Irony Classification of Online Dialogue

Kendall Lewis
kalewis@ucsc.edu

Abstract

The use of irony in online social media dialogue is a challenge to natural language processing systems aiming to analyze and gather information from online discussions. This project extends previous works on sarcasm (Justo et al., 2014; Lukin and Walker, 2013) by analyzing the effects of using different feature combinations and classifiers on the irony detection and classification tasks. The results show that the inclusion of different feature sets improves the performance of new classifiers to varying degrees (when using SVM and JRip classifiers, not previously tested on the dataset).

1 Introduction

Irony is becoming increasingly prevalent in online dialogue as people continue to incorporate technology into their daily lives. This prevalence of irony makes the problem of irony classification a relevant one. Irony can often change, and sometimes even negate, what language understanding systems might normally infer from the text. An example of an ironic comment, in context, might be: "Going to the dentist for a root canal this afternoon. Yay, I can't wait." (Riloff et al., 2013). Such an utterance could be completely misclassified if ironic characteristics are not taken into account.

It is obvious that irony is somewhat difficult to define. Some define irony as a type of echoic mention, where speakers "echo" something that has previously been said. Others define irony as needing pretense such that the speaker takes on a different persona when speaking ironically. Another definition is that of allusional pretense where the speaker must be both insincere and allude to some failed expectation or norm. One study (Gibbs, 2000) shows that none of these definitions or requirements hold all the time; that though they are all true sometimes, there are always exceptions. This same study goes on to describe six different types of irony: jocularity (or teasing), sarcasm, hyperbole, rhetorical questions, and understatement. All of these are types of verbal irony.

Although people tend to consider verbal irony to be sarcasm alone, Gibbs (2000) has shown that sarcasm is merely one of several ironic figures of speech. Therefore, limiting the task of this project to strictly "sarcasm" would not only be difficult, it would be inaccurate. Though it is likely that most ironic language that occurs in online dialogue is sarcasm, it may not always be the case. Though defining irony or sarcasm is rather difficult, most people have some idea of what is ironic or sarcastic, even if they can't express why. Since sarcasm is ironic, but jocularity, for example, is not necessarily sarcastic, it is more accurate to refer to this task as working with the broader topic of irony, rather than merely sarcasm.

The aim of this project is to re-evaluate and analyze several aspects of previous work on irony detection and classification in online dialogue, with respect to the classifiers used, the features chosen, and the evaluation metrics and representation.

2 Related Work

The problem of sarcasm detection and classification is a difficult one, and has been explored in only limited domains in previous work, including Twitter (using #sarcasm or #irony as a method for gathering labeled tweets (González-Ibánez et al., 2011; Reyes et al., 2013)), and searching Amazon product reviews for irony (Filatova, 2012; Tsur et al., 2010; Reyes and Rosso, 2011). These previous works use a variety of features, some of which are employed here, including emotional scenarios, punctuation-based features, and polarity profiling.

Following the works of Lukin and Walker (2013) and Justo et al. (2014), the focus of this project is on online web forums,

Quote	Not only good at spelling, but good at everything I'm involved in, that goes for understanding how life should be led, I know that homosexuality is wrong, and in the same way as I use a dictionary to proves you wrong, so too does the bible, which is like a life dictionary.
Response	You're FANTASTIC at being a bigot. Not too bad at being an idiot either. Gooooooood job, buddy. A great way to go through life. emoticonXBouncer
Quote	It seems to me the corollary of the phrase "abortion is murder" is "miscarriage is manslaughter", and that any consistent pro-life advocate must urge any mother who accidentally loses her child due to actions she undertook, be charged with manslaughter. I wonder, is this actually the case?
Response	That's a good point. We need to outlaw miscarriage. After all, women knew the risk of miscarriage when they had sex.

Table 1: Examples of quote-response pairs from the dataset

primarily in the context of debate, where we predict the use of irony is prevalent enough to warrant further study. Lukin and Walker (2013) present an algorithm aimed at high precision to test a bootstrapping method to classify subjective language in the form of sarcasm and nastiness in a corpus of online social media dialogue from debate forums.

Likewise, Justo et al. (2014) use a subset of the Lukin and Walker (2013) sarcasm-annotated data and Mechanical Turk cues, and a feature set including statistical features and linguistic category features using a Naive Bayes classifier. The authors report best results using statistical N-Gram cues under the Naive Bayes classifier, with a top accuracy of 68.7%.

The results presented in this project fall directly in the range of accuracies achieved by previous work.

3 Methodology

Below we describe the dataset used and the various features utilized, and discuss the experimental design.

3.1 Dataset

The dataset used is a subset of the Internet Argument Corpus (IAC) (Walker et al., 2012) used in several previous works on sarcasm detection and classification (Lukin and Walker, 2013; Justo et al., 2014). The dataset consists of 4,820 "quote-response" (Q-R) pairs from the debate website `http://www.4forums.com/`. Each Q-R pair was annotated using Amazon Mechanical Turk by 9 annotators. The data is equally divided into two labels: *ironic* and *not_ironic*, which refers to whether irony is present in the "response" text. The *ironic* data is composed of Q-R pairs that were annotated as ironic by 6 to 9 of the 9 annotators, and the *not_ironic* data is composed of Q-R pairs that were labeled as ironic by 0 or 1 of the annotators. Table 1 presents a few examples of quote-response pairs annotated as ironic.

3.2 Feature Details

The complete feature list explored is shown below and consists of features found to be most informative in related works.

3.2.1 $\mathbf{P}_{dal}$: Dictionary of Affect in Language

This feature utilizes the Dictionary of Affect in Language (DAL) described in Whissell (2009), and attempts to quantify the degree of pleasure suggested by individual words, as well as the emotional state of a data instance. Whissell (2009) represents emotional contexts in terms of three categories: activation, imagery, and pleasantness. Activation refers to how much response humans exhibit in an emotional state and is ranked from 1 (passive) to 3 (active). Imagery is how easily one can "form a mental picture" of a word, and is ranked from 1 (difficult to envision) to 3 (easy to envision). Pleasantness quantifies the degree of pleasantness a word suggests, and is ranked from 1 (unpleasant) to 3 (pleasant). This dictionary contains more than 8,000 English words that have been scored in each of these three categories. This feature was utilized by Reyes and Rosso (2011) and Reyes et al. (2013). In both cases it is found to have strong relevance to the irony detection task.

3.2.2 P_{echo} : Echoic Overlap

One of the observations of irony is that it is a type of echoic allusion where the ironic statement "echoes" an expectation that was not met (Gibbs, 2000). While all the other features only consider the "response", this feature requires the corresponding "quote" to be able to measure the overlap.

In order to measure the overlap, stopwords are removed, then the number of words that appear in both quote and response are totaled.

3.2.3 P_{itj} : Interjections

Interjections are certain words, like *ahh*, *gee*, or *hmm*, that appear at the beginning of a data instance and are often used to express emotion. We utilized a list of interjections from `http://enchantedlearning.com/wordlist/interjections.shtml` to identify this feature. Both Carvalho et al. (2009) and Kreuz and Caucci (2007) noted that interjections can be indicators of irony and sarcasm.

3.3 P_{laugh} : Laughter Expressions

The Internet is abound with expressions and symbols that indicate various emotions. This feature looks at how often words like haha or LOL (including variations on the capitalization) occur in a data instance. It also evaluates the frequency with which emoticons are used. Carvalho et al. (Carvalho et al., 2009) found this to be a good indicator of irony with an accuracy of 85.4%.

3.3.1 P_{msol} : Polarity

To determine the polarity of a data instance, we utilize the Macquarie Semantic Orientation Lexicon (MSOL), which contains 76,400 entries labeled as either positive or negative. Since sarcasm is often defined as using positive words to convey a negative meaning, use of this lexicon reveals the correlation between the number of positive elements and negative elements in each data instance to try to determine irony. This feature was used by Reyes and Rosso (2011) where they refer to it as "positive/negative profiling".

3.3.2 P_{pc} : Punctuation and Capitalization

This feature has five components: (i) the frequency of exclamation marks and question marks within all sentences in a data instance; (ii) how frequently those same characters are used in repetition, such as "!!!" or "???", as well as the frequency of ellipses ("..."); (iii) whether or not quotation marks are used; (iv) whether any special characters occur, such as "@" or "*"; and (v) capitalization, which refers to words with two or more characters where all characters are capitalized.

In the case of capitalization, some matches may indicate an abbreviation, but most often it's an indication of an implied emphasis that could represent an ironic statement. Various aspects of this feature have been shown to be indicators of irony and sarcasm by both Tsur et al. (2010) and Kreuz and Caucci (2007).

3.4 Experimental Design

For these experiments we used the built-in classifiers in the Weka toolkit. Naive Bayes was employed as a simple baseline, while SVM was employed as a more powerful, standard classifier, and JRip as a rule-based classifier that allows interpretation of which features appear to be the most useful.

Following previous work, the different classifiers and feature combinations discussed will be tested incrementally to gauge their effectiveness.

The standard metrics (precision, recall, and F-measure) will be used to evaluate the experiments against each others. We used 10-fold cross-validation to more closely resemble experiments performed in previous work.

4 Results

Below are presented the results of various experiments conducted, using different combinations of the features and classifiers under analysis.

It's clear from Table 2 that none of the feature types on their own do particularly well. We can see from Table 3 that even Naive Bayes achieves a higher F-measure when using multiple features in combination than any of the feature types do individually.

Laughter expressions achieves the highest precision when using Naive Bayes or JRip, but the lowest precision when using SVM. That is because SVM chooses to classify every example as *ironic*, which results in an accuracy of 50%.

After laughter expressions, polarity achieves the next highest precision. Polarity also has the highest recall for all classifiers, resulting in it having the highest F-measure for all classifiers. The feature type that achieves the third highest precision is echoic overlap. These observations are what

Feature	Naive Bayes			JRip			SVM		
	P	R	F-meas	P	R	F-meas	P	R	F-meas
P_{dal}	0.608	0.596	0.585	0.588	0.588	0.587	0.587	0.582	0.575
P_{echo}	0.637	0.593	0.556	0.626	0.626	0.625	0.635	0.596	0.565
P_{itj}	0.589	0.551	0.498	0.569	0.550	0.516	**0.636**	0.560	0.487
P_{laugh}	**0.718**	0.584	0.509	**0.721**	0.592	0.522	0.250	0.500	0.333
P_{msol}	0.688	**0.624**	**0.589**	0.664	**0.663**	**0.663**	0.686	**0.641**	**0.618**
P_{pc}	0.618	0.595	0.574	0.601	0.601	0.601	0.608	0.576	0.543

Table 2: Performance of various classifiers on individual feature types

Features	Naive Bayes			JRip			SVM		
	P	R	F-meas	P	R	F-meas	P	R	F-meas
P_{laugh},P_{msol}	0.699	0.634	0.602	0.682	0.682	0.682	0.691	0.653	0.635
P_{laugh},P_{msol},P_{echo}	0.693	0.638	0.609	0.678	0.678	0.678	0.689	0.663	0.651
P_{dal},P_{echo},P_{laugh},P_{msol},P_{pc},P_{itj}	0.705	0.664	0.647	0.708	0.708	0.708	0.733	0.733	0.733
P_{dal},P_{echo},P_{laugh},P_{msol},P_{pc}	0.702	0.651	0.628	**0.718**	**0.718**	**0.718**	0.725	0.725	0.725
20 Most Informative	**0.707**	**0.666**	**0.649**	0.711	0.711	0.711	**0.738**	**0.738**	**0.738**

Table 3: Performance of various classifiers on different feature combinations

motivated the first two feature combinations that appear in Table 3. Looking at these feature combinations, we can see that when using Naive Bayes or JRip, neither of the feature combinations can achieve a precision score as high as that of laughter expressions on its own. Additionally, we see that including the echoic overlap feature type in the combination only reduces the precision further.

The third feature combination in Table 3 includes all six feature types. Given that each feature type was evaluated individually, we thought it would be an interesting comparison to see how all feature types performed together. We can see that though it doesn't achieve a precision as high as that of laughter expressions, it is still an improvement over the previous combinations.

Since interjections most often performed the worst of the individual feature types, we wanted to see if removing it from the combination of all feature types had any impact. When using Naive Bayes or SVM, the precision, recall, and F-measure all decreased, however, when using JRip, all metrics increased.

Given that there are six feature types, we experimented with various combinations of them. Though additional feature combinations were tested, we found that the four combinations included in Table 3 performed the best. Interestingly enough, the choice of feature types included in those four combinations were motivated by the results presented in Table 2.

Using χ^2 feature selection across all feature types provided a ranking of all individual features, which in turn provided a list of the 20 most informative features, which appears in Table 4. The results of using these 20 features in combination can be seen in Table 3. We can see that both Naive Bayes and SVM produce their highest accuracy across all experiments when given these 20 most informative features.

1. P_{msol}: positive	11. P_{pc}: repeated
2. P_{msol}: negative	12. P_{dal}: activation
3. P_{echo}	13. P_{laugh}: words
4. P_{laugh}: emoticons	14. P_{itj}: oh
5. P_{dal}: pleasantness	15. P_{pc}: quotes
6. P_{pc}: contains "!"	16. P_{itj}: well
7. P_{dal}: imagery	17. P_{itj}: wow
8. P_{pc}: contains "?"	18. P_{itj}: yeah
9. P_{pc}: capitalized words	19. P_{itj}: no
10. P_{itj}: there	20. P_{itj}: hey

Table 4: 20 most informative features

When comparing the predictions made by each of the classifiers, we found that there is a subset of examples that were almost always misclassified, regardless of the feature type or classifier being used.

It's interesting to note that despite echoic overlap being one of the possible identifiers of irony, the majority of utterances with echoic overlap were actually labeled *not_ironic*. Additionally, though we expected utterances with positive polarity to be indicative of irony, utterances with higher scores were mostly labeled *not_ironic*.

5 Conclusion

The features used here were motivated by those presented in related works and theoretical litera-

ture. We've proven that these features still prove useful for this classification task, despite using a dataset taken from a relatively different source. We've been able to show that while none of the features on their own are particularly informative, they receive relatively high precision when used in combination, and that our classification task can achieve precision and F-measure scores as high as 0.738.

References

Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, pages 392–398.

Raymond W Gibbs. 2000. Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 581–586. Association for Computational Linguistics.

Raquel Justo, Thomas Corcoran, Stephanie M Lukin, Marilyn Walker, and M Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Roger J Kreuz and Gina M Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pages 1–4. Association for Computational Linguistics.

Stephanie Lukin and Marilyn Walker. 2013. 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*, pages 30–40.

Antonio Reyes and Paolo Rosso. 2011. Mining subjective knowledge from customer reviews: a specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 118–124. Association for Computational Linguistics.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2):509–521.