Are you serious?: Rhetorical Questions and Sarcasm in Social Media Dialog

Shereen Oraby[1], Vrindavan Harrison[1], Amita Misra[1], Ellen Riloff [2] and Marilyn Walker[1]

[1] University of California, Santa Cruz

[2] University of Utah

{soraby,vharriso,amisra2,mawalker}@ucsc.edu

riloff@cs.utah.edu

Abstract

Effective models of social dialog must understand a broad range of rhetorical and figurative devices. Rhetorical questions (**RQs**) are a type of figurative language whose aim is to achieve a pragmatic goal, such as structuring an argument, being persuasive, emphasizing a point, or being ironic. While there are computational models for other forms of figurative language, rhetorical questions received little attention to date. We expand a small dataset from previous work, presenting a corpus of 10,270 RQs from both debate forums and Twitter that represent different discourse functions. We show that we can clearly distinguish between RQs and sincere questions (0.76 F1), and then that RQs can be used both sarcastically and non-sarcastically. We observe that non-sarcastic (other) uses of RQs are frequently argumentative in forums, and persuasive in tweets. We present experiments to distinguish between these uses of RQs using the SVM and LSTM models that represent linguistic features and post-level context, achieving results as high as 0.76 F1 for SARCASTIC and 0.77 F1 for OTHER in forums, and 0.83 F1 for both SARCASTIC and OTHER in tweets. We supplement our quantitative experiments with an in-depth characterization of the linguistic variation in RQs.

1 Introduction

Theoretical frameworks for figurative language posit eight standard forms: *indirect questions, idiom, irony and sarcasm, metaphor, simile, hyperbole, understatement,* and *rhetorical questions*

1	**Then why do you call a politician who ran such measures liberal** *OH yes, it's because you're a republican and you're not conservative at all.*
2	**Can you read?** *You're the type that just waits to say your next piece and never attempts to listen to others.*
3	**Pray tell, where would I find the atheist church?** *Ridiculous.*
4	**You lost this debate Skeptic, why drag it back up again?** *There are plenty of other subjects that we could debate instead.*

(a) RQs in Forums Dialog

5	**Are you completely revolting?** *Then you should slide into my DMs, because apparently thats the place to be. #Sarcasm*
6	**Do you have problems falling asleep?** *Reduce anxiety, calm the mind, sleep better naturally [link]*
7	**The officials messed something up?** *I'm shocked I tell you.SHOCKED.*
8	**Does ANY review get better than this?** *From a journalist in New York.*

(b) RQs in Twitter Dialog

Table 1: RQs and Following Statements in Forums and Twitter Dialog

(Roberts and Kreuz, 1994). While computational models have been developed for many of these forms, rhetorical questions (**RQs**) have received little attention to date. Table 1 shows examples of RQs from social media in debate forums and Twitter, where their use is prevalent.

RQs are defined as utterances that have the structure of a question, but which are *not intended* to seek information or elicit an answer (Sadock, 1971; Rohde, 2006; Frank, 1990; Ilie, 1994). RQs are often used in arguments and expressions of opinion, advertisements and other persuasive domains (Petty et al., 1981), and are frequent in social media and other types of informal language.

Corpus creation and computational models for

some forms of figurative language have been facilitated by the use of hashtags in Twitter, e.g. the `#sarcasm` hashtag (Riloff et al., 2013; Bamman and Smith, 2015; Liebrecht et al., 2013). Other figurative forms, such as similes, can be identified via lexico-syntactic patterns (Qadir et al., 2015; Veale and Hao, 2007; Qadir et al., 2016). RQs are not marked by a hashtag, and their syntactic form is indistinguishable from standard questions (Han, 2002; Sadock, 1971).

Previous theoretical work examines the discourse functions of RQs and compares the overlap in discourse functions across all forms of figurative language (Roberts and Kreuz, 1994). For RQs, 72% of subjects assign *to clarify* as a function, 39% assign *discourse management*, 28% mention *to emphasize*, 56% percent of subjects assign negative emotion, and another 28% mention positive emotion.[1] The discourse functions of clarification, discourse management and emphasis are clearly related to argumentation. One of the other largest overlaps in discourse function between RQs and other figurative forms is between RQs and irony/sarcasm (62% overlap), and there are many studies describing how RQs are used sarcastically (Gibbs, 2000; Ilie, 1994).

To better understand the relationship between RQs and irony/sarcasm, we expand on a small existing dataset of RQs in debate forums from our previous work (Oraby et al., 2016), ending up with a corpus of 2,496 RQs and the self-answers or statements that follow them. We use the heuristic described in that work to collect a completely novel corpus of 7,774 RQs from Twitter. Examples from our final dataset of 10,270 RQs and their following self-answers/statements are shown in Table 1. We observe great diversity in the use of RQs, ranging from sarcastic and mocking (such as the forum post in Row 2), to offering advice based on some anticipated answer (such as the tweet in Row 6).

In this study, we first show that RQs can clearly be distinguished from sincere, information-seeking questions (0.76 F1). Because we are interested in how RQs are used sarcastically, we define our task as distinguishing sarcastic uses from other uses RQs, observing that non-sarcastic RQs are often used argumentatively in forums (as opposed to the more mocking sarcastic uses), and persua-

sively in Twitter (as frequent advertisements and calls-to-action). To distinguish between sarcastic and other uses, we perform classification experiments using SVM and LSTM models, showing that adding linguistic features and post-level context improve classification results in both domains.

This paper provides the first in-depth investigation of the use of RQs in different forms of social media dialog. We present a novel task, dataset[2], and results aimed at understanding how RQs can be recognized, and how sarcastic and other uses of RQs can be distinguished.

2 Related Work

Much of the previous work on RQs has focused on RQs as a form of figurative language, and on describing their discourse functions (Roberts and Kreuz, 1994; Gibbs, 2000; Petty et al., 1981; Frank, 1990; Schaffer, 2005). Related work in linguistics has primarily focused on the differences between RQs and standard questions (Han, 2002; Ilie, 1994; Han, 1997). For example Sadock (1971) shows that RQs can be followed by a *yet* clause, and that the discourse cue *after all* at the beginning of the question leads to its interpretation as an RQ. Phrases such as *by any chance* are primarily used on information seeking questions, while negative polarity items such as *lift a finger* or *budge an inch* can only be used with RQs, e.g. *Did John help with the party?* vs. *Did John lift a finger to help with the party?*

RQs were introduced into the DAMSL coding scheme when it was applied to the Switchboard corpus (Jurafsky et al., 1997). To our knowledge, the only computational work utilizing that data is by Battasali et al. (2015), who used n-gram language models with pre- and post-context to distinguish RQs from regular questions in SWBD-DAMSL. Using context improved their results to 0.83 F1 on a balanced dataset of 958 instances, demonstrating that context information could be very useful for this task.

Although it has been observed in the literature that RQs are often used sarcastically (Gibbs, 2000; Ilie, 1994), previous work on sarcasm classification has not focused on RQs (González-Ibáñez et al., 2011; Riloff et al., 2013; Liebrecht et al., 2013; Filatova, 2012; Davidov et al., 2010; Tsur et al., 2010; Bamman and Smith, 2015). Riloff

---

[1] Subjects could provide multiple discourse functions for RQs, thus the frequencies do not add to 1.

[2] The Sarcasm RQ corpus will be available at: `https://nlds.soe.ucsc.edu/sarcasm-rq`.

et al. (2013) investigated the utility of sequential features in tweets, emphasizing a subtype of sarcasm that consists of an expression of positive emotion contrasted with a negative situation, and showed that sequential features performed much better than features that did not capture sequential information. More recent work on sarcasm has focused specifically on sarcasm identification on Twitter using neural network approaches (Poria et al, 2016; Ghosh and Veale, 2016; Zhang et al., 2016; Amir et al., 2016).

Other work emphasizes features of semantic incongruity in recognizing sarcasm (Reyes et al., 2012; Joshi et al., 2015). Sarcastic RQs clearly feature semantic incongruity, in some cases by expressing the certainty of particular facts in the frame of a question, and in other cases by asking questions like *"Can you read?"* (Row 2 in Table 1), a competence which a speaker must have, prima facie, to participate in online discussion.

To our knowledge, our previous work is the first to consider the task of distinguishing sarcastic vs. not-sarcastic RQs, where we construct a corpus of sarcasm in three types: generic, RQ, and hyperbole, and provide simple baseline experiments using ngrams (0.70 F1 for SARC and 0.71 F1 for NOT-SARC) (Oraby et al., 2016). In this work, we adopt the same heuristic for gathering RQs and expand their corpus in debate forums, and also collect a novel Twitter corpus. We show that we can distinguish between SARCASTIC and OTHER uses of RQs, as we observe many non-sarcastic uses of RQs, such as argumentation and persuasion in forums and Twitter, respectively. We show that linguistic features aid in the classification task, and explore the effects of context, using traditional and neural models.

3 Corpus Creation

Sarcasm is a prevalent discourse function of RQs. In previous work, we observe both sarcastic and not-sarcastic uses of RQs in forums, and collect a set of sarcastic and not-sarcastic RQs in debate by using a heuristic stating that an RQ is a question that occurs in the middle of a turn, and which is answered immediately by the speaker themselves (Oraby et al., 2016). RQs are thus defined *intentionally*: the speaker indicates that their intention is not to elicit an answer by not ceding the turn.[3]

---

[3]We acknowledge that this method may miss RQs that do not follow this heuristic, but opt to use this conservative pat-

SARCASTIC	
1	**Do you even read what anyone posts?** *Try it, you might learn something.......maybe not.......*
2	**If they haven't been discovered yet, HOW THE BLOODY HELL DO YOU KNOW?** *Ten percent more brains and you'd be pondlife.*

OTHER	
3	**How is that related to deterrence?** *Once again, deterrence is preventing through the fear of consequences.*
4	**Well, you didn't have my experiences, now did you?** *Each woman who has an abortion could have innumerous circumstances and experiences.*

(a) SARC vs. OTHER RQs in Forums

SARCASTIC	
5	**When something goes wrong, what's the easiest thing to do?** *Blame the victim! Obviously they had it coming #sarcasm #itsajoke #dontlynchme*
6	**You know what's the best?** *Unreliable friends. They're so much un. #sarcasm #whatever.*

OTHER	
7	**And what, Socrates, is the food of the soul?** *Surely, I said, knowledge is the food of the soul. Plato*
8	**Craft ladies, salon owners, party planners?** *You need to state your #business [link]*

(b) SARC vs. OTHER RQs in Twitter

Table 2: Sarcastic vs. Other Uses of RQs

In this work, we are interested in doing a closer analysis of RQs in social media. We use the same RQ-collection heuristic from previous work to expand our corpus of SARCASTIC vs. OTHER uses RQs in debate forums, and create another completely novel corpus of RQs in Twitter. We observe that the other uses of RQs in forums are often argumentative, aimed at structuring an argument more emphatically, clearly, or concisely, whereas in Twitter they are frequently persuasive in nature, aimed at advertising or grabbing attention. Table 2 shows examples of sarcastic and other uses of RQs in our corpus, and we describe our data collection methods for both domains below.

Debate Forums: The Internet Argument Corpus (IAC 2.0) (Abbott et al., 2016) contains a large number of discussions about politics and social issues, making it a good source of RQs. Following our previous work (2016), we first extract RQs in

---

tern for expanding the data to avoid introducing extra noise.

posts whose length varies from 10-150 words, and collect five annotations for each of the RQs paired with the context of their following statements.

We ask Turkers to specify whether or not the RQ-response pair is sarcastic, as a binary question. We count a post as "sarcastic" if the majority of annotators (at least 3 of the 5) labeled the post as sarcastic. Including the 851 posts per class from previous work (Oraby et al., 2016), this resulted in 1,248 sarcastic posts out of 4,840 (25.8%), a significantly larger percentage than the estimated 12% sarcasm ratio in debate forums (Swanson et al., 2014). We then balance the 1,248 sarcastic RQs with an equal number of RQs that 0 or 1 annotators voted as sarcastic, giving us a total of 2,496 RQ pairs. For our experiments, all annotators had above 80% agreement with the majority vote.

Twitter: We also extract RQs defined as above from a set of 80,000 tweets with a `#sarcasm`, `#sarcastic`, or `#sarcastictweet` hashtag. We use the hashtags as "labels", as in other work (Riloff et al., 2013; Reyes et al., 2012). This yields 3,887 sarcastic RQ tweets, again balanced with 3,887 RQ pairs from a set of random tweets (not containing any sarcasm-related hashtags). We remove all sarcasm-related hashtags and username mentions (prefixed with an "@") from the posts, for a total of 7,774 total RQ tweets.

4 Experimental Results

In this section, we present experiments classifying rhetorical vs. information-seeking questions, then sarcastic vs. other uses of RQs.

4.1 RQs vs. Information-Seeking Qs

By definition, fact-seeking questions are not RQs. We take advantage of the annotations provided for subsets of the IAC, in particular the sub-corpus that distinguishes FACTUAL posts from EMOTIONAL POSTS (Oraby et al., 2015; Abbott et al., 2016).[4] Table 3 shows examples of FACTUAL/INFO-SEEKING questions.

To test whether RQ and FACTUAL/INFO-SEEKING questions are easily distinguishable, we randomly select 1,020 questions from our forums RQ corpus, and balance them with the same number of questions from FACT corpus. We divide the question data into 80% train and 20% test, and use

---

[4]https://nlds.soe.ucsc.edu/factfeel

	FACTUAL/INFO-SEEKING QUESTIONS
1	How do you justify claims about covering only a fraction more ?
2	If someone is an attorney or in law enforcement, would you please give an interpretation?

Table 3: Examples of Information-Seeking Questions

an SVM classifier (Pedregosa et al., 2011), with GoogleNews Word2Vec (W2V) (Mikolov et al., 2013) features. We perform a grid-search on our training set using 3-fold cross-validation for parameter tuning, and report results on our test set. Table 4 shows the precision (P), recall (R) and F1 scores we achieve, showing good classification performance for distinguishing both classes, at 0.76 F1 for the RQ class, and 0.74 F1 for the FACTUAL/INFO-SEEKING class.

#	Class	P	R	F1
1	RQ	0.74	0.79	0.76
2	FACT	0.77	0.72	0.74

Table 4: Supervised Learning Results for RQs vs. Fact/Info-Seeking Questions in Debate Forums

4.2 Sarcastic vs. Argumentative/Persuasive RQs

Next, we focus on distinguishing SARCASTIC from OTHER uses of RQs in forums and Twitter. We divide the full RQ data from each domain (2,496 forums and 7,774 tweets, balanced between the two classes) into 80% train and 20% test data. We experiment with two models, an SVM classifier from Scikit Learn (Pedregosa et al., 2011), and a bidirectional LSTM model (Chollet, 2015) with a TensorFlow backend (Abadi et al., 2016). We perform a grid-search using cross-validation on our training set for parameter tuning, and report results on our test set.

For each of the models, we establish a baseline with W2V features (Google News-trained Word2Vec size 300 (Mikolov et al., 2013) for the debate forums, and Twitter-trained Word2Vec size 400 (Godin et al., 2015), for the tweets). We experiment with different word embedding representations, finding that we achieve best results by averaging the word embeddings for each input when using SVM, and creating an embedding matrix
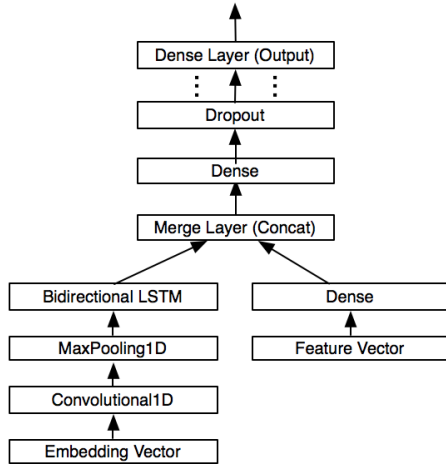
Figure 1: LSTM Network Architecture

(number of words by embedding size for each input) as input to an embedding layer when using LSTM.[5]

For our LSTM model, we experiment with various different layer architectures from previous work (Poria et al, 2016; Ghosh and Veale, 2016; Zhang et al., 2016; Amir et al., 2016). For our final model (shown in Figure 1), we use a sequential embedding layer, 1D convolutional layer, max-pooling, a bidirectional LSTM, dropout layer, and a sequence of dense and dropout layers with a final sigmoid activation for the output.

For additional features, we experiment with using post-level scores (frequency of each category in the input, normalized by word count) from the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2001). We experiment with which LIWC categories to include as features on our training data, and end up with a set of 20 categories for each domain[6], as shown in Table 5. When adding features to the LSTM model, we include a dense and merge layer to concatenate features, followed by the dense and dropout layers and sigmoid output.

We experiment with different levels of textual context in training for both the forums and Twitter data (keeping our test set constant, testing on only the RQ and self-answer portion of the text). We are motivated by the intuition that training on larger context will help us identify more informa-

---

[5]In future work, we plan to further explore the effects of different embedding representations on model performance.

[6]We discuss some of the highly-informative LIWC categories by domain in Sec. 5.

Debate Forums	Tweets
2nd PERSON	2nd PERSON
3rd PERSON PLURAL	3rd PERSON PLURAL
3rd PERSON SINGULAR	ARTICLES
ADVERBS	AUXILIARY VERBS
AFFILIATION	CERTAINTY
ASSENT	COLON
AUXILIARY VERBS	COMMA
COMPARE	CONJUNCTION
EXCLAMATION MARKS	FRIENDS
FOCUS FUTURE	MALE
FRIENDS	NEGATIONS
FUNCTION	NEGATIVE EMOTION
HEALTH	PARENTHESIS
INFORMAL	QUOTE MARKS
INTERROGATIVES	RISK
NETSPEAK	SADNESS
NUMERALS	SEMICOLON
QUANTIFIERS	SWEAR WORDS
REWARDS	WORD COUNT
SADNESS	WORDS PER SENTENCE

Table 5: LIWC Features by Domain

tive segments of RQs in test. Specifically, we test four different levels of context representation:

- RQ: only the RQ and its self-answer
- $Pre+RQ$: the preceding context and the RQ
- $RQ + Post$: the RQ and following context
- $FullText$: the full text or tweet (all context)

Table 6 presents our results on the classification task by model for each domain, showing P, R, and F1 scores for each class (Forums in Table 6a and Twitter in Table 6b). For each domain, we present the same experiments for both models (SVM and LSTM), first showing a W2V baseline (Rows 1 and 6 in both tables), then adding in LIWC (Rows 2 and 7), and finally presenting results for W2V and LIWC features on different context levels (Rows 2-5 for SVM and Rows 7-10 for LSTM).

Debate Forums: From Table 6a, for both models, we observe that the addition of LIWC features gives us a large improvement over the baseline of just W2V features, particularly for the SARC class (from 0.72 F1 to 0.76 F1 SARC and 0.73 F1 to 0.77 F1 OTHER for SVM in Row 2, and from 0.68 F1 to 0.72 F1 SARC and 0.74 F1 to 0.75 F1 OTHER for LSTM in Row 7). Our best results come from the SVM model, with best scores of 0.76 F1 for SARC

						SARCASTIC			OTHER		
#	Domain	Model	Features	Training		P	R	F1	P	R	F1
1	**Forums**	SVM	$W2V_{Google}$	RQ		0.74	0.70	0.72	0.71	0.75	0.73
2			$W2V_{Google}+LIWC$	RQ		0.78	0.74	**0.76**	0.75	0.79	**0.77**
3				$Pre+RQ$		0.73	0.72	0.74	0.73	0.78	0.76
4				$RQ+Post$		0.75	0.76	0.75	0.76	0.74	0.75
5				$Full\ Text$		0.75	0.77	**0.76**	0.76	0.74	0.75
6		LSTM	$W2V_{Google}$	RQ		0.76	0.62	0.68	0.68	0.80	0.74
7			$W2V_{Google}+LIWC$	RQ		0.76	0.68	0.72	0.71	0.79	0.75
8				$Pre+RQ$		0.81	0.60	0.69	0.68	0.86	0.76
9				$RQ+Post$		0.74	0.76	0.75	0.76	0.74	0.75
10				$Full\ Text$		0.76	0.67	0.71	0.70	0.78	0.74

(a) Supervised Learning Results on Debate Forums

						SARCASTIC			OTHER		
#	Domain	Model	Features	Training		P	R	F1	P	R	F1
1	**Twitter**	SVM	$W2V_{Tweet}$	RQ		0.77	0.85	0.80	0.83	0.74	0.78
2			$W2V_{Tweet}+LIWC$	RQ		0.80	0.86	**0.83**	0.85	0.79	0.82
3				$Pre+RQ$		0.80	0.87	**0.83**	0.86	0.78	0.82
4				$RQ+Post$		0.79	0.87	**0.83**	0.86	0.77	0.81
5				$Full\ Text$		0.80	0.86	**0.83**	0.85	0.79	0.82
6		LSTM	$W2V_{Tweet}$	RQ		0.76	0.70	0.73	0.72	0.78	0.75
7			$W2V_{Tweet}+LIWC$	RQ		0.80	0.82	0.81	0.82	0.79	0.80
8				$Pre+RQ$		0.78	0.84	0.81	0.83	0.76	0.80
9				$RQ+Post$		0.83	0.81	0.82	0.82	0.84	**0.83**
10				$Full\ Tweet$		0.80	0.83	0.82	0.83	0.79	0.81

(b) Supervised Learning Results on Twitter

Table 6: Supervised Learning Results for RQs in Debate Forums and Twitter

and 0.77 OTHER from using only the RQ and self-response in training.

We observe that while the SVM results with LIWC features do not change significantly depending on the training context (Rows 3-5), the LSTM model is highly sensitive to context changes for the SARC class (Rows 8-10). Some interesting findings emerge when training on different context granularities for LSTM: our best LSTM results for the SARC class come from training on the $RQ+Post$ context (0.75 F1 in Row 9), and for the $Pre+RQ$ context for the ARGUE class (0.76 F1 in Row 8). We note that this increase in the SARC class from plain word embeddings to word embeddings combined with LIWC and context is larger than the increase in the OTHER class, indicating that post-level context for SARC captures more diverse instances in training. We also

note that these results beat our previous baselines using only ngram features on the smaller original dataset of 851 posts per class (0.70 F1 for SARC, 0.71 F1 for NOT-SARC) (Oraby et al., 2016).

We investigate why certain context features benefit each class differently for LSTM. Table 7 shows examples of single posts, divided into Pre, RQ, and $Post$. Looking at Row 1, it is clear that while the RQ and self-answer portion may not appear to be sarcastic, the $Post$ context makes the sarcasm much more pronounced. This is frequent in the case of sarcastic debate posts, where the speaker ends with a sharp remark or an interjection (like *"gasp!!!"*), or emoticons (like winking *;)* or roll-eyes *8-)*). In the case of the OTHER forums posts, the RQ is often nestled within sequences of a questions, or other RQ and self-answer pairs seeking to inform (Row 2).

SARCASTIC

1	Pre	[...] the argument I hear most often from so-called 'pro-choicers' is that you cannot legislate morality.
	RQ	**Well then what can you legislate?** *Every law in existence is legislation of morality!*
	Post	By that way of thinking, then we should have no laws. If someone kidnaps and murders your 3-year-old child, then let's hope the murderer goes free because we cannot legislate morality!

OTHER

2	Pre	what that man did isn't illegal in the us? you couldn't claim self defence if someone running away like that.
	RQ	**you think that the fact that man had a gun stopped people getting shot?** *what would have happened if he hadn't would be that the robbers got away with some money.*
	Post	nothing to do with taking lives. [...]

(a) SARC vs. OTHER RQs in Context on Forums

SARCASTIC

3	Pre	Gasp!
	RQ	**Two football players got into it with each other?!** *How uncivilized!*
	Post	Lets make a big deal about it! #NFLlogic #cowboys

OTHER

4	Pre	
	RQ	**Are you willing to succeed?** *The answer isn't as simple as you may think.*
	Post	Read my blog post and you'll see why.... [link]

(b) SARC vs. OTHER RQs in Context on Twitter

Table 7: Sarcastic vs. Other Uses RQs in Context

Twitter: From Table 6b, we observe that the best result of 0.83 F1 for the SARC class come from the SVM, while the best result of 0.83 F1 for the OTHER class comes from the LSTM model. We observe a strong performance increase from adding in LIWC features for both models, even more pronounced than for forums (0.80 F1 to 0.83 F1 SARC and 0.78 F1 to 0.82 F1 OTHER for SVM, and 0.73 F1 to 0.81 F1 SARC and 0.75 F1 to 0.80 F1 OTHER for LSTM).

Again, while the SVM results do not vary significantly due to context, there is a large improvement in the OTHER class for LSTM when using $RQ + Post$ level context, giving us our best OTHER class results. From Table 9 Row 4, we see

an example of a "call-to-action" that are frequent and distinctive in persuasive Twitter RQs, asking users to visit a link at the end of a tweet ($Post$ RQ). In the case of the SARC tweet in Row 3, the extra tweet-level context (such as initial exclamations/interjections) aids in highlighting the sarcasm, but is limited in length compared to the forums posts, explaining the smaller gain from context in the Twitter domain for SARC.

Comparing both domains, we observe that the results for tweets in Table 6b are much higher than the results for forums in Table 6a, noting several possible reasons that we plan to explore more extensively in future work: less lexical diversity and a larger amount of data, making them more distinguishable than the more varied forums posts.

5 Linguistic Characteristics of RQs by Class and Domain

In this section, we discuss linguistic characteristics we observe in our SARCASTIC vs OTHER uses of RQs using the most informative LIWC (Pennebaker et al., 2001) features.

Previous work has observed that FACTUAL utterances are often very heavy on technical jargon (Oraby et al., 2015): this is also true of factual questions. When analyzing differences in LIWC categories in our factual vs. RQ data, that our factual questions are longer on average than the RQs (14 words on average compared to 12). We also find significant differences in "function" word categories ($p < 0.05$, unpaired t-test) in LIWC (Pennebaker et al., 2001), marking use of personal references, and "affective processes" ($p < 0.005$). Both categories are more prevalent in the RQs than in the FACT questions, indicating more emotional language that is targeted towards the second party.

A qualitative analysis of our SARCASTIC vs. OTHER data shows that sarcastic RQs in forums are often followed by short statements that serve to point attention or mock, whereas the argumentative RQ-self-response pairs serve as a technique to concisely structure an argument. RQs in Twitter are frequently advertisements (persuasive communication) (Petty et al., 1981), making them more distinguishable from the more diverse sarcastic instances. Tables 8 and 9 show examples of LIWC features that are most characteristic of each domain and class based on our experiments. For ranking, we show the learned feature

Table 8: Forums LIWC Categories

SARCASTIC

#	FW	Feature	Example
1	15.19	2nd Person	**Do you ever read headers?** *You got a mouth on you as big as grand canyon.*
2	12.09	Informal	**The hate you're spewing is palpable, yet you can't even see that can you?** *Hypocrites, ya gotta luv em.*
3	8.92	Exclamation	**Force the children to learn science?** *How obscene!!*
4	4.66	Netspeak	**To make fun of my title?** *lol, how that stings...*

OTHER

#	FW	Feature	Example
5	8.98	Interrog.	**How do you know it's the truth?** *If it were definitive [...]*
6	8.54	3rd Person Plural	**what's the difference?** *both are imposing their ideologies*
7	3.93	Quantifiers	**[...] we have minimum wage, why can't we have a maximum wage?** *some of [...]*
8	3.88	Health	**When will the people press congress to take up abortion?** *It's the job of congress [...]*

Table 9: Tweet LIWC Categories

SARCASTIC

#	FW	Feature	Example
1	15.71	Comma	**Wait, wait, I can't...it's impossible...NO WAY?!** *- a stiffer track pad?!*
2	6.86	Word Count	**Shouldn't you be in power?** *You know best after all.*
3	5.89	Negations	**Can't we do that already without brain imaging?** *I think it's called empathy*
4	3.91	3rd Person Plural	**How intelligent, they make the laws and then violate [them]?** *That is absurd!*

OTHER

#	FW	Feature	Example
5	4.51	Swear Words	**Idk why I'm fighting my sleep?!** *Ain't shit else to do*
6	3.60	Risk	**Have their been launch pad explosions?** *That would be a risk.*
7	3.01	2nd Person	**Do you want a great deal on [...]?** *Check out the latest*
8	2.83	Friends	**Can I get 12.7k followers today?** *:) xo Thanks to everyone who is following me.*

weight (FW) for each class, found by performing 10-fold cross-validation on each training set using an SVM model with only LIWC features.

In Table 8, Row 1, we observe that 2nd person mentions are frequent in the sarcastic debate forums posts (referring to the other person in the debate), while in the Twitter domain, they come up as significant features in the *non-sarcastic* tweets, where they are used as methods to persuade readers to interact: click a link, like, comment, share (Table 9, Row 6). Likewise, "informal" words and more "verbal speech style" non-fluencies, including exclamations and social media slang ("netspeak"), also appear in sarcastic debate (Table 8, Rows 2 and 4). Features of sarcastic tweets include exclamations (Table 9, Rows 3), often used in a hyperbolic or figurative manner (McCarthy and Carter, 2004; Roberts and Kreuz, 1994).

Table 8 shows that "interrogatives" are a strong feature of argumentative forums (Row 7), as well as the use of technical jargon (including quantifiers health words with some domain-specific topics, such as abortion) (Row 8). Table 9 indicates that persuasive tweets frequently contain forms of advertisement and calls-to-action involving 2nd person references (Row 7). Similarly, RQ tweets are sometimes used to express frustration ("swear words" in Row 5), or increase engagement with references to "friends" and followers (Row 8).

6 Conclusions

In this study, we expand on a small corpus from previous work to create a large corpus of RQs in two domains: debate forums and Twitter. To our knowledge, this is the first in-depth study dedicated to sarcasm and other uses of RQs in social media. We focus on two domains where RQs are prevalent, debate forums and Twitter, and present supervised learning experiments using traditional and neural models to classify sarcasm in each domain, providing analysis of unique features across domains and classes.

We first show that we can distinguish between information-seeking and rhetorical questions (0.76 F1). We then focus on classifying sarcasm in only the RQs, showing that there are distinct linguistic differences between the methods of expression used in RQs across forums and Twitter. For forums, we show that we are able to distinguish between the sarcastic and other uses (noting they are often argumentative) in forums with 0.76 F1 for SARC and 0.77 F1 for NOT-SARC, improving on our baselines from previous work on a smaller dataset (2016).

We also explore sarcastic and other uses of RQs on Twitter, noting that other non-sarcastic uses of RQs are often advertisements, a form of persuasive communication not represented in debate dialog. We show that we can distinguish between sarcastic and other uses of RQ in Twitter with scores of 0.83 F1 for both the SARC and PERSUADE classes. We observe that tweets are generally more easily discriminated than the more diverse forums, and that the addition of linguistic categories from LIWC greatly improves classification performance. We also note that the LSTM model is more sensitive to context changes than the SVM model, and plan to explore the differences between the models in greater detail in future work.

Our future work also includes expanding our dataset to capture more instances of what may characterize RQs across these domains to improve performance, and also to analyze other interesting domains, such as Reddit. We believe that it will be possible to improve our results by using more robust models, and also by developing features to represent the *sequential* properties of RQs by further utilizing the larger context of the surrounding dialog in our analysis.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*.

Robert Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Language Resources and Evaluation Conference, LREC2016*.

Silvio Amir, Byron Wallace, Hao Lyu, Paula Carvalho, and Mario Silva. 2016. Modelling Context with User Embeddings for Sarcasm Detection in Social Media *The SIGNLL Conference on Computational Natural Language Learning (CoNLL2016)*.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*.

Francois Chollet. 2015. Keras. https://github.com/fchollet/keras

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.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Language Resources and Evaluation Conference, LREC2012*.

Jane Frank. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics* 14(5):723–738.

Aniruddha Ghosh and Tony Veale. 2016. Fracking Sarcasm using Neural Network In *Proc. of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. WASSA 2016.

Raymond Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol* 15(1):5–27.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations. *ACL-IJCNLP* 2015:146–153.

Roberto González-Ibáñez, 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*. Citeseer, volume 2, pages 581–586.

Chung-hye Han. 1997. Deriving the interpretation of rhetorical questions. In *The Proc. of the Sixteenth West Coast Conference on Formal Linguistics, WCCFL16*.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua* 112(3):201–229.

Cornelia Ilie. 1994. *What else can I tell you?: a pragmatic study of English rhetorical questions as discursive and argumentative acts*. Acta Universitatis Stockholmiensis: Stockholm studies in English. Almqvist & Wiksell International. https://books.google.com/books?id=T2wiAQAAIAAJ.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. volume 2, pages 757–762.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Swbd-damsl labeling project coder's manual. Technical report, University of Colorado. Available as http://stripe.colorado.edu/ jurafsky/manual.august1.html.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. WASSA 2013.

Michael McCarthy and Ronald Carter. 2004. 'There's millions of them': hyperbole in everyday conversation *Journal of Pragmatics* 36:149–184.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Shereen Oraby, Vrindavan Harrison, Ernesto Hernandez, Lena Reed, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proc. of the SIGDIAL 2015 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. And thats a fact: Distinguishing factual and emotional argumentation in online dialogue. *NAACL HLT 2015* page 116.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincet Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

James Pennebaker, Martha Francis, and Rojer Booth. 2001. *LIWC: Linguistic Inquiry and Word Count*.

Richard Petty, John Cacioppo, and Martin Heesacker. 1981. Effects of rhetorical questions on persuasion: A cognitive response analysis. *Journal of personality and social psychology* 40(3):432.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *26th International Conference on Computational Linguistics (COLING2016)*

Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2015. Learning to recognize affective polarity in similes. In *Conferenece on Empirical Methods in NLP, EMNLP-2015*.

Ashequl Qadir, Ellen Riloff, and Marilyn A Walker. 2016. Automatically inferring implicit properties in similes. In *Proceedings of NAACL-HLT*. pages 1223–1232.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering* .

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 *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological Science* 5(3):159–163.

Hannah Rohde. 2006. Rhetorical questions as redundant interrogatives. *Department of Linguistics, UCSD* .

Jerrold M Sadock. 1971. Queclaratives. In *Seventh Regional Meeting of the Chicago Linguistic Society*. volume 7, pages 223–232.

Deborah Schaffer. 2005. Can rhetorical questions function as retorts? : Is the pope catholic? *Journal of Pragmatics* 37:433–600.

Reid Swanson, Stephanie Lukin, Luke Eisenberg, Thomas Chase Corcoran, and Marilyn A Walker. 2014. Getting reliable annotations for sarcasm in online dialogues. In *Language Resources and Evaluation Conference, LREC 2014*.

Oren Tsur, Dmitry Davidov, and Ari 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.

Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. In *Proceedings of the Cognitive Science Society*. volume 29.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet Sarcasm Detection Using Deep Neural Network. In *26th International Conference on Computational Linguistics (COLING2016)*