Knowledge-Based Systems 69 (2014) 124–133

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web

Raquel Justo [a], Thomas Corcoran [b], Stephanie M. Lukin [b], Marilyn Walker [b], M. Inés Torres [a,*]

[a] Speech Interactive Research Group, Universidad del País Vasco UPV/EHU, Depto. Electricidad y Electrónica, Fac. Ciencia y Tecnología, Sarriena s/n, 48940 Bilbao, Spain
[b] Natural Language and Dialogues Systems Lab, University of California, Santa Cruz, Department of Computer Science, 1156 N. High, SOE-3, Santa Cruz, CA 95064, USA

ARTICLE INFO

Article history:
Available online 17 June 2014

Keywords:
Emotional language
Social web
Feature extraction
Sarcasm
Nastiness

ABSTRACT

Automatic detection of emotions like sarcasm or nastiness in online written conversation is a difficult task. It requires a system that can manage some kind of knowledge to interpret that emotional language is being used. In this work, we try to provide this knowledge to the system by considering alternative sets of features obtained according to different criteria. We test a range of different feature sets using two different classifiers. Our results show that the sarcasm detection task benefits from the inclusion of linguistic and semantic information sources, while nasty language is more easily detected using only a set of surface patterns or indicators.

1. Introduction

Dialogic language on the web in interactive forms of media such as social networks and online forums is very different than the newspaper articles or task-oriented dialogs typically studied in work on natural language processing [33,23,10,36,37,43]. Online conversation is both more informal and more subjective: users tend to express their opinions with highly subjective and often emotional language. Moreover, in many cases context is needed in order to understand what people are saying.

Only recently have large corpora of this type of subjective dialog language become available, including labeled corpora of tweets, user reviews, online conversations and chats [40,12,39,32,22]. However because the number of studies using these corpora are limited, in many cases there are few baselines establishing how difficult it is to understand user utterances in these contexts. In particular, our work draws on the recently released the Internet Argument Corpus (IAC), a publicly available corpus of online forum conversations on a range of social and political topics [40]. The IAC includes a large set of conversations from *4forums.com*, a website for political debate and discourse. This site is a fairly typical internet forum where people post a discussion topic, other people post responses, and a treelike conversation structure is created. The corpus comes with annotations of different types of social language categories including sarcastic vs. not sarcastic, nasty vs. not nasty, rational vs. emotional and respectful vs. insulting. Figs. 1 and 2 provide examples of posts and post pairs from the IAC.

We focus on two types of social dialogic language for which annotations are provided in the IAC distribution, namely SARCASM and NASTINESS. Our primary goal is simply to test how difficult it is to automatically classify sarcastic and nasty utterances using supervised learning techniques, independently of topic. Both sarcasm and nastiness are highly subjective utterance types, but previous work suggests that they are likely to differ in detection difficulty [24]. We hypothesize that nastiness is presented more overtly, using less figurative language, and requiring less world knowledge to recognize. See Figs. 1 and 2.

We present a set of supervised learning experiments on detection of sarcasm and nastiness in online dialog. We compare a range of feature sets developed using different criteria. One of our foci is to test whether it is possible to automatically obtain a set of features valuable for identifying different forms of social language in online conversations regardless of the topic, style, speaker, or affordances of the online forum. To do so, we integrate statistic, linguistic, semantic and emotional information into our features, as a richer alternative to purely statistical or syntactically motivated surface patterns. These feature sets are then used to establish a baseline for a rule-based classifier and for a Naive Bayes classifier. The unique contributions of this paper include:

- Methods for discovering an appropriate set of features for different types of social language.

* Corresponding author. Address: Depto. Electricidad y Electrónica, Fac. Ciencia y Tecnología, Universidad del País Vasco, Sarriena s/n, 48940 Leioa, Spain. Tel.: +34 94601 2715.

E-mail addresses: raquel.justo@ehu.es (R. Justo), tcorcora@soe.ucsc.edu (T. Corcoran), slukin@soe.ucsc.edu (S.M. Lukin), mwalker@soe.ucsc.edu (M. Walker), manes.torres@ehu.es, manes@ehu.es (M.I. Torres).

http://dx.doi.org/10.1016/j.knosys.2014.05.021
0950-7051/

Category	Post or Post Pair
Sarcastic	**P1**: That's it? That was your post? To attack a source without anything backing you up? Bravo!
Sarcastic	**P2**: But...but...I just swallowed 56 zygotes so that they could vote by proxy through me for the Ripofflican(tm) party this upcoming election so they can win. Now I'm going to have all that indigestion for nothin'
Sarcastic	**P3**: OK, I get it! It's not being gay that is a sin; it's the gay sex that is a sin. You are such an expert on gay people! How about this EZ, what if two gay men live together, love each other, sleep in separate beds, never have sex but maybe give each other a little French kiss now and then? Is that OK, EZ? Here let me answer that question for you since your famous for avoiding the issue. Why yes Mr. Monster that is perfectly fine and a totally realistic thing to ask people to do! I know why you're avoiding answering the topic question directly. Believe me I am going to rip you a new one regardless (figuratively speaking.) You may as well let it all out. You know you want to. So go ahead, make my day.
Sarcastic	**P4-1**: "God" Does NOT play dice with Sexual Orientation. It is this way, because it was made this way before there was even the notion of a Bible, or Society. It is this way throughout the Animal Kingdom and remains so with us. **P4-2**: You need to experiment with the notion that you will be sexually attracted to your own sex. After youŕe done, come back and report your findings!
Sarcastic	**P5-1**: I simply mean a member of the species of "human." If something is a member of that species, it is a human being. **P5-2** If that is a human being, then is my kidney a human being too?
Sarcastic	**P6-1**: Let's support intrauterine civil registration. **P6-2**: Right, Cybererratus, uteruses are the property of church and state, not of the woman in whose body they are found.

Fig. 1. Sarcastic Posts and Post Pairs from `4forums.com`. Sarcastic examples were all reliably rated sarcastic: 4 or more turkers voted sarcastic, greater than 80%sarcastic yes count.

Category	Post Pair
Nasty	**P7-1**: Read what i actually wrote for a change Yank. "This is what socially inept losers do when they know their argument is ####, deny they ever made one. **P7-2**: I've asked repeatedly for an example of the evolution of a new body part in a population and all you've done is confirm the fact that you're an absolute #####."
Nasty	**P8-1**: I maintain that anyone who disagrees with this point of view is a depraved monster, a moral leper who is unworthy to be called a true human being. **P8-2**: And I maintain that your belief is dangerous to America and people like you should be deported to less civilized countries so you can truly appreciate what you had.
Nasty	**P9-1**: You can ignore the obvious question. This is what I find most evolutionists do, they ignore other branches of science that militate against their belief system, and legitimate questions raised by others are dismissed as being ignorant, or worse "religious". **P9-2**: Actually what they are really doing is ignoring silly irrelevent questions from clueless people with a religious agenda to promote.
Nasty	**P10-1**: i believe you just get angry when soemone doesn't agree with you 100%. **P10-2**: I'm not angry. In fact, I'm laughing at you right now.

Fig. 2. Nasty Post Pairs from `4forums.com`. Nastiness was annotated on a scale of $-5 \ldots 5$, with -5 being very nasty. The selected examples all had average nastiness ratings less than -2.5.

- Classification methods that explicitly consider the possibility of different forms of sarcasm such as hyperbole, understatement and irony [14,4].
- Feature sets which explicitly considers the semantic meaning and the problem of long utterances that include **both** the target category of sentences, e.g., sarcastic, **as well as** sentences not in the target category (e.g., not sarcastic).
- Comparison of ease of detection of two types of social language (sarcasm and nastiness) in an identical context.

2. Related work

Social networks, blogs, forums and many other websites allow people to share information. This social use of the web can provide valuable information to companies, which are therefore interested in opinion mining and sentiment analysis. Developing tools to select and analyze opinions is now a challenging goal for many companies. However, this information is informal and unstructured so that it is also a challenging topic for research in natural language processing. Truly understanding natural language requires computational models that can decoding the semantic meaning of utterances as well as the sentics, requiring methods that go beyond words to deal with concepts [8]. Cambria et al. [7] present a review of the evolution of research on these topics. According to their work, sentiment analysis has typically been performed over on-topic documents [7]. The review includes a discussion of the most relevant text features for sentiment classification, such as term and n-gram frequencies and presence, certain adjectives as indicators, as well as phrases chosen by POS patterns and sentiment lexicons. Revised methods included keyword spotting, lexical affinity, and concept-based approaches using web ontologies, as well as Bayesian inference and support vector machines as statistical classifiers. Recent techniques also include concept-level analysis for both knowledge-based [9,3] and statistical approaches [13]. However, concept-based approaches to date have mainly dealt with polarity detection [9,13]. Only irony, sometimes theorized as an utterance that assumes the opposite of the actual situation, has been addressed as playing the role of a polarity reverser [3,32]. In this situation, detection of irony is assumed to require a representation of the dialog context.

Previous theoretical work on sarcasm defines it as mocking, contemptuous, or ironic language intended to convey scorn or insult [14,4]. However while it may employ ambivalence, it is not necessarily ironic. For example, if a person requires a lot of time to obtain the solution of a simple mathematical problem, one might ask "How many days does he need?". This is sarcastic but not ironic. However, "He is like Einstein" is a sarcastic sentence that uses irony. More importantly, almost any utterance can be sarcastic in the correct context, e.g., "you are early" will be interpreted as sarcastic in a context where the addressee is clearly late. Some sarcastic forms also require sociocultural knowledge to detect, e.g., "The Nazis really brought good to the world".

In spoken language, sarcasm is manifested partly in vocal inflections [19]; however, in written language there are no standard cues for signaling the speaker's intention. Several form of punctuation have been proposed, like percontation point ⸮, but are not usually employed. Twitter users sometimes employ hashtags like '#sarcasm' as an explicit marker of sarcasm, and [22] show that extra-linguistic elements such as hashtags can be seen as the social media equivalent of non-verbal expressions that people employ in live interaction. Unfortunately, all sarcastic tweets are not marked by a hashtag. Previous work suggests that hashtags: (1) are not used reliably; and (2) may be used only for the hardest forms of sarcasm that would not be recognized without the hashtag [15,32]. In written texts, scare quotes may be used to denote skepticism or disagreement with the quoted terminology, but this cue is not reliable either. Clearly, new methods are needed to detect the different forms in which sarcasm is expressed in social media.

Previous work on sarcasm detection has primarily focused on product reviews and tweets rather than online dialogs. Ref. [38] proposes a semi-supervised algorithm based on *k*-nearest neighbors to detect sarcasm in online product reviews using pattern and punctuation based features. Subsequently the same algorithm applied to millions of tweets collected from Twitter [15] produced significantly better results. They suggest that this is due to the contextless nature of Twitter, which forces tweeters to express sarcasm in a way that is easy to understand from individual sentences. In contrast, sarcastic utterances in Amazon reviews co-appear with other sentences (in the same review) and sarcastic meaning may emerge from the context. Other work on sarcasm detection in tweets explored the contribution of linguistic and pragmatic features of tweets to the automatic separation of positive and negative polarity messages from sarcastic messages [15]. This work found that the three pragmatic features (*ToUser*, *smiley* and *frown*) were among the ten most discriminating features in the classification task. Ref. [22] uses *n*-gram features to build a Winnow classifier to identify sarcastic tweets in Dutch, based on the hypothesis that hashtags can, and do, replace linguistic markers typically needed to mark sarcasm. They also conclude that people tend to be more sarcastic towards specific topics such as school, homework, and family life. Ref. [32] presents a method to detect a common form of sarcasm consisting of contrasting a positive sentiment with a negative situation in tweets. They use a bootstrapping method to recognize the contrasting sentiments in syntactic structures related to sarcasm.

Lukin and Walker (2013) is the only other work we are aware of that focuses on sarcasm detection in online dialog, using a bootstrapping method to automatically find sarcastic utterances in unannotated dialog [24]. Their work explores using a weakly supervised bootstrapping approach, rather than establishing accuracies for a range of supervised classifiers based on different feature sets as we do here. As above, this task is more difficult than sarcasm detection in Twitter due to the contextless nature of the tweets and their limited length. In contrast, in product reviews and online dialog, a sarcastic utterance may also include sentences that are not sarcastic at all, introducing noise to the sarcasm classification task. Moreover, other challenges of online dialog include the diversity of topics and the dialogic nature of the utterances. These factors lead to more colloquial vocabulary and language style, and the appearance of many different forms of sarcasm [14,4].

Some examples of these situations are provided in Fig. 1. For example, while **P1** uses the word "Bravo" that could be identified as a sarcastic cue, **P2** does not seem to contain any surface sarcastic indicators: recognizing it as sarcastic depends on semantic interpretation and sociocultural knowledge. We also see long sarcastic posts (**P3**) where some sentences are not sarcastic. Finally, example **P5–2** shows a form of sarcasm that might be difficult to identify because there appear to be no explicit cues: rather recognizing **P5–2** as sarcasm appears to rely on world knowledge to recognize that it makes a ridiculous analogy, equating a kidney to a human being.

Turning to our other category of social language, we also aim to automatically identify utterances from the IAC that are categorized as nasty. Fig. 2 provides examples of nasty posts and post pairs from IAC. Post **P7–2** contains potentially identifiable nasty patterns or keywords such as "you're an" along with the presence of a curse word [35]. **P9–2** also contains keywords like "silly" and "clueless" but context may be necessary to recognize that these words are used as an insult with respect to the third party evolutionists mentioned in **P9–1**. More interestingly, identifying **P10–2** as a nasty utterance requires understanding that when the poster of **P10–2** is "laughing", they are actually amused at the expense of the poster of **P10–1**.

Early work in this area in [35] focuses on identifying abusive messages, known as flames, in emails. This approach builds features from the syntax and semantics of each sentence and uses a decision tree for rule-based classification. These features were hypothesized to be useful and then automatically extracted. Ref. [31] examine flame detection in text messages based on statistical models and automatically create rule-based patterns based on word sequences and inserting wild-cards. They found that previous work was unable to determine agreement between human expert annotators on what a flame is. They define a tolerance margin for flames based on certain conditions or different contexts, thus creating a dictionary of words and phrases with different degrees of flame intensity. Ref. [44] use a semi-supervised technique for detecting insults in Tweets based on statistical topical modeling techniques for feature induction rather than keyword or pattern matching as in the two techniques previously described. By bootstrapping this process, they are able to classify many Tweets using only a small set of seed words. Ref. [34] use crowdsourcing to label comments from an online community form that contain profanities and insults. They then use Support Vector Machines to identify negative comments, which are often coupled with insults, and furthermore to whom the comment is directed: to another user in the forum or to a third party. Their best performing system for insult detection is a multistep classifier that utilizes both valence and relevance analyses.

Other work that is closely related is the detection of cyberbullying or trolling (one of the multiple forms of cyberbullying). Cyberbullying is generally defined as the repeated injurious use of harassing language to insult or attack another individual in an online community and it is widely recognized as a serious social problem, especially for adolescents. In [5] a troll detector is trained by first identifying the concepts most commonly used by trolls and then expanding the resulting knowledge base with semantically related concepts. This work uses *Sentic Computing* (a new opinion mining and sentiment analysis paradigm) to analyse the texts. Ref. [11] provides a new concept-level approach for bullying detection that outperforms state-of-the-art natural language processing and statistical approaches. Ref. [18] determined the terms most

commonly used in cyberbullying and developed queries that can be used to detect cyberbullying content using a combination of bag-of-words language models and a supervised machine learning approach based on *Latent Semantic Indexing*.

More generally, all types of semantic features of user generated content in social media can be useful. This includes trustworthiness of information and deception detection, as well as language that indicates social or political affiliation between conversants [1,25,41,16,2]. For example previous work focuses on using supervised techniques to identify web, email and opinion spam [28,27,20,21,17]. However, because deceptive opinion spam is not easily identified by humans, it is difficult to construct a gold-standard corpus. Several lines of previous work therefore crowdsource deceptive reviews using Amazon Mechanical Turk [28,20,21]. Other work [26] proposed a method to exploit observed reviewing behaviors to detect opinion spammers (fake reviewers) in an unsupervised Bayesian inference framework. Other approaches are based on graph models [42], or analysis of the generality or specificity of the reviews, as well as some aspects of the content [20,21].

3. Empirical approach

In order to identify sarcasm or nastiness in online written conversation, humans make use of different types of knowledge that is related to the type of language that is used. Thus, in this work, we would like to provide the system with this kind of knowledge by using different information sources. Specifically, we were interested in exploring the ability of statistic, linguistic, semantic, length and emotional information to heuristically approximate the ability of humans to detect sarcasm and nastiness.

3.1. Feature extraction

In this section we present the way in which the required information is given to the system. Specifically, different sets of features were obtained according to the following criteria:

- *Mechanical Turk Cues:* We first consider a set of features made up of cues that were identified as indicators of sarcasm or nastiness in our previous work [24], e.g., "oh really", "I get it", and "no way" associated with sarcasm, and others associated with nastiness like "idiot", "you're an", "your ignorance is". These were identified in a Mechanical Turk procedure applied to an independent set of 617 posts for sarcasm and a different set of 510 posts for nastiness. The sarcastic set of cues consists of 1815 *n*-grams (1629 unigrams, 147 bigrams, 39 trigrams) and the nasty set of cues consists of 3256 *n*-grams (289 unigrams, 966 bigrams and 2001 trigrams). Lukin and Walker (2013) demonstrated that these cues for sarcasm and nastiness were reliable selected across a number of different annotators [24]. Table 1 provides detailed examples of some of the top unigram, bigram and trigram cues for both sarcasm and nastiness, along with the interannotor agreement scores for each cue and the frequency of the cue in the annotated dataset (see column **MT** for the cues and column **IA** for interannotator agreement scores). Note the very high agreement among the human annotators, even though the frequency of each cue is relatively low. Thus, although these cues might also appear in other types of sentences, they appear to be recurrent resources in written sarcastic/nasty utterances that appear in forum dialogs, regardless of the specific context and situation. Table 1 also indicates some of the features that would be selected from the same dataset if a feature selection procedure based on χ^2 was used. Examination of the selected features shows that χ^2 feature selection appears to overfit on this small dataset, and indeed Lukin and Walker show that the cues selected by χ^2 do not perform as well in their bootstrapping approach as the **MT** cues selected by the human annotators. See [24] for more details about how the annotations were collected.
- *Statistical Cues:* Since manually identifying the cues associated with each emotion is costly, the set of *Mechanical Turk Cues* might not be general enough to correctly identify the required emotions in different corpora. Thus, a second set of features was automatically extracted from the training set. This set of

Table 1
Mechanical Turk (MT) and χ^2 indicators for Sarcasm and Nastiness, shown with Interannotator Agreement (**IA**) and Frequency of Occurrence (**N**). From Lukin and Walker (2013), and used in the experiments below.

SARCASM CUES				NASTINESS CUES			
χ^2	MT	IA	*N*	χ^2	MT	IA	*N*
unigram				*unigram*			
right	ah	0.95	2	like	idiot	0.9	3
oh	relevant	0.85	2	them	unfounded	0.85	2
we	amazing	0.8	2	too	babbling	0.8	2
same	haha	0.75	2	oh	lie	0.72	11
all	yea	0.73	3	mean	selfish	0.7	2
them	thanks	0.68	6	just	nonsense	0.69	9
mean	oh	0.56	56	make	hurt	0.67	3
bigram				*bigram*			
the same	oh really	0.83	2	of the	don't expect	0.95	2
mean like	oh yeah	0.79	2	you mean	get your	0.9	2
trying to	so sure	0.75	2	yes,	you're an	0.85	2
that you	no way	0.72	3	oh,	what's your	0.77	4
oh yeah	get real	0.7	2	you are	prove it	0.77	3
I think	oh no	0.66	4	like a	get real	0.75	2
we should	you claim	0.65	2	I think	what else	0.7	2
trigram				*trigram*			
you mean to	I get it	0.97	3	to tell me	get your sick	0.75	2
mean to tell	I'm so sure	0.65	2	would deny a	your ignorance is	0.7	2
have to worry	then of course	0.65	2	like that?	make up your	0.7	2
sounds like a	are you saying	0.6	2	mean to tell	do you really	0.7	2
to deal with	well if you	0.55	2	sounds like a	do you actually	0.65	2
I know I	go for it	0.52	2	you mean to	doesn't make it	0.63	3
you mean to	oh, sorry	0.5	2	to deal with	what's your point	0.60	2

cues consists of the unigrams, bigrams and trigrams extracted from each utterance in the training set. Then a feature selection procedure is used to reduce the number of features removing those that are not relevant.

- *Linguistic information:* Even when considering *Statistical Cues*, we need a large training set to automatically extract a set of cues general enough to be used with different test data. However, large annotated sets of training data are not always available. Therefore we also consider what types of generalizations of features are possible through the use of linguistic categories. Specifically, in this feature set, we generalize using the part-of-speech (POS) labels associated with each word in an utterance. These labels correspond to the classical 8 lexical categories in English: Noun, Pronoun, Adjective, Verb, Adverb, Preposition, Conjunction and Interjection. The inclusion of this kind of linguistic information provides a more general way of defining patterns. For example, "Really? well" is a sarcastic feature appearing in the training set along with the sarcastic POS pattern "ADV ADV". Then, if "Really? then", which could also be a sarcastic marker, does not appear in the training set, its corresponding POS pattern "ADV ADV" could still be used to identify this phrase as a useful feature. Thus, in this work, a set POS *n*-grams was considered where $n = 1, 2, 3$.
- *Semantic information:* Although some patterns may be good cues to different emotions, there is additional information in online conversation that is more related to semantic categories than to specific words or POS types. Here, we explore whether LIWC classes can provide generalizations of patterns based on semantic information [30]. The LIWC dictionary includes 64 different lexical categories including sets of words related to *anger*, *affect*, *feelings*, *social*, *number*, and *health* but also more general categories such as *pronoun*, *verb* or *future*. In order to make more meaningful feature patterns, we associate each word in an utterance with its corresponding LIWC semantic category or categories. For example, the word "address" is associated with the LIWC category "home" while "adult" is associated with LIWC categories "social" and "humans". Then, a vector is built with the number of times each category appears in the utterance.
- *Length information:* Given the specific corpus we are dealing with, in which there are very different length posts, we noticed that length information can be a very useful factor for the correctly interpretation of previous sets of features. Thus, in this work, a vector with the following information is also considered for each utterance: number of words, number of characters, number of sentences, average words per sentence, average character per sentence.
- *Concept and Polarity Information:* Considering that both, sarcasm and nastiness, are closely tied to specific emotional states, new feature sets including emotional information were also considered. Specifically, *SenticNet-3.0* [6] was used. This tool aims to infer cognitive and affective information from natural language text. More specifically, it provides the semantics and sentics associated with more than 14,000 common sense concepts. In this work, we wrote the sentences of the posts in terms of those concepts and obtained the semantics and polarity associated with each one. If we consider the sentence *"I love chocolate but ice cream is better"* the corresponding concept sentence would be *"love better"*. The semantic information associated with this concept sentence would be: **love**: *lust, love_another_person, sexuality, beloved, show_empathy* and **better**: *tall_play_basketball, reliable, good_quality, all_right, best*. Thus, the feature set is made up of vectors with the number of times each semantic item appears in the utterance. On the other hand, polarity information was also obtained for the concept sentence: **love**: 0.667, **better**: 0.132. A feature set considering this kind of information was also built. Concretely, a vector of two features, *Positive Polarity* (the sum of positive polarities in the utterance) and *Negative Polarity* (the sum of the negative polarities) was built for each utterance. In our example that vector would be (0.799, 0). Note that instead of using only one score with the total polarity of the sentence, both positive and negative polarities are considered. This allows us to differentiate between sentences with neutral words and those that have a similar number of words with positive and negative polarity.

3.2. Sarcasm detection

In order to tackle the problem of classifying a post as sarcastic or not sarcastic, different combinations of the features above were defined.

First, the set of *Mechanical Turk cues* (**MT_C**), as illustrated in Table 1, and the set of *Statistical cues* (**ST_C**) were considered. The idea is to see whether a set of cues of a specific emotion can be automatically extracted from the corpus.

In a second stage, *Semantic Information* was considered. The use of this kind of information can help to detect some sarcastic forms that are related to the correct interpretation of the post's meaning. For example, in the following sarcastic utterance: *"is there any reason to respond to this guy at all? if we ignore him, may be he'll go again"*, there are no specific terms that indicate sarcasm. Additionally, there are some uses of irony, like "Oh how I love *being ignored*", where a positive sentiment is contrasted with a negative situation [32]. In this kind of sarcasm, we hypothesize that a balanced number of categories associated with positive and negative emotions will be a cue. Thus, we first considered a set of features with the information coming from LIWC semantic classes (**Liwc**). Then, we also used a set of semantic features extracted from *SenticNet 3.0* tool, **SEM**, so that emotional semantic information can be compared to more general semantic information obtained from LIWC classes. The feature sets were built as explained in Section 3.1. Then, and in order to see whether these two sources are complementary, a new set, **Liwc_SEM**, with the combination of the two semantic information sources was built.

Then, in a third stage, the automatically obtained **ST_C** set was combined with other types of information. Our goal is to enrich the statistical set in order to detect additional forms of sarcasm that are directly related to linguistic and semantic information and that are not only expressed by using specific patterns like "oh really":

First of all, *Statistical cues* and *Linguistic information* were gathered in a new set **ST_Li**. The inclusion of *Linguistic information* is useful to detect more general linguistic patterns found in the training data as described above.

Alternatively, **ST_Liwc** and **ST_Liwc_SEM** sets were built including *Statistical cues* + *Semantic information* extracted from LIWC classes and enriched with emotional semantic information.

On the other hand, the set of utterances we are attempting to classify has a great deal of variation, including long sarcastic posts where sarcasm is only used in part of the post, as well as short sarcastic posts, where the whole post is sarcastic. The posts that are not sarcastic also vary a great deal in length. Thus, it is possible for the features for an utterance to include a balanced number of positive and negative emotions if the post is long, but not sarcastic. In shorter posts, such as those in Twitter [32], the probability of being sarcastic in that situation is higher. In the same way, there could be a specific cue in a short post than can be a very good indicator of sarcasm, whereas it could not indicate sarcasm at all in a longer post where it is not relevant for the whole utterance. Thus, *Length information* was also considered in **ST_Lg** set that combines *Statistical cues* + *Length information*.

Then, we defined a feature set that combines *Statistical cues* with *Mechanical Turk cues* which we label **ST_MT**, in order to test whether the human provided annotations from Mechanical Turk

can significantly improve the results. That is, the idea is to see to what extent there are specific patterns in **MT_C** not included in **ST_C** that might be discriminant in the classification process.

In a fourth stage of our work, we included *Polarity Information*, extracted from *SenticNet 3.0* tool, together with semantic and statistical information leading to **Liwc_SEM_Pol** and **ST_Liwc_Pol** sets. The inclusion of polarity in the feature set could help detect the emotional attitude of the poster. It could also capture the uses of irony where positive situations are contrasted with negative ones as explained above.

Finally, two new different sets were also considered: **Comb** that included *Statistical cues*, *Linguistic information*, *Semantic information* and *Length information* and **All** set including **Comb** + **MT_C**. These new sets could be helpful to see if different information sources can cooperate to provide better results.

3.3. Nastiness detection

Since some sarcastic forms can express nastiness in an utterance, there should be similarities between the previous task and this one; so the same sets of features suggested for sarcasm detection were also considered here. However, there are some expressions or cue words that indicate that an utterance is insulting or nasty without any ambiguity, i.e. as mentioned above we hypothesized that nastiness is often expressed very overtly. See Fig. 2. For example, when a post contains the "you idiot" bigram it is unambiguously nasty, while it is possible for the "really? well" bigram to occur in a not sarcastic sentence, even though it appears more frequently in sarcastic posts. Thus, we expect that the features derived from patterns extracted from the training data (automatically or manually) should be more robust for nastiness.

Summing up, the different feature sets employed here were also **MT_C**, **ST_C**, **Liwc**, **SEM**, **Liwc_SEM** on the one hand and **ST_Li**, **ST_Liwc**, **ST_Liwc_SEM**, **ST_Lg**, **ST_MT** on the other hand. Finally, **Liwc_SEM_Pol**, **ST_Liwc_Pol**, **Comb** and **All** sets were also considered.

4. Classification methods

We evaluate the utility of the cues and features defined above for the task of detecting either sarcastic or nasty posts in online conversation through classification experiments. We utilize two different classifiers: a rule-based classifier as a baseline and a classical Naive Bayes classifier.

4.1. Baseline classifier

The rule-based classifier is designed to allow our results to be comparable to previous work aimed at developing a high-precision classifier for bootstrapping [24,33]. However in contrast to the bootstrapping approach, we use a much larger training set to learn cues and we apply different methods to learn more general cues as described in Section 3.1.

The classifier is based on computing two metrics for each feature: the number of times each feature appears in the training set (**freq**) and the percentage of cases that the feature occurs in a sarcastic utterance (%**sarc**) for sarcasm detection, or the percentage of cases that the feature occurs in a nasty utterance (%**nast**) for nastiness detection. Then two different thresholds were defined for these metrics as follows: $\theta_1 \leqslant$ **freq** and $\theta_2 \leqslant$ %**sarc** or $\theta_2 \leqslant$ %**nasty**. A test post is then classified as sarcastic if it contains at least two features x_1 and x_2 such that ($\mathbf{freq}(x_1) > \theta_1 \wedge \mathbf{freq}(x_2) > \theta_1$) and ($\%\mathbf{sarc}(x_1) > \theta_2 \wedge \%\mathbf{sarc}(x_2) > \theta_2$). In the same way a test post is classified as nasty if it contains at least two features x_1 and x_2 such that ($\mathbf{freq}(x_1) > \theta_1 \wedge \mathbf{freq}(x_2) > \theta_1$) and ($\%\mathbf{nast}(x_1) > \theta_2 \wedge \%\mathbf{nast}(x_2) > \theta_2$). In this work, θ_1 and θ_2 parameters were chosen in a tuning step where the classifier was run over the training set.

4.2. Naive Bayes classifier

A Naive Bayes classifier implements Bayes' theorem under the assumption of independence between pairs of features. Given a class ω and a feature vector $\mathbf{x} = x_1, \ldots, x_d$ the Bayes' formula states:

$$P(\omega|, x_1, \ldots, x_d) = \frac{p(x_1, \ldots, x_d|\omega)P(\omega)}{\sum_{j=1}^{c} p(x_1, \ldots, x_d|\omega_j)P(\omega_j)} \quad (1)$$

Using the naive independence assumption $p(x_1, \ldots, x_d|\omega) = \prod_{i=1}^{d} p(x_i|\omega)$ the Naive Bayes classification rule is set as:

$$\hat{\omega} = \arg\max_{\omega} P(\omega|x_1, \ldots, x_d) \approx \arg\max_{\omega} (P(\omega)\prod_{i=1}^{d} p(x_i|\omega)) \quad (2)$$

Usually Naive Bayes classifiers assume classic forms for the distribution $p(x_i|\omega)$ such as Gaussian, multinomial or Bernouilli. In this work, we assume our data to be multinomially distributed, Naive Bayes Multinomial (NBM). The distribution is parametrized by vectors $\theta_\omega = (\theta_{\omega_1}, \ldots, \theta_{\omega_d})$ for each class ω, where d is the number of features and $\theta_{\omega_i} = p(x_i|\omega)$ is the probability of feature i appearing in a sample in class ω. Parameters θ_ω are estimated as:

$$\hat{\theta}_{\omega_i} = \frac{N_{\omega_i} + \alpha}{N_\omega + \alpha d} \quad (3)$$

where $N_{\omega_i} = \sum_{\mathbf{x} \in \mathcal{H}} x_i$ is the number of times feature i appears in a sample of class ω in the training set $\mathcal{H}$, $N_\omega = \sum_{i=1}^{|\mathcal{H}|} N_{\omega_i}$ is the total count of all features for class ω, and $\alpha \geqslant 0$ implements a smoothed maximum likelihood estimation that prevents zero probabilities.

When the number of features is high, a previous selection procedure is required in order to extract the most significant features. Here, the feature selection procedure is based on the χ^2 distribution. The χ^2 test determines whether there is a significant association between two variables. In this way χ^2 statistics is also used to select a set of significant features for classification. The set of the n features with the highest values for the χ^2 statistic is selected from the data, which are assumed to be multinomially distributed among classes. This selection was carried out in a tuning step where the classifier was run over the training data set.

5. Experimental evaluation

The Cues and Features defined in Sections 3.2 and 3.3 for sarcasm and nastiness detection were evaluated in two main sets of classification experiments. The first one aims to evaluate the basic sets of cues, **MT_C** and **ST_C**, using a simple baseline classifier. The second one, presented in Section 5.3, evaluates all the features with a NBM classifier. Classifiers results are reported in terms of Precision (*P*), Recall (*R*), *F*-measure (*F*) and Accuracy (*Ac*).

5.1. Corpus and annotations

For our experiments we used the IAC [40], along with the annotations it provides for different types of social language categories including sarcastic vs. not sarcastic, nasty vs. not nasty, rational vs. emotional and respectful vs. insulting. We focus specially on sarcastic and nasty utterances labeled by Mechanical Turkers as described in [24]. From this corpus we selected the sarcasm and nastiness data sets used in our experiments:

- Sarcasm Data Set: a subset of IAC consisting of 6,461 instances balanced between sarcastic and not sarcastic posts.
- Nastiness Data Set: a subset of the IAC consisting of 2765 instances balanced between nasty and not nasty posts.

A cross-validation procedure was employed in all experiments. Both the Sarcasm and the Nastiness Data Sets were previously split into 10 disjoint subsets. At each iteration of the cross validation procedure 9 subsets were used as training and the last subset is used as an independent test-set.

5.2. Experiments with the baseline classifier

The values of θ_1 and θ_2 that maximized the *F*-measure over the training set (consisting of 9 subsets of the cross-validation partition) were chosen as thresholds for the final classification of a completely independent test set. This procedure was carried out at each iteration of the cross-validation procedure for both sarcasm and nastiness detection.

In this first series of experiments, **MT_C** and **ST_C** cues were employed and compared. The results are given in Table 2 and Table 3 respectively. These results show that the **ST_C** feature set provides better results than the **MT_C** features in terms of all the evaluation metrics, for both sarcasm and nastiness detection. However, the most important difference can be observed in Recall values. This could be explained by looking at the number of sarcastic features that are above the mentioned threshold (θ_1, θ_2) in the training set for each case. When considering **MT_C** this value was 345 for sarcasm and 696 for nastiness, while it was much higher (10,824 and 3432) for **ST_C** case. As discussed above, and illustrated by Table 1, the **MT_C** features are better when the training set is small, but for larger data sets feature selection based on χ^2 reveals more useful cues.

If we directly compare the performance of sarcasm and nastiness detection, we note that the nastiness detector works better with both **MT_C** and **ST_C** feature sets. This supports our initial hypothesis that nasty language is easier to identify than sarcastic language. Further support for this hypothesis is that the training set for sarcastic detection was much larger than the training set for nastiness (6461 vs. 2765), but better results were still obtained for nastiness.

This difference is more evident for **ST_C** set and it could be related to the specific sarcastic and nasty **MT_C** sets that were employed, as illustrated in Table 1. While there are many more sarcastic unigrams than sarcastic bigrams or trigrams, the set of nasty trigrams is much higher than the sets of unigrams and bigrams. Since it is more difficult to statistically select predefined bigrams and trigrams than unigrams in a corpus, the sarcastic **MT_C** set is more efficient than the nasty **MT_C** set.

Table 2
Precision, recall, *F*-measure and accuracy results for the baseline classifier for sarcasm detection and **MT_C** and **ST_C** sets.

	MT_C	ST_C
Sarcasm		
θ_1, θ_2	2, .55	2, .75
# sarc. feat.	345	10,824
P (%)	53.5	54.4
R (%)	46.7	55.3
F (%)	49.8	54.8
Ac (%)	53.0	54.5

Table 3
Precision, recall, *F*-measure and accuracy results for the baseline classifier for nastiness detection and **MT_C** and **ST_C** sets.

	MT_C	ST_C
Nastiness		
θ_1, θ_2	2, .55	2, .80
# nasty feat.	696	3,432
P (%)	56.6	61.6
R (%)	47.1	57.4
F (%)	51.3	59.3
Ac (%)	55.4	60.7

5.3. Experiments with the Naive Bayes classifier

The second series of experiments aims to evaluate the ability of the features defined in Section 3 to detect sarcasm and nastiness through a NBM classifier. We first selected the most relevant features and for this purpose, different sets with increasing number of features were obtained according to the χ^2 statistic. The number of features added to each step was the total number divided by 25. Then the training set was classified, and the *F* measure computed, using each of the 25 sets. The feature set that maximized *F* was selected, and the test set, which was not seen during feature selection and tuning, is evaluated using the selected features. This procedure was repeated in each iteration of the cross-validation procedure for both sarcasm and nastiness detection. For these experiments we employed the implementation provided in the Scikit-learn python package for χ^2 statistics and NBM classifier [29].

5.3.1. Sarcasm detection

Table 4 shows the obtained *P*, *R*, *F* and *Ac* results for the different feature sets described in Section 3.2. The average number of features selected for each of the 10 cross-validation iterations is also given (1st column).

Table 4 shows that the NBM classifier provides better results than the Baseline classifier. The improvements in terms of *F* are higher than 20% for both the **MT_C** and **ST_C** feature sets. However, focusing on the comparison between **MT_C** and **ST_C** the same tendency is observed for both classifiers. That is, **ST_C** provides better results in terms of all the measures, but the most significant improvements are associated with *R*.

Note that if we directly compare the different types of *Semantic Information* used here (**SEM** vs. **Liwc**), that we get slightly better results with more general semantic classes (**Liwc**). However, the combination of the two types of features (**Liwc_SEM**) provided a significant improvement in terms of F. Thus, it seems that the emotional semantic information complements the general semantic information provided by **Liwc**, resulting in better performance. However, note that the **ST_C** features still perform better.

In the third stage of our work, different information sources were added to **ST_C**, namely linguistic and semantic information. Table 4 shows that the four new sets (**ST_Li**, **ST_Liwc**, **ST_Liwc_**

Table 4
F-measure (*F*), accuracy (*Ac*), precision (*P*) and recall (*R*) results obtained with NBM classifier for sarcasm detection and different sets of features.

	# Feat.	*P* (%)	*R* (%)	*F* (%)	*Ac* (%)
Sarcasm					
MT_C	1089	65.1	67.3	66.2	65.6
ST_C	239 K	**67.8**	72.3	69.7	**68.7**
Liwc	35	60.2	64.5	62.2	60.7
SEM	5093	59.8	63.8	61.6	60.1
Liwc_SEM	5492	59.3	73.6	65.6	61.5
ST_Li	374 K	65.6	75.9	70.0	67.5
ST_Liwc	334 K	65.3	**77.4**	**70.7**	67.9
ST_Liwc_SEM	370 K	66.1	75.1	70.1	68.0
ST_Lg	157 K	65.2	**77.4**	**70.5**	67.8
ST_MT	353 K	66.1	76.2	**70.6**	68.3
Liwc_SEM_Pol	5454	59.3	73.7	65.7	61.5
ST_Liwc_Pol	348 K	65.5	76.9	70.7	68.1
Comb	303 K	65.3	76.7	70.1	67.3
All	280 K	65.3	76.7	70.2	67.5

SEM, **ST_Lg**) provide significant improvements in terms of *R* when compared to **ST_C** set, i.e. although *P* values are lower *F* increases in all cases. Therefore it is clear that the addition of linguistic, semantic and length information improves our ability to identify different forms of sarcasm. The best results are obtained with **ST_Liwc** and **ST_Lg**, the feature sets including *Semantic* and *Length* Information respectively. Although **Liwc_SEM** provided better results than the **Liwc** feature set, the combined feature set **ST_Liwc_SEM** does not result in an *F* value as high as that obtained with the **ST_Liwc** feature set. This might be due to the fact that feature selection that has more influence on the results when the number of features is high. More precise ways of carrying out feature selection should be considered in future work, in order to obtain more balanced sets of statistical and other types of features. However, the results obtained with these semantic feature sets highlights the influence of semantic meaning on sarcasm detection in online dialog, when combined with the effect of the differences in post lengths. Finally, when considering the **ST_MT** feature set, which combines both statistical and manually obtained cues, the *R* is lower than that obtained with **ST_Liwc** but the *P* does not drop as much: this set has a *F* value similar to those obtained with **ST_Lg** and **ST_Liwc**. These results suggest that adding human knowledge to automatically obtained cues leads to higher performance.

The integration of polarity information in the feature set does not significantly improve the results, as can be seen when comparing **Liwc_SEM** to **Liwc_SEM_Pol** and **ST_Liwc** to **ST_Liwc_Pol**. However, given that polarity information typically helps with other sentiment classification tasks, we posit that its effect might be hidden by feature sets with more features, and plan to explore different ways of integrating polarity features in future work.

Table 4 also shows that combining different information sources together in **Comb.** yields performance improvements over the **ST_C** feature set in terms of *R* and *F*, but performance lower than **ST_Liwc**. Therefore it is clear that additional features do not always improve performance. The **MT_C** feature set also achieves very similar results (see the column of the **All** set).

Over all the different feature sets, the best *P* and *Ac* values were obtained with **ST_C**, while **ST_Liwc** provides the best *R* and *F* values.

5.3.2. Nastiness detection

The same experiments were carried out for the nastiness detection task with results shown in Table 5. Again, a comparison of **MT_C** and **ST_C** shows that better results are achieved with **ST_C** for all measures, but for nastiness the improvement is even greater.

Table 5
F-measure (*F*), accuracy (*Ac*), precision (*P*) and recall (*R*) results obtained with NBM classifier for nastiness detection and different sets of features.

	# Feat.	*P* (%)	*R* (%)	*F* (%)	*Ac* (%)
Nastiness					
MT_C	922	73.2	63.8	67.9	70.0
ST_C	246 K	**77.4**	81.2	**79.1**	78.5
Liwc	41	66.3	66.9	66.4	66.2
SEM	4913	67.9	67.1	67.2	67.4
Liwc_SEM	5056	64.8	77.9	70.6	67.7
ST_Li	229 K	76.9	76.8	76.7	76.7
ST_Liwc	269 K	74.3	**84.7**	**79.0**	77.4
ST_Liwc_SEM	262 K	75.9	80.2	77.9	77.2
ST_Lg	149 K	75.4	79.7	77.2	76.4
ST_MT	235 K	76.9	82.3	**79.4**	**78.6**
Liwc_SEM_Pol	5056	64.8	77.7	70.5	67.6
ST_Liwc_Pol	269 K	75.7	83.1	78.8	77.6
Comb.	129 K	74.1	77.8	75.6	75.1
All	194 K	74.3	78.3	76.0	75.3

Note that in this case a fewer number of cues were obtained after the feature selection procedure (922 vs. 1089) while the number of statistical cues is higher (246 K vs. 239 K).

For nastiness detection, the effect of semantic information from different sources is similar to that for sarcasm detection. A comparison of emotional semantic information (**SEM**) to more general semantic information (**Liwc**), shows that **SEM** provides slightly better results: a higher value of *P* was achieved for a similar *R* value. However, the combination of the two sources also results in better performance (see **Liwc_SEM** set).

Table 5 shows that combining different information sources from **ST_Li**, **ST_Liwc**, **ST_Liwc_SEM** and **ST_Lg**, does not result in better performance than **ST_C**. This suggests that specific patterns are good features for predicting nastiness and that the integration of additional information adds little value. **ST_Liwc** provides considerably better results than **Set_Lg** in this case, suggesting that length information does not contribute to nastiness detection but that semantic information does. Indeed, the fact that new nasty expressions were detected using this set is demonstrated by the high *R* value (84.7) that was achieved. Interestingly, combining the manually obtained cues from **ST_MT** with statistical features shows a small improvement over **ST_C** alone, providing the best results overall. However, again the polarity features fail to improve performance.

When all types of features are combined into the **Comb** and **All** feature sets however, the results are worse. In this case these results are also worse than **ST_C**.

6. Discussion

Our experimental results confirm our hypothesis that while sarcasm and nastiness share some properties, there are also key differences in the way they are expressed. While sarcastic language is often subtle and requires world knowledge to recognize, nastiness is often expressed more overtly and can be unambiguously detected.

However, in both cases we saw that the cost involved with obtaining manual patterns meant that the **MT_C** set is much smaller and provides worse results than **ST_C**. Although a bigger set of manually chosen patterns could possibly improve performance, it is difficult to predict how many posts would have to be hand-annotated to obtain a better set of manually chosen cues. Since the MT cues were extracted from an independent set, as explained in Section 3.1, some of these cues are not retrieved by the statistical method, and they significantly enrich the **ST_C** set. The results with the **ST_MT** feature set include a high *F*-measure that is very near to the best *F*-measure for sarcasm and is the best *F*-measure for nastiness. For sarcasm detection, there were 1815 manually obtained cues of which 252 are not in the statistical set of features. Some of them correspond to spelling issues like *"thankyou"*, *"diagree"*..., plurals (*"absolutes"* vs. *"absolute"*) or third person derivations (*"enhance"* vs. *enhances*). However, there are also some cues, not included in **ST_C** set, that are good sarcasm indicators, like "an idiot", "don't you think", "then of course". Alternatively, there were 3256 nasty cues of which about 1900 were not in the statistical set of features. This difference in the overlapping percentage might result from higher variability in the *n*-grams that express nastiness, which means they can be unambiguously identified more easily, leading to a bigger feature set. However, data sparsity is also more noticeable in this case, so the performance of the **ST_MT** set is lower than that for sarcasm.

On the other hand, the inclusion of additional information (linguistic, semantic and length) improves our ability to detect different forms of sarcastic forms, with the semantic and length information providing the most discriminant information sources.

As mentioned above, there are some forms of sarcasm that are not associated with specific cues: they can only be detected by considering the meaning of utterance. Thus, the **ST_C** feature set is able to successfully detect sarcastic posts such as *"oh they are? that's strange...i could not find where exactly they stand on any of their numerous websites, just more restrictions ...and the old hci was not moderate"* where *n*-grams like *"they are?"* or *"that's strange"* are frequently found in sarcastic posts. However, other posts such as *"because...(1) that texas has a higher rate of violent crime? (2) that texas fails to prevent such crimes? (3) that texas sees a huge fiscal deficit from such cases? strange sense of honor you got there"* was wrongly labeled as not sarcastic when **ST_C** was used but correctly detected as sarcastic with **ST_Liwc**. Furthermore, specific patterns can be used to identify sarcastic posts only when considered in combination with information about the length of the post. For instance, an isolated pattern, in a very long post, with a lot of different sentences, might not indicate sarcasm, while the same pattern would indicate sarcasm in a very short post. For example, the following post was correctly labeled as not sarcastic by **ST_Lg** whereas **ST_C** labeled it as sarcastic *"well, there are some groups that say only some gun control, but they have a hidden agenda. the brady campaign and the violence policy center, both favor complete prohibition of guns, even though they do not advocate it openly. yes i believe in some gun control, but only that which is logical. if you make it illegal, it will just create a big black market, and we all know how that is. la in california has a big black market thanks to gun control"*. In this case there are some *n*-grams like *"well,"* that can signal sarcasm in some cases, but do not function as a strong indicator when the whole post is considered. Thus, the inclusion of length information in the feature set improves performance in these cases.

An informal error analysis using the best performing classifiers suggests that misclassified posts are often those where context is needed. For example, *"your proof?"* was wrongly labeled as not sarcastic because the feature sets currently do not capture any context. There are other misclassified examples where world knowledge is essential, like *"oh it applies, but if we were to look at the bible that way, god wouldn't allow one single bad thing to happen"*. Thus, in future work it will be important to develop methods for representing context and some types of world knowledge.

Our results clearly indicate that the specific properties of the data have to be taken into account when developing feature sets for different sentiment detection tasks. Our results for the same type of online conversation suggests that different sets of features should be used to detect sarcasm in tweets, where all the utterances have a limited length, vs. in online forums which vary much more in length.

For nastiness detection, additional features representing linguistic information do not lead to better results than those obtained from statistical cues based on *n*-grams. This confirms the hypothesis of Section 3.3 and shows that nastiness is expressed more overtly, and indicated with specific cues. Thus finding such cues is sufficient for nastiness detection, regardless of the length of the utterance. Furthermore, adding linguistic and semantic information introduces noise and reduces performance. An error analysis of the misclassified examples suggests that nasty posts that are wrongly classified as nice using the **ST_C** feature set result from cases where sarcasm or figurative language are used to express nastiness, e.g., *"if the world was a figment of your imagination it would be a lot more pornographic"*. This post was properly classified as nasty when the **ST_Liwc** feature set was used. Indeed, **ST_Liwc** provide the best *R* value, although **ST** was better in terms of F. The remaining misclassified posts, when **ST_C** was used, result from the lack of contextual information.

Additionally, while it seems plausible that combining all information sources should yield better performance than pairwise combinations, this is not borne out by our experiments. This could possibly be due to the heterogeneous origin of the different sets, or possibly that the feature selection procedure significantly prunes the number of features.

7. Concluding remarks and future work

In this work, we report on the performance of different feature sets and different classifiers for the automatic detection of sarcasm and nastiness in online forums dialog. The results show that the best features are different for each emotion, although the two tasks appear to be similar a priori. Sarcasm can occur in many different forms, so the results were better when multiple sources of information were combined. Specifically, semantic information provides the best results although length information also results in significant improvement. Two semantic feature sets, one representing emotion and the other general lexical categories were shown to be complementary to each other, with their combination yielding a significant improvement. In future work, alternative ways of combining relevant knowledge should be explored, e.g., balancing features from different sources when combining them might improve our results.

Acknowledgements

Partial funding for this work was provided by NSF CISE Grant #1302668, by the Basque Government Grant IT685-13 and by CICYT Grant TIN2011-28169-C05-04.

References

[1] R. Abbott, M. Walker, P. Anand, J.E. Fox Tree, R. Bowmani, J. King, How can you say such things?!?: recognizing disagreement in informal political argument, in: Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011, pp. 2–11.
[2] A. Abu-Jbara, M. Diab, B. King, D. Radev, Identifying opinion subgroups in arabic online discussions, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2013, pp. 829–835.
[3] C. Bosco, V. Patti, A. Bolioli, Developing corpora for sentiment analysis: the case of irony and senti-tut, IEEE Intell. Syst. 28 (2) (2013) 55–63.
[4] G. Bryant, J. Fox Tree, Recognizing verbal irony in spontaneous speech, Meta. Symb. 17 (2) (2002) 99–119.
[5] E. Cambria, P. Chandra, A. Sharma, A. Hussain, Do not feel the trolls, in: Proceedings of the Third International Workshop on Social Data on the Web (SDoW2010), vol. I, 2010..
[6] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press, 2014.
[7] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intell. Syst. 28 (2) (2013) 15–21.
[8] E. Cambria, B. White, Jumping NLP curves: a review of natural language processing research, IEEE Comput. Intell. Magaz. 9 (2) (2014) 48–57.
[9] A. Charng-Rurng Tsai, C.E. Wu, R. Tzong-Han Tsai, J. Yung-jen Hsu, Building a concept-level sentiment dictionary based on commonsense knowledge, IEEE Intell. Syst. 28 (2) (2013) 22–30.
[10] B. Di Eugenio, J.D. Moore, M. Paolucci, Learning features that predict cue usage, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, vol. 97, Association for Computational Linguistics, ACL/EACL, 1997, pp. 80–87.
[11] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, R. Picard, Common sense reasoning for detection, prevention, and mitigation of cyberbullying, ACM Trans. Interact. Intell. Syst. 2 (3) (2012) 18:1–18:30.
[12] E. Forsyth, C. Martell, Lexical and discourse analysis of online chat dialog, in: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), IEEE Computer Society, 2007, pp. 19–26.
[13] L. García-Moya, H. Anaya-Sánchez, R. Berlanga-Llavori, Retrieving product features and opinions from customer reviews, IEEE Intell. Syst. 28 (3) (2013) 19–27.
[14] R. Gibbs, Irony in talk among friends, Meta. Symb. 15 (1) (2000) 5–27.
[15] R. González-Ibáñez, S. Muresan, N. Wacholder, Identifying sarcasm in twitter: a closer look, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT' 11, vol. 2, Association for Computational Linguistics, 2011, pp. 581–586.
[16] A. Hassan, V. Qazvinian, D. Radev, What's with the attitude?: identifying sentences with attitude in online discussions, in: Proceedings of the 2010

Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics; 2010, pp. 1245–1255.
[17] N. Jindal, B. Liu, Opinion spam and analysis, in: Proceedings of the 2008 International Conference on Web Search and Data Mining. Association for Computing Machinery, WSDM '08, 2008, pp. 219–230.
[18] A. Kontostathis, K. Reynolds, A. Garron, L. Edwards, Detecting cyberbullying: query terms and techniques, in: H.C. Davis, H. Halpin, A. Pentland, M. Bernstein, L.A. Adamic (Eds.), ACM Web Science 2013, Association for Computing Machinery, 2013, pp. 195–204.
[19] C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Transactions on Speech and Audio Processing 13(2) (2005) 293–303 (IEEE Signal Processing Society Best Paper Award 2009).
[20] J.M. Li, C. Cardie, Identifying manipulated offerings on review portals, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Wash., 2013, pp. 1933–1942.
[21] J. Li, C. Cardie, S. Li, Topicspam: a topic-model-based approach for spam detection, in: Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2013, pp. 217–221.
[22] C. Liebrecht, F. Kunneman, A. Van den Bosch, 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, Association for Computational Linguistics, 2013, pp. 29–37.
[23] D. Litman, J. Allen, Recognizing and relating discourse intentions and task-oriented plans, in: P. Cohen, J. Morgan, M. Pollack (Eds.), Intentions in Communication, MIT Press, 1990.
[24] S. Lukin, M. Walker, Really? Well. Apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue, in: Proceedings of the Workshop on Language Analysis in Social Media, Association for Computational Linguistics, 2013, pp. 30–40.
[25] A. Misra, M. Walker, Topic independent identification of agreement and disagreement in social media dialogue, in: Proceedings of the SIGDIAL 2013 Conference, Association for Computational Linguistics, 2013, pp. 41–50.
[26] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, R. Ghosh, Spotting opinion spammers using behavioral footprints, in: The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, 2013, pp. 632–640.
[27] M. Ott, C. Cardie, J. Hancock, Estimating the prevalence of deception in online review communities, in: Proceedings of the 21st International Conference on World Wide Web, Association for Computing Machinery, 2012, pp. 201–210.
[28] M. Ott, Y. Choi, C. Cardie, J.T. Hancock, Finding deceptive opinion spam by any stretch of the imagination, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, 2011, pp. 309–319.
[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[30] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic inquiry and word count: Liwc [computer software], Austin, TX, liwc net 2007.
[31] A. Razavi, D. Inkpen, S. Uritsky, S. Matwin, Offensive language detection using multi-level classification, Advan. Artif. Intell. (2010) 16–27.
[32] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, Sarcasm as contrast between a positive sentiment and negative situation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013, pp. 704–714.
[33] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03, Association for Computational Linguistics, 2003, pp. 105–112.
[34] S. Sood, E. Churchill, J. Antin, Automatic identification of personal insults on social news sites, J. Am. Soc. Inform. Sci. Technol. 63 (2) (2011) 270–285.
[35] E. Spertus, Smokey: automatic recognition of hostile messages, in: Proceedings of the National Conference on Artificial Intelligence, John Wiley & Sons Ltd., 1997, pp. 1058–1065.
[36] A. Stent, A conversation acts model for generating spoken dialogue contributions, Comp. Speech Lang.: Spec. Iss. Spok. Lang. Gener. 16 (3–4) (2002) 313–352.
[37] R. Subba, B. Di Eugenio, An effective discourse parser that uses rich linguistic information, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, Association for Computational Linguistics, 2009, pp. 566–574.
[38] O. Tsur, D. Davidov, A. Rappoport, 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, The AAAI Press, 2010, pp. 162–169.
[39] P. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, 2002, pp. 417–424.
[40] M. Walker, J.F. Tree, P. Anand, R. Abbott, J. King, A corpus for research on deliberation and debate, in: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), 2012, pp. 23–25.
[41] M.A. Walker, P. Anand, R. Abbott, J.E.F. Tree, C. Martell, J. King, That is your evidence?: Classifying stance in online political debate, Dec. Supp. Syst. 53 (4) (2012) 719–729.
[42] G. Wang, S. Xie, B. Liu, P.S. Yu, Identify online store review spammers via social review graph, ACM Trans. Intell. Syst. Technol. (TIST) 3 (4) (2012) 61.
[43] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Opinionfinder: a system for subjectivity analysis, in: Proceedings of HLT/EMNLP on Interactive Demonstrations, HLT-Demo '05, Association for Computational Linguistics, 2005, pp. 34–35.
[44] G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 1980–1984.