seen as its confidence in that label. We rank its predictions by the classifier's confidence on the 'sarcasm' label and inspect manually which of the top-ranking tweets is indeed sarcastic. Based on this manual annotation we can compute the precision at different rank numbers, which may reveal whether the top-ranked false positives are in fact sarcastic tweets.

4. Results

Results for the first evaluation are displayed in Table 2. Of the 353 tweets explicitly marked with the '#sarcasm' or its pseudo-synonyms on the test day, 307 (87%) are identified as sarcastic, in addition to 376,144 tweets not containing such a hashtag. Because this latter amount is not a big part of the 2.25 million tweets in the test set, the FPR is fairly low and a good AUC of 0.85 is achieved.

Besides generating an absolute winner-take-all classification, our Balanced Winnow classifier assigns scores to each label that can be seen as its confidence in that label. We can rank its predictions by the classifier's confidence on the 'sarcasm' label and inspect manually which of the top-ranking tweets that do not contain any of the four target hashtags is indeed sarcastic. We generated a list of the 250 most confident 'sarcasm'-labeled tweets. Three annotators (three of the authors of this paper)

[3] Tokenization was carried out with Ucto, http://ilk.uvt.nl/ucto.
[4] http://www.phasar.cs.ru.nl/LCS/.

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

judged these tweets as being either sarcastic or not. The instructions beforehand were to positively annotate tweets that were clearly expressing a positive or negative valence that is shifted by the language use. In case of doubt, for example due to the lack of context, a tweet should be annotated as non-sarcastic. The sarcasm should be clear from the text in a tweet; the annotator was not allowed to enquire into the conversational context when a tweet was addressed to one or more twitter users.

When taking the majority vote of the three annotators as the golden label, a curve of the precision at all points in the ranking can be plotted. This curve is displayed in Fig. 1. The overall performance at the end of the plotted curve is about 0.35. After peaking at a precision of 0.6 after 10 tweets, precision decreases rapidly before stabilizing after rank 50. Precision scores are lower if sarcasm is only labeled with unanimous agreement between the annotators, ending below 0.3.

In order to test for intercoder reliability, Cohen's Kappa was used. In line with Siegel and Castellan (1988), we calculated a mean Kappa based on pair-wise comparisons of all possible coder pairs. The mean intercoder reliability between the three possible coder pairs is moderate at $\kappa = .53$. The average mutual F-score over all annotator pairs is 0.72, indicating that annotators disagree in about a quarter of all cases.

5. Analysis

5.1. Reliability of the training set

An important additional check on our results concerns the reliability of the user-generated sarcastic hashtags as golden labels, as Twitter users cannot all be assumed to understand what sarcasm is, or be versed in using tropes. The three annotators who annotated the ranked classifier output also coded a random sample of 250 tweets with sarcastic hashtags from the training set. The tweets were sampled proportional to the percentage of the four hashtags in the training set (e.g.: 162 '#not'-tweets, 86 '#sarcasm'-tweets and 2 '#irony'-tweets). The instructions beforehand were to decide whether a tweet contains a positive or negative valence, which is shifted by means of the hashtag at the end.

The average score of agreement between the three possible coder pairs turned out to be moderate ($\kappa = .44$), but due to the majority of the tweets being genuinely sarcastic, the mutual F-score between the annotators is 0.94, indicating a disagreement on a fairly random 6% of cases. Taking the majority vote over the three annotations as the reference labeling, 212 of the 250 annotated sarcastic tweets, about 90%, were found to be sarcastic. Using hashtags as golden labels thus introduces about 10% noise into the labeled training data. The outcome of the annotated tweets for '#not' and '#sarcasm' separately is in balance, with respective scores of 90% and 91%. These outcomes are in line with a similar annotation that was conducted in Liebrecht et al. (2013), sampling 250 tweets that were all labeled with the hashtag '#sarcasm'. Of these tweets, 85% were judged as being actually sarcastic.

5.2. Predictors of a sarcastic tweet

While the classifier performance gives an impression of its ability to detect sarcastic tweets, the strong indicators of sarcasm as discovered by the classifier may provide additional insights into the usage of sarcasm by Twitter users. Including only word unigrams, bigrams, and trigrams as features brings about an unbiased classifier model to be analysed. We set out to analyse the feature weights assigned by the Balanced Winnow classifier ranked by the strength of their connection to the sarcasm label, taking into account the 500 tokens and n-grams with the highest positive weight towards the sarcasm class. These words and n-grams provide insight into the topics Twitter users are talking about. Even though Liebrecht et al.

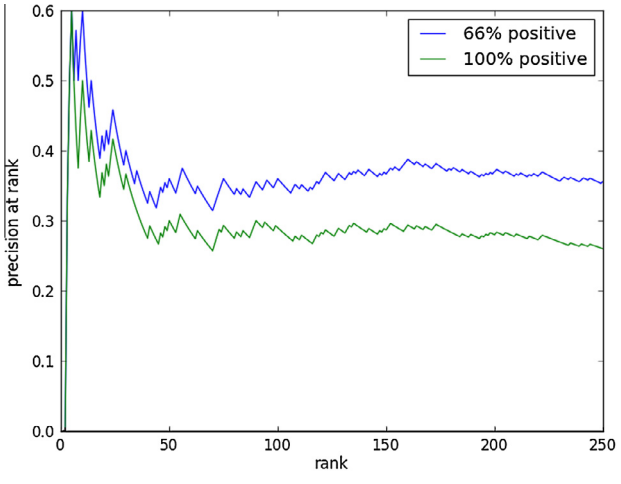

Fig. 1. Precision at $\{1 \ldots 250\}$ on the sarcasm class.

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

(2013) reported on topical words appearing as strong predictors, relating to school, the weather, holidays, public transport, soccer, and television programs, in our current study, which is based on a significantly larger training set, such topics are hardly present in the top-500. The 500 words and *n*-grams are mostly adverbs and adjectives that realize a positive evaluation (including intensifiers), exclamations, and non-sarcastic hashtags for meta-communication.

We suspected that the sarcastic utterances contained many intensifiers to make the tweets hyperbolic. The list of strongest predictors shows that some intensifiers are indeed strong predictors of sarcasm, such as (with and without capitals) *geweldig* (awesome), *heerlijk* (lovely), *prachtig* (wonderful), *boeiend* (fascinating), *allerleukste* (most fun), *perfect*, and *super*. However, besides these intensifiers many unintensified positive adverbs and adjectives occur in the list of strongest predictors as well, such as *interessant* (interesting), *gezellig* (cozy), *leuk* (fun), *handig* (handy), *slim* (smart), *charmant* (charming) and *nuttig* (useful). Considerably less negative words occur as strong predictors. This supports our hypothesis that the utterances are mostly positive, while the opposite meaning is meant. This finding corresponds with the results of Burgers et al. (2012b), who show that 77% of the ironic utterances in Dutch communication are literally positive. It also concurs with the observation that a sarcastic utterance always implies an evaluation: these (positive) adverbs and adjectives explicitly indicate (and thus mark) that the sender intentionally conveys an attitude towards his or her message.

A substantial set of positive exclamations are found by the classifier as strong predictors. Exclamations are another means to make an utterance hyperbolic and thereby sarcastic. Examples of Dutch exclamations within the top-500 of most predictive features are (with and without # or capitals): *jippie, yes, goh, joepie, jeej, jeuj, yay, woehoe*, and *wow*.

The fourth group of features in the top-500 are non-sarcastic hashtags that signal meta-communication, such as '#humor', '#lml' (love my life), '#wehebbenerzinin' (we are looking forward to it), '#gaatgoed' (all is well), '#bedankt' (thanks), and '#grapje' (joke).

To inspect in more detail the actual occurrence of the four types of words that constitute the top-500 of most predictive features, we further analyse the sarcastic tweets without sarcastic hashtags that our classifier correctly identifies in the top-250 ranked tweets of our test day, and contrast this with the tweets in our training set that do have a sarcastic hashtag. Fig. 2 displays two proportional Venn diagrams of occurrences in these two sets of tweet of the four aforementioned categories of markers: intensified and unintensified adverbs and adjectives, exclamations, and non-sarcastic hashtags. The Venn diagram on the right in Fig. 2 visualizes the proportions of tweets without a hashtag correctly identified as being sarcastic that have one or more of these four categories, or none of them (the circle labeled 'None'). The overlap between circles in the Venn diagram visualize which proportion of tweets have a combination of two or more of the four marker categories. The left diagram represents all tweets in the training set with a sarcastic hashtag, while the right diagram represents sarcastic tweets without a sarcastic hashtag. As can be seen in the figure, most sarcastic tweets without a hashtag have unintensified evaluative words. The other three categories occur less frequently, and some of these occur in combination, the most frequent combination being between unintensified evaluative words and exclamations.

The left diagram differs in three aspects from the right diagram: first, the number of tweets containing none of the major linguistic sarcasm markers is the largest category; second, the four categories almost never co-occur. The overall third observation is that the presence of a sarcastic tag in the training tags appears to mute the occurrence of non-sarcastic hashtags. These differences suggests that the presence of an explicit sarcasm hashtag requires fewer other clues to signify sarcasm.

5.3. #sarcasme in French tweets

We have shown that a polarity shift between the actual and intended valence of a message can to a certain extent be recognized automatically in the case of Dutch tweets, by means of hashtag labels. This complements previous findings for

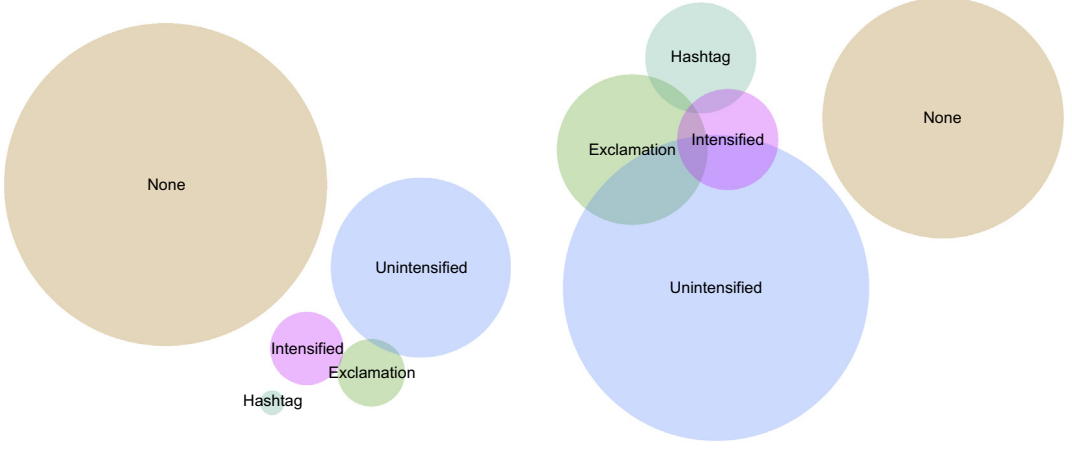

Fig. 2. Proportional Venn diagrams of the co-occurrences of linguistic markers and hashtags in all sarcastic training tweets (left) and detected sarcastic test tweets without a hashtag #sarcasm (right).

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

English tweets. Thus, in both languages such hashtags are predominantly applied in the same way. Future research would be needed to chart the prediction of sarcasm in languages that are more distant to Dutch. As the findings in this analysis suggests, sarcasm may be signaled rather differently in other cultures (Goddard, 2006). Languages may use the same type of marker in different ways, like a different intonation in spoken sarcasm by English and Cantonese speakers (Cheang & Pell, 2009). Such a difference between languages in the use of the same marker may also apply to written sarcastic utterances, such as tweets.

To investigate the potential success of leveraging hashtag marked tweets in other language regions, we set out to annotate a sample of 500 French tweets ending with #sarcasme (French for 'sarcasm'). The tweets were harvested from topsy.com, by means of the otter API.[5] We queried tweets containing #sarcasme, setting the language to French and including all days in 2012 and 2013. This resulted in 8301 tweets. From this sample, we removed retweets and tweets that did not end in #sarcasme, and took a random sample of 500 tweets. The tweets were annotated by one of the authors and a second person. Both annotators were L1 speakers of Dutch with a French L2 near-native proficiency. The instructions beforehand were the same as for the annotation of the Dutch #sarcasm tweets sampled from the training data.

The annotators marked 63% of the tweets both as sarcastic, with a moderate κ of .43 and a mutual F-score of .85. The percentage of sarcastically marked tweets by both annotators is smaller than the 90% attained with the Dutch tweets. When we split the three annotators of the Dutch tweets in pairs of two, allowing a better comparison with the two annotators of the French tweets, the percentages are also higher (.85, .82, and .84).

Speakers of French seem to be more lenient with the hasthag #sarcasme than speakers of Dutch (or English for that matter), because they also use it to signal other rhetorical figures, such as paradoxes, rhetorical questions and other types of humor. Since the instruction explicitly asked annotators to look for a shift in evaluative valence for a tweet being labeled as sarcastic, the percentage of polarity shifting tweets is accordingly lower. Moreover, the number of tweets without an explicit evaluation was also considerable. Apparently, users of French in tweets more heavily rely on context (and on the receiver being able to interpret the tweet correctly) than Dutch or English users do. The difference between Dutch and French sarcastic tweets suggests that culture also influences the use and reception of sarcasm (and especially the use of '#sarcasme'). This is in line with Holtgraves (2005) who also argues that the use and interpretation of non-literal meanings can be culture-specific.

6. Conclusion

In this study we developed and tested a system that detects sarcastic tweets in a realistic sample of 2.25 million Dutch tweets posted on a single day, trained on a set of 406 thousand tweets, harvested over time, marked by the hashtags '#sarcasm', '#irony', '#cynicism', or '#not' by the senders, plus 406 thousand tweets without these tags. The classifier attains an AUC score of .84 and is able to correctly spot 309 of the 353 tweets among the 2.25 million that were explicitly marked with the hashtag, with the hashtag removed. Testing the classifier on the top 250 of the tweets it ranked as most likely to be sarcastic, but that did not have a sarcastic hashtag, it attains only a 35% average precision. We can conclude that it is fairly hard to distinguish sarcastic tweets from literally intended tweets in an open setting, though the top of the classifier's ranking does identify many sarcastic tweets not explicitly marked with a hashtag.

An additional linguistic analysis provides some insights into the characteristics of sarcasm on Twitter. We found that most tweets contain a literally positive message, and contain four types of markers for sarcasm: intensified as well as unintensified evaluative words, exclamations, and non-sarcastic hashtags. Intensified evaluative words and exclamations induce hyperbole, but they occur less frequently in sarcastic tweets than unintensified evaluative words. Note that we based our selection of marker categories on the top-500 of most predictive features; other linguistic markers from Burgers et al. (2012b) did not occur in this set and were not included in this study. The differences between the occurrence or absence of markers displayed in the two Venn diagrams of Fig. 2 indicate that the inclusion of a sarcastic hashtag reduces the use of linguistic markers that otherwise would be needed to mark sarcasm. Arguably, extralinguistic elements such as hashtags can be seen as the social media equivalent of non-verbal expressions that people employ in live interaction when conveying sarcasm. As Burgers et al. (2012a) show, the more explicit markers an ironic utterance contains, the better the utterance is understood, the less its perceived complexity is, and the better it is rated. Many Twitter users already seem to apply this knowledge.

To investigate the usefulness of a sarcastic hashtag to train sarcasm detection in other language regions, we annotated 500 French tweets containing #sarcasme, finding that the majority of French tweets could indeed be labeled as sarcastic, but to a lesser extent than Dutch tweets. In other words: also in French, the hashtag signals a polarity switch in most cases. Apart from hashtag usage, markers of sarcasm can be language-specific. In Dutch, for example, diminutives can mark irony (Burgers et al., 2012a), while the neighbor language English does not have this device. Dedaić (2005) shows other language-specific markers that have been associated with irony in the Croatian language; as did Bennett-Kastor (1992) for the Ghanese language Sissala. Knowledge of specific sarcasm markers in a language could be used as explicitly added features to our system.

[5] https://code.google.com/p/otterapi/wiki/Resources.

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

Although this study provides insights into the use of sarcasm in Twitter messages and training a machine learning classifier to detect this rhetoric device automatically, we must emphasize that we focused in our analysis on the hyperbole, one of the linguistic markers people use for sarcasm (Burgers et al., 2012b), and we did find evidence for the fact that typical hyperbole inducers, i.e. exclamations and intensifiers, appear among the most predictive features of our classifier. In future work we need to explore means to allow our classifier to detect other linguistic markers, such as rhetorical questions, repetition, echo, change of register, interjections, or diminutives, many of which cannot simply be inferred from the presence of words or *n*-grams of words. Since gender, education and profession can be predictors of the use of irony, just like previous use of irony, we also need to further explore the possibilities of including speaker and context characteristics in the model (Wallace, 2013).

Another strand of future research would be to expand our scope from sarcasm to other more subtle variants of irony, such as understatements, euphemisms, and litotes (Burgers et al., 2012b). Following Giora, Fein, Ganzi, Levi, and Sabah (2005), there seems to be a spectrum of degrees of irony from the sarcastic 'Max is exceptionally bright' via the ironic 'Max is not exceptionally bright', the understatement 'Max is not bright' to the literal 'Max is stupid'. In the first three utterances a gap exists between what is literally said and the intended meaning of the sender. The greater the gap or contrast, the easier it is to perceive the irony. But the negated *not bright* is still perceived as ironic; more ironic than the literal utterance (Giora et al., 2005). We may need to combine the sarcasm detection task with the problem of the detection of negation and hedging markers and their scope (Morante & Daelemans, 2009; Morante, Liekens, & Daelemans, 2008) in order to arrive at a comprehensive account of polarity-reversing mechanisms, which in sentiment analysis is still highly desirable.

References

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics, 32*(6), 793–826.
Attardo, S. (2007). Irony as relevant inappropriateness. In R. W. Gibbs, Jr. & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 135–170). New York, NY: Lawrence Erlbaum.
Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Visual markers of irony and sarcasm. *Humor, 16*(2), 243–260.
Bennett-Kastor, T. (1992). Relevance relations in discourse: A study with special reference to Sissala. *Journal of Linguistic Anthropology, 2*(2), 240–242.
Bowers, J. W. (1964). Some correlates of language intensity. *Quarterly Journal of Speech, 50*(4), 415–420.
Brown, R. L. (1980). The pragmatics of verbal irony. In R. W. Shuy & A. Shnukal (Eds.), *Language use and the uses of language* (pp. 111–127). Washington, DC: Georgetown University Press.
Bryant, G. A., & Tree, J. E. F. (2005). Is there an ironic tone of voice? *Language and Speech, 48*(3), 257–277.
Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 161–164). Association for Computational Linguistics.
Burgers, C., Van Mulken, M., & Schellens, P. J. (2011). Finding irony: An introduction of the verbal irony procedure (vip). *Metaphor and Symbol, 26*(3), 186–205.
Burgers, C., Van Mulken, M., & Schellens, P. J. (2012a). Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics, 44*(3), 231–242.
Burgers, C., Van Mulken, M., & Schellens, P. J. (2012b). Verbal irony. *Journal of Language and Social Psychology, 31*(3), 290–310.
Chang, H.-C. (2010). A new perspective on twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology, 47*, 1–4.
Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6*(1), 1–6.
Cheang, H. S., & Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America, 126*, 1394.
Colston, H. L. (2007). What figurative language development reveals about the mind. *Mental States: Volume 2: Language and Cognitive Structure, 93*, 191.
Davidov, D., Tsur, O., & Rappoport, A. (2010a). Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: Posters. COLING '10* (pp. 241–249). Stroudsburg, PA, USA: Association for Computational Linguistics.
Davidov, D., Tsur, O., & Rappoport, A. (2010b). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the fourteenth conference on computational natural language learning. CoNLL '10* (pp. 107–116). Stroudsburg, PA, USA: Association for Computational Linguistics.
Dedaić, M. N. (2005). Ironic denial: Tobože in Croatian political discourse. *Journal of Pragmatics, 37*(5), 667–683.
Eisinger, R. M. (2000). Questioning cynicism. *Society, 37*(5), 55–60.
Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Tech. Rep. HPL-2003-4, Hewlett Packard Labs.
Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General, 115*(1), 3.
Gibbs, R. W. (2007). On the psycholinguistics of sarcasm. In R. W. Gibbs, Jr. & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 173–200). New York, NY: Lawrence Erlbaum.
Gibbs, R. W., & Colston, H. (2007). Irony as persuasive communication. In R. W. Gibbs, Jr. & H. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 581–595). New York, NY: Lawrence Erlbaum.
Gibbs, R. W., & O'Brien, J. (1991). Psychological aspects of irony understanding. *Journal of Pragmatics, 16*(6), 523–530.
Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford University Press.
Giora, R., Fein, O., Ganzi, J., Levi, N., & Sabah, H. (2005). On negation as mitigation: The case of negative irony. *Discourse Processes, 39*(1), 81–100.
Goddard, C. (2006). lift your game Martina!: Deadpan jocular irony and the ethnopragmatics of Australian English. *Applications of Cognitive Linguistics, 3*, 65.
González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers. HLT '11* (Vol. 2, pp. 581–586). Stroudsburg, PA, USA: Association for Computational Linguistics.
Grice, H. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Pragmatics: Syntax and semantics* (pp. 113–127). New York, NY: Academic Press.
Holtgraves, T. (2005). Context and the comprehension of nonliteral meanings. In H. Colston & A. N. Katz (Eds.), *Figurative language comprehension: Social and cultural influence* (pp. 73–98). Associates, New Jersey: Lawrence Erlbaum.
Jahandarie, K. (1999). *Spoken and written discourse: A multi-disciplinary perspective*. Greenwood Publishing Group.
Kreuz, R. J., & Roberts, R. M. (1993). The empirical study of figurative language in literature. *Poetics, 22*(1), 151–169.
Kreuz, R. J., & Roberts, R. M. (1995). Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbol, 10*(1), 21–31.
Kreuz, R., Roberts, R., Johnson, B., & Bertus, E. (1996). Figurative language occurrence and co-occurrence in contemporary literature. In R. Kreuz & M. MacNealy (Eds.), *Empirical approaches to literature and aesthetics* (pp. 83–97). Norwood, NJ: Ablex.
Leigh, J. H. (1994). The use of figures of speech in print ad headlines. *Journal of Advertising*, 17–33.
Liebrecht, C., Kunneman, F. & Van den Bosch, A. (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* (pp. 29–37).
Liebrecht, C. (in preparation). Intens krachtig. stilistische intensiveerders in evaluatieve teksten. Ph.D. thesis, Radboud University Nijmegen.

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning, 2*, 285–318.
Livnat, Z. (2004). On verbal irony, meta-linguistic knowledge and echoic interpretation. *Pragmatics & Cognition, 12*(1), 57–70.
Mizzau, M. (1984). *L'ironia: la contraddizione consentita*. Milan, Italy: Feltrinelli.
Mohammad, S. M. (2012). #Emotional tweets. In *Proceedings of the first joint conference on lexical and computational semantics – Vol. 1: Proceedings of the main conference and the shared task, and Vol. 2: Proceedings of the sixth international workshop on semantic evaluation, SemEval '12* (pp. 246–255). Stroudsburg, PA, USA: Association for Computational Linguistics.
Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*.
Morante, R., & Daelemans, W. (2009). Learning the scope of hedge cues in biomedical texts. In *Proceedings of the workshop on BioNLP* (pp. 28–36). Association for Computational Linguistics.
Morante, R., Liekens, A. & Daelemans, W. (2008). Learning the scope of negation in biomedical texts. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 715–724).
Muecke, D. C. (1969). *The compass of irony*. Oxford University Press.
Muecke, D. C. (1978). Irony markers. *Poetics, 7*(4), 363–375.
Qadir, A. & Riloff, E. (2013). Bootstrapped learning of emotion hashtags #hashtags4you. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 2–11).
Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems, 53*(4), 754–760.
Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*.
Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation, 47*(1), 239–268.
Rockwell, P. (2003). Empathy and the expression and recognition of sarcasm by close relations or strangers. *Perceptual and Motor Skills, 97*(1), 251–256.
Rockwell, P. (2007). Vocal features of conversational sarcasm: A comparison of methods. *Journal of Psycholinguistic Research, 36*(5), 361–369.
Siegel, S., & Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw Hill.
Srinarawat, D. (2005). Indirectness as a politeness strategy of Thai speakers. In R. Lakoff & S. Ide (Eds.), *Broadening the horizon of linguistic politeness* (pp. 175–193). Amsterdam, The Netherlands: John Benjamins.
Tjong Kim Sang, E., & Van den Bosch, A. (2013). Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal, 3*, 121–134.
Tsur, O., Davidov, D. & Rappoport, A. (2010). Icwsm–a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the fourth international AAAI conference on Weblogs and social media* (pp. 162–169).
Van Mulken, M., & Schellens, P. J. (2012). Over loodzware bassen en wapperende broekspijpen. Gebruik en perceptie van taalintensiverende stijlmiddelen. *Tijdschrift voor taalbeheersing, 34*(1), 26–53.
Wallace, B. C. (2013). Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, 1–17.
Yoos, G. E. (1985). The rhetoric of cynicism. *Rhetoric Review, 4*(1), 54–62.

Please cite this article in press as: Kunneman, F., et al. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management* (2014), http://dx.doi.org/10.1016/j.ipm.2014.07.006