Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis

Diana Maynard, Mark A. Greenwood

University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{diana,m.greenwood}@dcs.shef.ac.uk

Abstract

Sarcasm is a common phenomenon in social media, and is inherently difficult to analyse, not just automatically but often for humans too. It has an important effect on sentiment, but is usually ignored in social media analysis, because it is considered too tricky to handle. While there exist a few systems which can detect sarcasm, almost no work has been carried out on studying the effect that sarcasm has on sentiment in tweets, and on incorporating this into automatic tools for sentiment analysis. We perform an analysis of the effect of sarcasm scope on the polarity of tweets, and have compiled a number of rules which enable us to improve the accuracy of sentiment analysis when sarcasm is known to be present. We consider in particular the effect of sentiment and sarcasm contained in hashtags, and have developed a hashtag tokeniser for GATE, so that sentiment and sarcasm found within hashtags can be detected more easily. According to our experiments, the hashtag tokenisation achieves 98% Precision, while the sarcasm detection achieved 91% Precision and polarity detection 80%.

Keywords: sarcasm, opinion mining, social media

1. Introduction

Sarcasm occurs frequently in user-generated content such as blogs, forums and microposts, especially in English, and is inherently difficult to analyse, not only for a machine but even for a human. One needs to have a good understanding of the context of the situation, the culture in question, and perhaps the very specific topic or people involved in the sarcastic statement. This kind of real-world knowledge is almost impossible for a machine to make use of. Furthermore, even correctly identifying a statement as sarcastic is often insufficient to be able to analyse it, especially in terms of sentiment, due to issues of scope.

Almost all current research on sarcasm detection has only studied the issue of classifying sentences (usually tweets) as sarcastic or not. The French company Spotter recently hit the news for its state of the art sarcasm detection tool, reputed to achieve an 80% success rate in finding sarcastic utterances, as part of its reputation analysis platform[1]. But it is not entirely clear how it then processes them. Furthermore, sarcastic statements are particularly prone to novelty in vocabulary use, which makes it hard to train Machine Learning algorithms to successfully spot them when no external clues (such as hashtags) are present.

In this work, we investigate the use of sarcasm in tweets, and in particular their effect on sentiment analysis. Our early experiments show that correctly detecting sarcasm can improve sentiment detection by nearly 50 percentage points, but that even when a tweet is correctly identified as being sarcastic, accuracy of sentiment analysis is still far from perfect. We perform an analysis of the effect of sarcasm scope on the polarity of tweets. Existing work on sarcasm detection, as discussed in Section 6. focuses only on the presence or absence of sarcasm, and not on how to deal with it in sentiment analysis. As part of the analysis of scope, we focus particularly on hashtags. For this, we have developed a hashtag tokeniser for GATE (Cunningham et al., 2002), such that sentiment and sarcasm found within hashtags can be detected. We have also compiled a number of rules which enable us to improve the accuracy of sentiment analysis when sarcasm is known to be present.

2. Sarcasm and its effect on sentiment

While not restricted to English, sarcasm is an inherent part of British culture: so much so that the BBC has its own webpage on sarcasm designed to teach non-native English speakers how to be sarcastic successfully in conversation[2]. The Oxford English Dictionary defines sarcasm as *"a sharp, bitter, or cutting expression or remark; a bitter gibe or taunt."* However, these days it is generally used to mean a statement when people *"say the opposite of the truth, or the opposite of their true feelings in order to be funny or to make a point"*, as defined on the BBC sarcasm webpage mentioned above. Bousfield (Bousfield, 2007) describes it as *"the use of strategies which, on the surface appear to be appropriate to the situation, but are meant to be taken as meaning the opposite in terms of face management. That is, the utterance which appears, on the surface, to maintain or enhance the face of the recipient actually attacks and damages the face of the recipient. ... sarcasm is an insincere form of politeness which is used to offend one's interlocutor."* Clearly, there are no clear-cut delimiters to sarcasm.

There is much confusion between sarcasm and verbal irony: traditionally, the distinction between the two was that irony was indirect, whereas sarcasm was direct, i.e. that irony involved a meaning opposite to the literal interpretation. The OED defines irony as a kind of special case of sarcasm: *"a figure of speech in which the intended meaning is the opposite of that expressed by the words used; usually taking*

[1] http://www.bbc.co.uk/news/technology-23160583

[2] http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/1210_how_to_converse/page13.shtml

the form of sarcasm or ridicule in which laudatory expressions are used to imply condemnation or contempt." Given the confusion, and the fact that sarcasm has come to take on this idea of "opposite meaning", it is hard to make a distinction, and so like many other researchers, we do not differentiate between the two.

In this work, we therefore define a sarcastic statement as one where the opposite meaning is intended, because this is the dominant usage in such research, and it is also what tends to impact the polarity of the sentiment being expressed. For example, *"I love walking to work in the rain"* would be interpreted with negative sentiment in its sarcastic sense. Furthermore, tweets labelled with the hashtag #irony typically do not refer to verbal irony, but to situational irony, for example:

> *Fat woman, wearing a tracksuit, walking into Greggs #irony*

To validate this, we collected a corpus of 257 tweets containing the hashtag #irony, and found that only 2 tweets contained clear instances of verbal irony, about 25% involved clear situational irony, while about 75% referred to extra-contextual information, so that the meaning was not clear, e.g. *"That was pure irony."*

Concerning its effect on sentiment, a naïve interpretation of sarcasm would blindly assume that it acts similarly to negation: for example, if we have a phrase such as *"this project is great!"*, we would reverse the polarity of the sentiment from positive to negative if we knew that it was meant sarcastically. However, the scope of sarcasm is not always easy to determine. Take the following example:

> *I am not happy that I woke up at 5:15 this morning.. #greatstart #sarcasm*

If we reversed the polarity of the opinion *"not happy"*, we would end up with a positive sentiment, but this is incorrect. The sarcasm applies only to the hashtag #greatstart, and not to the previous sentence.

In the next example, however, the sarcasm refers to the main utterance *"you are really mature"* and not to the other hashtag #lying.

> *You are really mature. #lying #sarcasm*

Furthermore, sarcasm does not always reverse polarity. Take the example

> *It's not like I wanted to eat breakfast anyway. #sarcasm*

When uttered sarcastically, this statement indicates negative sentiment, but without the sarcasm, it does not particularly indicate positive sentiment. In the rest of this paper, we examine the impact of this phenomenon on our system for sentiment analysis.

3. Analysing hashtags

3.1. Hashtag retokenisation

Much useful sentiment information is contained within hashtags, but this is problematic to identify because hashtags are typically tokenised as a single token, although they contain multiple words, e.g. #notreally. We therefore developed an algorithm to extract the individual tokens from the hashtag. First, we try to form a token match against the Linux dictionary which we have converted into a GATE gazetteer. We also try to match against our (limited) existing gazetteers of entity types such as Locations, Organisations etc., and against the dictionary of common slang words used in our Twitter normalisation tool (Bontcheva et al., 2013). After some initial experimentation, we manually edited the slang dictionary to remove most single-character "words" such as "h" (but leaving "real" words such as "i", "a" etc.) and a few other entries that we considered non-words.

Working from left to right, we use a Viterbi-like algorithm to look for the best possible match that combines a set of known words from the lookups, and completes the tokenisation to the end of the hashtag. If a combination of matches can be found without a break, the individual components are converted to tokens and the original single Token annotation covering the whole span of the hashtag is removed.

In our example, *#notreally* would be correctly identified by these rules. Figure 1 shows an example of some retokenised hashtags in GATE: we can see, for example, that *#conflictpalmoil* has been tokenised as "conflict" + "palm" + "oil" (the faint lines in the red shading indicate Token boundaries, while the popup window shows the details of each Token). However, some hashtags are ambiguous: for example, *#greatstart* gets split wrongly into the two tokens "greats" + "tart" rather than "great" + "start". These problems are hard to deal with. We could make use of a language modelling approach based on unigram / bigram frequencies for the ambiguity problem, though there is no guarantee that this would improve results since hashtags are often novel. We could also consider using contextual information to assist, or looking at the POS tags, for example adjective+noun combinations are more probable than verb+noun combinations.

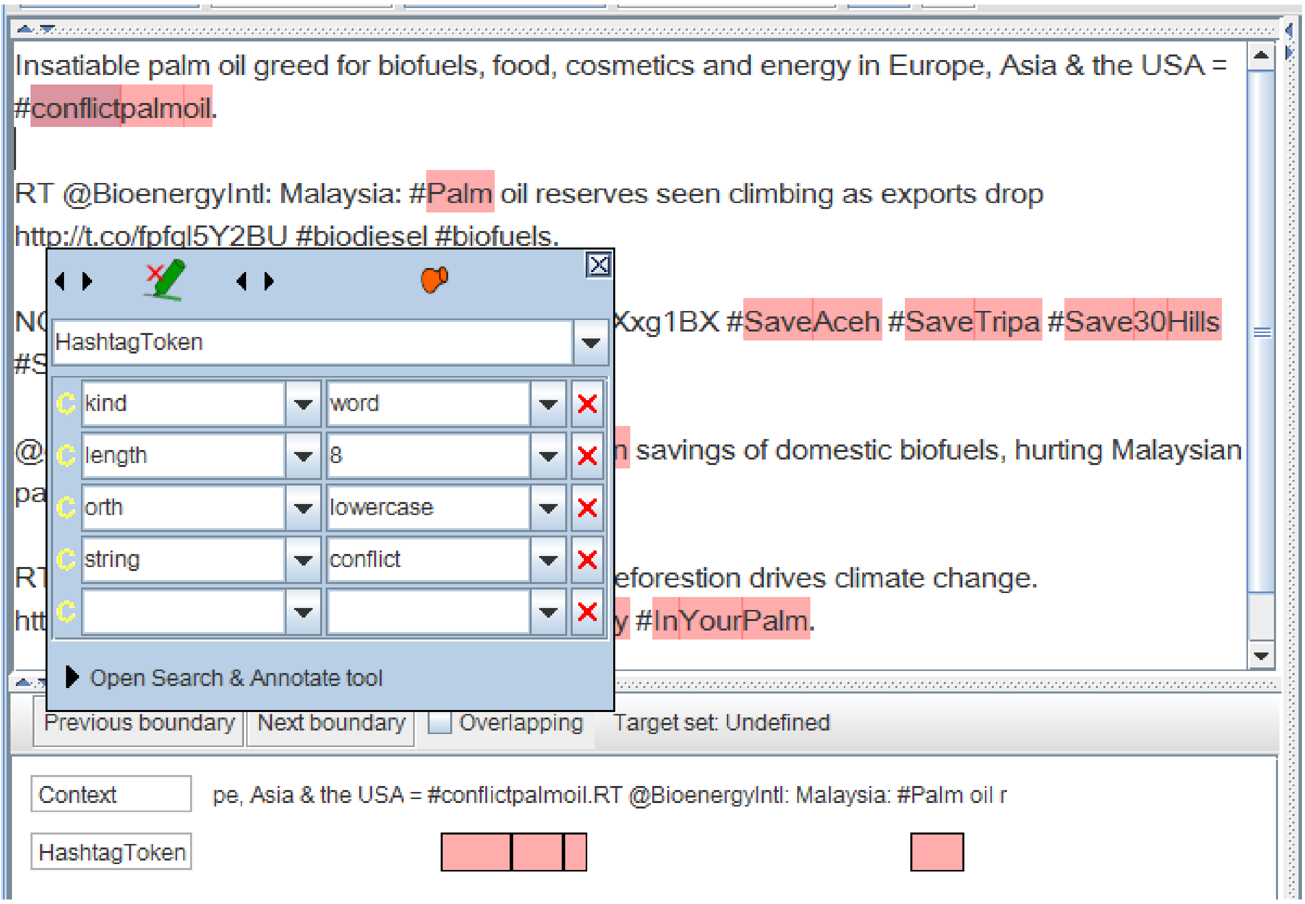
3.2. Using hashtags for sentiment scope detection

Now that we have tokenised (most of) the hashtags correctly, we can make use of the information contained within them for sentiment detection. For example, we can recognise individual words that might be used to denote sarcasm, and we can recognise positive and negative words within a hashtag.

As a first step, we simply reversed the polarity of an opinion whenever a sarcastic statement was found. To identify sarcastic statements, we manually collected a list of sarcastic hashtags from a corpus of random tweets, and then extended this by automatically collecting pairs of hashtags where one hashtag contained an existing sarcasm hashtag (e.g. #sarcasm), using the GazetteerListCollector GATE plugin[3]. For example, the following tweets contain pairs of sarcastic hashtags:

> *I love living with a 7 year old #NotReally #sarcasm'*

[3] http://gate.ac.uk/userguide/sec:gazetteers:listscollector



Insatiable palm oil greed for biofuels, food, cosmetics and energy in Europe, Asia & the USA = #conflictpalmoil.

RT @BioenergyIntl: Malaysia: #Palm oil reserves seen climbing as exports drop http://t.co/fpfql5Y2BU #biodiesel #biofuels.

N(Xxg1BX #SaveAceh #SaveTripa #Save30Hills
#S

@ n savings of domestic biofuels, hurting Malaysian
pa

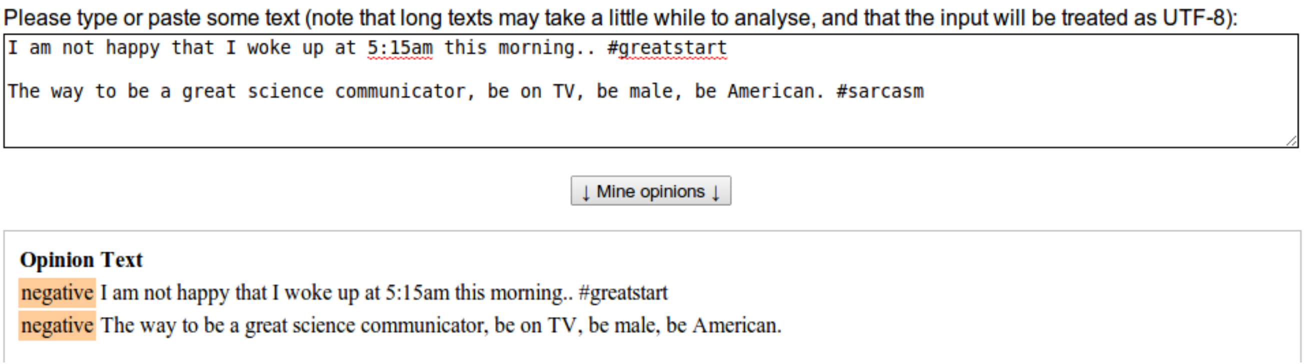
RT eforestion drives climate change.
htt y #InYourPalm.

HashtagToken

kind	word
length	8
orth	lowercase
string	conflict

Context: pe, Asia & the USA = #conflictpalmoil.RT @BioenergyIntl: Malaysia: #Palm oil r

Figure 1: Screenshot of retokenised hashtags in GATE



Please type or paste some text (note that long texts may take a little while to analyse, and that the input will be treated as UTF-8):

```
I am not happy that I woke up at 5:15am this morning.. #greatstart

The way to be a great science communicator, be on TV, be male, be American. #sarcasm
```

Opinion Text

negative I am not happy that I woke up at 5:15am this morning.. #greatstart

negative The way to be a great science communicator, be on TV, be male, be American.

Figure 2: Example of sarcastic tweets in the demonstrator

> *The best feeling in the world is definitely being ignored. I love it #keepdoingit #bestthingever #sarcasm*

We then added the other hashtag to our list of sarcasm indicators, e.g. #notreally. These indicators were used to signify when sarcasm was present in a sentence: if one or more of these was found within the same sentence as an opinion, or within the same tweet, the existing polarity was reversed. Figure 2 depicts an example of two sarcastic sentences analysed by our improved application. Here we show the results via our online demo system[4].

As discussed in Section 2, in order to understand the scope of sarcasm correctly, we often need to investigate the hashtags. We have developed a set of rules which attempt to cover these issues of scope; some examples are given below:

- If there is a single hashtag denoting sarcasm, and the original sentiment is positive or neutral, we flip the polarity to negative.

[4] demos.gate.ac.uk/arcomem/opinions/

- If there is more than one hashtag, we look at any sentiment contained in those hashtags.
- If two hashtags both contain sarcasm indicators, we treat them as one single sarcasm indicator, e.g. *"#lying #notreally"*
- If a positive hashtag is followed by a sarcasm indicator, and the polarity of the tweet is positive or neutral, we flip the polarity of the sentiment of the positive hashtag to negative, and then apply this sentiment to the text (flipping the polarity of the tweet from positive or neutral to negative), e.g. *"Heading to the dentist. #great #notreally"*
- If a negative hashtag is followed by a sarcasm indicator, and the polarity of the tweet is positive or neutral, we treat both hashtags as negative and flip the polarity of the tweet to negative.

If there are no hashtags that indicate sarcasm, there are other clues we can look for. One indicator might be word combinations with opposite polarity, e.g. a typically negative word such as "rain" or "delay" together with a positive sentiment word such as "brilliant". This is particularly likely if the sentiment word is a strong one ("rules" for sarcasm dictate that one should normally use a highly emphatic positive word such as "excellent" rather than just a mildly positive one such as "good" if one wants to use it sarcastically). Modifying a positive sentiment word with a swear word also increases the likelihood of sarcasm. Another possibility is to include world knowledge, especially in combination with the above. For example, knowing that people typically do not enjoy going to the dentist, or that people typically like weekends better than weekdays, could help indicate that a positive sentiment about such things is likely to be sarcastic. However, these kind of issues are incredibly hard to solve, and we should assume that there are some instances of sarcasm that a machine is highly unlikely to ever identify.

4. Evaluating hashtag tokenisation

We conducted an experiment to measure the accuracy of hashtag tokenisation, using a gold standard set of tokenised hashtags (extracted from a larger corpus of general tweets) that we annotated manually. The gold standard set contained 2010 hashtags and 4538 tokens. The system achieved 98.12% Precision and 96.41% Recall, and an F1 of 97.25%.

For a fairly simple solution, these initial results are pleasing. One error is due to the presence of unknown named entities (people, locations and organisations) forming part of the hashtag. While some of these are recognised by our gazetteer lookup (especially locations), many of them are unknown. The named entity recognition component in GATE cannot identify these until they have been correctly tokenised, so we have a circular problem. However, in collaboration with the SemanticNews project[5], we have been working on a solution for disambiguating user names in Twitter by means of DBpedia; we are planning to experiment with adapting this technique to hashtags also, so that such entities can be recognised. We could also investigate using a language modelling approach based on unigram or bigram frequencies, such as used by Berardi et al. (Berardi et al., 2011).

5. Sentiment analysis of tweets

For the experiments in this paper, we have used the GATE-based sentiment analysis system developed as part of the Arcomem project (Maynard et al., 2012). This adopts a lexicon-based approach similar to that of Taboada et al. (Taboada et al., 2011), incorporating a series of intensifiers and negators which alter the polarity score.

To investigate the effect of sarcasm on tweets, we collected a corpus of 134 tweets containing the hashtag #sarcasm, from a larger set acquired via GardenHose on Oct 16 2012 (Preotiuc-Pietro et al., 2012), and manually annotated the sentences with a sentiment (positive, negative or no sentiment). Out of 266 sentences, 68 were found to be opinionated (approximately 25%), and of these 62 were negative while 6 were positive. Of these 68 opinionated sentences, 61 were deemed to be influenced by sarcasm while 8 were not (note that it is possible for a tweet to be sarcastic, or to at least have a sarcastic hashtag, without every sentence contained in that tweet being sarcastic). Unsurprisingly, we can see that the vast majority of sarcastic tweets have negative polarity. However, more than 10% of sarcastic tweets involved sentiment-containing sentences not directly affected by the sarcasm, i.e. whose polarity was unaltered.

We first evaluated the performance of our regular sentiment analyser (SA-Reg) and the analyser which considered sarcasm (SA-Sar); results are shown in Table 1. We measured detection of opinionated sentences, detection of opinions with the correct polarity of sentiment, and detection of correct polarity of sentiment only for those opinionated sentences correctly identified. Note that the results for opinion detection are identical for both analysers.

Sarcastic corpus	P	R	F1
Opinion	74.58	63.77	68.75
Opinion+polarity: SA-Reg	20.34	17.39	18.75
Opinion+polarity: SA-Sar	57.63	49.28	53.13
Polarity: SA-Reg	27.27	27.27	27.27
Polarity: SA-Sar	77.28	77.28	77.28

Table 1: Experiments on Sarcastic Corpus

We performed a second experiment on a set of general tweets collected and manually annotated with named entities as part of the TrendMiner project[6]. We selected from this a random sample of 400 tweets and performed manual double-annotation with opinion, polarity and sarcasm information. Inter-annotator agreement showed 87.5% observed agreement for polarity detection and 91.07% for sarcasm detection. The comparison of system annotation with the gold standard is shown in Table 2. The sarcasm detection performed very well at 91% Precision and Recall. Out

[5] http://semanticnews.org.uk

[6] http://www.trend-miner.eu

of 400 tweets, there were 91 sarcastic sentences, which is quite a high proportion, and many of these were not indicated by any kind of sarcasm marker.

Twitter corpus	P	R	F1
Opinionated	65.69	77.31	71.02
Opinion+polarity	52.61	61.92	56.89
Polarity only	80.08	80.03	80.05
Sarcasm detection	91.03	91.04	91.03

Table 2: Experiments on General Tweets

6. Related Work

There have been a few recent works attempting to detect sarcasm in tweets and other user-generated content. Tsur et al. (Tsur et al., 2010) use a semi-supervised approach to classify sentences in online product reviews into different sarcastic classes, and report an F-measure of 82.7% on the binary sarcasm detection task (although Precision is much higher than Recall). Liebrecht et al. (Liebrecht et al., 2013) use the Balanced Winnow Algorithm to classify Dutch tweets as sarcastic or not, with 75% accuracy, training over a set of tweets with the #sarcasm hashtag. Reyes et al. (Reyes et al., 2013) use a similar technique on English tweets to detect ironic tweets, using the #irony hashtag, with 70% accuracy. Davidov et al. (Davidov et al., 2010) use Tsur's algorithm for sarcasm detection, and achieve 82.7% F1 on tweets and 78.8% on Amazon reviews. Interestingly, they claim that the sarcasm hashtag is not used frequently in their corpus; however, perhaps this usage has become more common in the last 3 years. It appears that none of these approaches go beyond this step: even when a statement is known to be sarcastic, one cannot necessarily predict how this will affect the sentiment expressed.

With respect to the issue of hashtag tokenisation, there exist already plenty of methods for resolving this issue. Berardi et al. (Berardi et al., 2011) describe a Viterbi-based method which, unlike our method, takes word frequency into account. Given the word distribution model and a hashtag, they convert the hashtag to a vector of possible tokens, eventually returning just a single sequence of words equal to the original hashtag. In calculating the most likely sequence of words they only consider unigram frequency, however. Much work on segmentation has been done for other languages, especially Asian languages where white space is not typically used to delineate words, and e.g. for segmenting audio transcriptions. We have not proposed new research in this area here, simply the implementation of a fairly quick and dirty algorithm that works very well for hashtags.

7. Conclusions

In this paper, we have investigated some of the characteristics of sarcasm on Twitter, and described some preliminary experiments which study the effect of sarcasm on sentiment analysis. In particular, we are concerned not just with identifying whether tweets are sarcastic or not, but also considering the range of the sarcastic modifier on the meaning of the tweet and on the polarity of the sentiment expressed.

Our initial observations are that there are many interesting phenomena to be observed, and that detection of sarcasm in tweets, while useful, is not sufficient for accurate sentiment analysis of such tweets. Adding rules to deal with the scope of sarcastic hashtags does, however, improve performance considerably, though further improvements could still be made. We also do not deal currently with sarcasm when it is not mentioned in the hashtags.

In terms of our sentiment detection tools overall, there is also still much further work possible. Opinion mining from text, and particularly from social media which is difficult to analyse, is still very much in its infancy in terms of research, while very much a hot topic. This means that our tools are far from perfect, although they exhibit advances over the state-of-the-art in certain aspects, and there remain therefore a number of issues which have not yet been handled and which form part of our ongoing work. Such improvements include use of more detailed discourse analysis in order to provide better mechanisms for scope handling – not only of things like negation and sarcasm, but also of the opinions themselves.

8. Acknowledgements

This work was supported by the European Union under grant agreements No. 270239 Arcomem[7] and No. 610829 DecarboNet[8].

9. References

Giacomo Berardi, Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. 2011. ISTI@ TREC Microblog Track 2011: Exploring the Use of Hashtag Segmentation and Text Quality Ranking. In *TREC*.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

D. Bousfield. 2007. Never a truer word said in jest: A pragmastylistic analysis of impoliteness as banter in Henry IV, Part I. *Contemporary Stylistics*, pages 195–208.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an Architecture for Development of Robust HLT Applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7–12 July 2002*, ACL '02, pages 168–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

[7] http://www.arcomem.eu

[8] http://www.decarbonet.eu

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013*, page 29.

Diana Maynard, Kalina Bontcheva, and Dominic Rout. 2012. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012*, Turkey.

D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the workshop on Real-Time Analysis and Mining of Social Streams*.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, pages 1–30.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 1(September 2010):1–41.

O. Tsur, D. Davidov, and A. Rappoport. 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*, pages 162–169.