Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus

Courtney Napoles,[1] **Joel Tetreault**,[2] **Enrica Rosato**,[3] **Brian Provenzale**,[4] and **Aasish Pappu**[3]
[1] Johns Hopkins University, napoles@cs.jhu.edu
[2] Grammarly, joel.tetreault@grammarly.com
[3] Yahoo, {aasishkp|enricar}@yahoo-inc.com
[4] Accellion, bprovenzale@gmail.com

Abstract

This work presents a dataset and annotation scheme for the new task of identifying "good" conversations that occur online, which we call ERICs: Engaging, Respectful, and/or Informative Conversations. We develop a taxonomy to reflect features of entire threads and individual comments which we believe contribute to identifying ERICs; code a novel dataset of Yahoo News comment threads ($2.4k$ threads and $10k$ comments) and $1k$ threads from the Internet Argument Corpus; and analyze the features characteristic of ERICs. This is one of the largest annotated corpora of online human dialogues, with the most detailed set of annotations. It will be valuable for identifying ERICs and other aspects of argumentation, dialogue, and discourse.

1 Introduction

Automatically curating online comments has been a large focus in recent NLP and social media work, as popular news outlets can receive millions of comments on their articles each month (Warzel, 2012). Comment threads often range from vacuous to hateful, but good discussions *do* occur online, with people expressing different viewpoints and attempting to inform, convince, or better understand the other side, but they can get lost among the multitude of unconstructive comments. We hypothesize that identifying and promoting these types of conversations (ERICs) will cultivate a more civil and constructive atmosphere in online communities and potentially encourage participation from more users.

ERICs are characterized by:

- A respectful exchange of ideas, opinions, and/or information in response to a given topic(s).
- Opinions expressed as an attempt to elicit a dialogue or persuade.
- Comments that seek to contribute some new information or perspective on the relevant topic.

ERICs have no single identifying attribute: for instance, an exchange where communicants are in total agreement throughout can be an ERIC, as can an exchange with heated disagreement. Figures 1 and 2 contain two threads that are characterized by continual disagreement, but one is an ERIC and the other is not. We have developed a new coding scheme to label ERICs and identify six dimensions of comments and three dimensions of threads that are frequently seen in the comments section. Many of these labels are for characteristics of online conversations not captured by traditional argumentation or dialogue features. Some of the labels we collect have been annotated in previous work (§2), but this is the first time they are aggregated in a single corpus at the dialogue level.

In this paper, we present the Yahoo News Annotated Comments Corpus (YNACC), which contains $2.4k$ threads and $10k$ comments from the comments sections of Yahoo News articles. We additionally collect annotations on $1k$ threads from the Internet Argument Corpus (Abbott et al., 2016), representing another domain of online debates. We contrast annotations of Yahoo and IAC threads, explore ways in which threads perceived to be ERICs differ in this two venues, and identify some unanticipated characteristics of ERICs.

This is the first exploration of how characteristics of individual comments contribute to the dialogue-level classification of an exchange. YNACC will facilitate research to understand ERICS and other aspects of dialogue. The corpus and annotations will be available at https://github.com/cnap/ynacc.

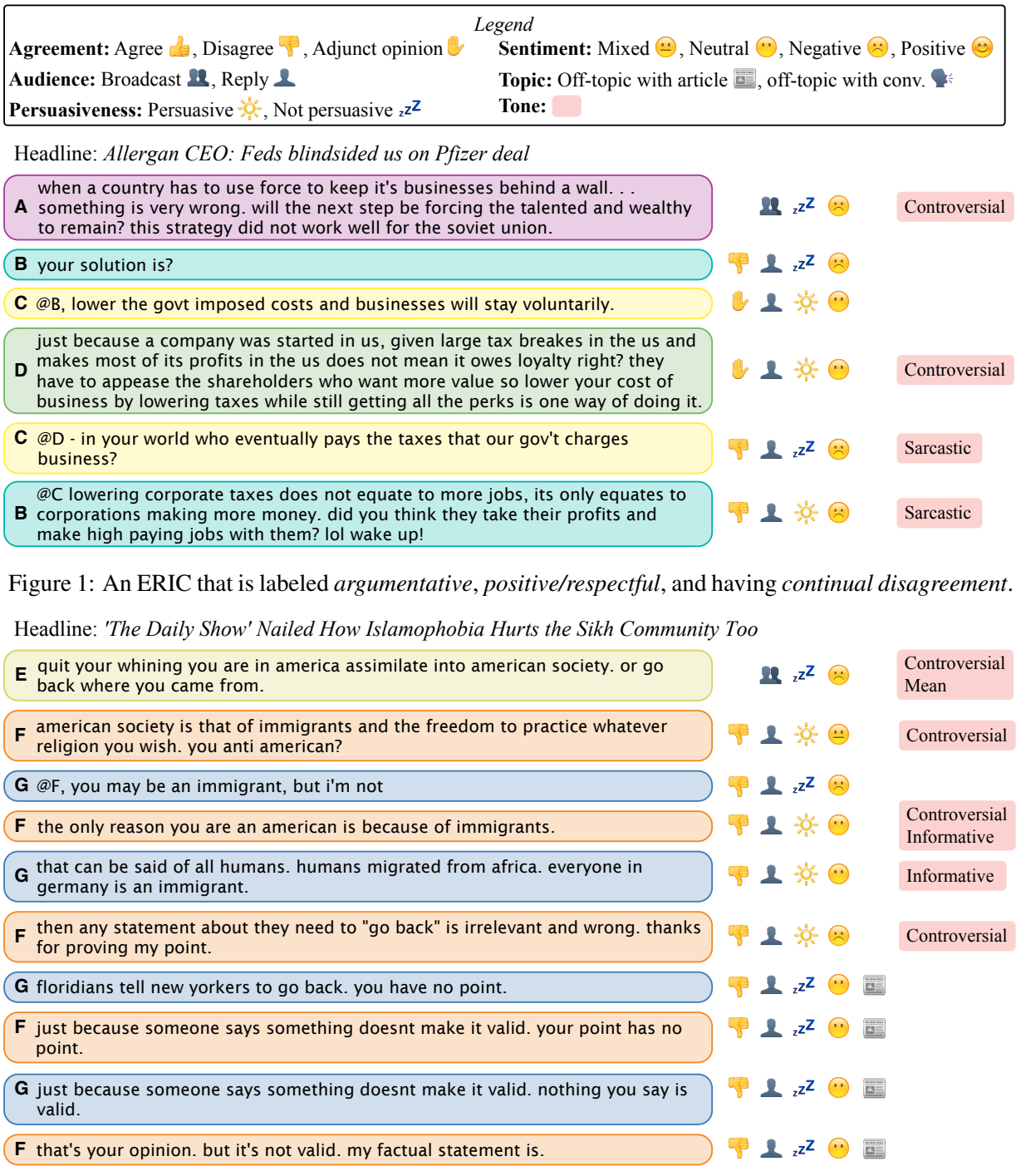

Figure 1: An ERIC that is labeled *argumentative*, *positive/respectful*, and having *continual disagreement*.

Figure 2: A non-ERIC that is labeled *argumentative* and *off-topic* with *continual disagreement*.

2 Related work

Recent work has focused on the analysis of user-generated text in various online venues, including labeling certain qualities of individual comments, comment pairs, or the roles of individual commenters. The largest and most extensively annotated corpus predating this work is the Internet Argument Corpus (IAC), which contains approximately $480k$ comments in $16.5k$ threads from online forums in which users debate contentious issues. The IAC has been coded for for topic ($3k$ threads), stance ($2k$ authors), and agreement, sarcasm, and hostility ($10k$ comment pairs) (Abbott et al., 2016; Walker et al., 2012). Comments from online news articles are annotated in the SENSEI corpus, which contains human-authored summaries of $1.8k$ comments posted on *Guardian* articles (Barker et al., 2016). Participants described

each comment with short, free-form text labels and then wrote a 150–250-word comment summary with these labels. Barker et al. (2016) recognized that comments have diverse qualities, many of which are coded in this work (§3), but did not explicitly collect labels of them.

Previous works present a survey of how editors and readers perceive the quality of comments posted in online news publications (Diakopoulos and Naaman, 2011) and review the criteria professional editors use to curate comments (Diakopoulos, 2015). The latter identifies 15 criteria for curating user-generated responses, from online and radio comments to letters to the editor. Our annotation scheme overlaps with those criteria but also diverges as we wish for the labels to reflect the nature of all comments posted on online articles instead of just the qualities sought in editorially curated comments. ERICs can take many forms and may not reflect the formal tone or intent that editors in traditional news outlets seek.

Our coding scheme intersects with attributes examined in several different areas of research. Some of the most recent and relevant discourse corpora from online sources related to this work include the following: Concepts related to *persuasiveness* have been studied, including annotations for "convincing-ness" in debate forums (Habernal and Gurevych, 2016), influencers in discussions from blogs and Wikipedia (Biran et al., 2012), and user relations as a proxy of *persuasion* in reddit (Tan et al., 2016; Wei et al., 2016). *Politeness* was labeled and identified in Stack Exchange and Wikipedia discussions (Danescu-Niculescu-Mizil et al., 2013). Some previous work focused on detecting agreement has considered blog and Wikipedia discussions (Andreas et al., 2012) and debate forums (Skeppstedt et al., 2016). *Sarcasm* has been identified in a corpus of microblogs identified with the hashtag #sarcasm on Twitter (González-Ibánez et al., 2011; Davidov et al., 2010) and in online forums (Oraby et al., 2016). *Sentiment* has been studied widely, often in the context of reviews (Pang and Lee, 2005), and in the context of user-generated exchanges, positive and negative attitudes have been identified in Usenet discussions (Hassan et al., 2010). Other qualities of user-generated text that are not covered in this work but have been investigated before include metaphor (Jang et al., 2014) and tolerance (Mukherjee et al., 2013) in online discussion threads, "dogmatism" of reddit users (Fast and Horvitz, 2016), and argumentation units in discussions related to technology (Ghosh et al., 2014).

3 Annotation scheme

This section outlines our coding scheme for identifying ERICs, with labels for comment threads and each comment contained therein.

Starting with the annotation categories from the IAC and the curation criteria of Diakopoulos (2015), we have adapted these schemes and identified new characteristics that have broad coverage over 100 comment threads (§4) that we manually examined.

Annotations are made at the *thread-level* and the *comment-level*. Thread-level annotations capture the qualities of a thread on the whole, while comment-level annotations reflect the characteristics of each comment. The labels for each dimension are described below. Only one label per dimension is allowed unless otherwise specified.

3.1 Thread labels

Agreement The overall agreement present in a thread.

- *Agreement throughout*
- *Continual disagreement*
- *Agreement* $\rightarrow$ *disagreement*: Begins with agreement which turns into disagreement.
- *Disagreement* $\rightarrow$ *agreement*: Starts with disagreement that converges into agreement.

Constructiveness A binary label indicating when a conversation is an ERIC, or has a clear exchange of ideas, opinions, and/or information done so somewhat respectfully.[1]

- *Constructive*
- *Not constructive*

Type The overall type or tone of the conversation, describing the majority of comments. Two labels can be chosen if conversations exhibit more than one dominant feature.

- *Argumentative*: Contains a lot of "back and forth" between participants that does not necessarily reach a conclusion.
- *Flamewar*: Contains insults, users "yelling" at each other, and no information exchanged.

[1]Note that this definition of *constructive* differs from that of Niculae and Danescu-Niculescu-Mizil (2016), who use the term to denote discrete progress made towards identifying a point on a map. Our definition draws from the more traditional meaning when used in the context of conversations as "intended to be useful or helpful" (Macmillan, 2017).

- *Off-Topic/digression*: Comments are completely irrelevant to the article or each other, or the conversation starts on topic but veers off into another direction.
- *Personal stories*: Participants exchange personal anecdotes.
- *Positive/respectful*: Consists primarily of comments expressing opinions in a respectful, potentially empathetic manner.
- *Snarky/humorous*: Participants engage with each other using humor rather than argue or sympathize. May be on- or off-topic.

3.2 Comment labels

Agreement Agreement expressed with explicit phrasing (e.g., *I disagree...*) or implicitly, such as in Figure 2. Annotating the target of (dis)agreement is left to future work due to the number of other codes the annotators need to attend to. Multiple labels can be chosen per comment, since a comment can express agreement with one statement and disagreement with another.

- *Agreement with another commenter*
- *Disagreement with another commenter*
- *Adjunct opinion*: Contains a perspective that has not yet been articulated in the thread.

Audience The target audience of a comment.

- *Reply to specific commenter*: Can be explicit (i.e., @HANDLE) or implicit (not directly naming the commenter). The target of a reply is not coded.
- *Broadcast message*: Is not directed to a specific person(s).

Persuasiveness A binary label indicating whether a comment contains persuasive language or an intent to persuade.

- *Persuasive*
- *Not persuasive*

Sentiment The overall sentiment of a comment, considering how the user feels with respect to what information they are trying to convey.

- *Negative*
- *Neutral*
- *Positive*
- *Mixed*: Contains both positive and negative sentiments.

Tone These qualities describe the overall tone of a comment, and more than one can apply.

- *Controversial*: Puts forward a strong opinion that will most likely cause disagreement.
- *Funny*: Expresses or intends to express humor.
- *Informative*: Contributes new information to the discussion.
- *Mean*: The purpose of the comment is to be rude, mean, or hateful.
- *Sarcastic*: Uses sarcasm with either intent to humor (overlaps with *Funny*) or offend.
- *Sympathetic*: A warm, friendly comment that expresses positive emotion or sympathy.

Topic The topic addressed in a comment, and more than one label can be chosen. Comments are on-topic unless either *Off-topic* label is selected.

- *Off-topic with the article*
- *Off-topic with the conversation*: A digression from the conversation.
- *Personal story*: Describes the user's personal experience with the topic.

4 Corpus collection

With the taxonomy described above, we coded comments from two separate domains: online news articles and debate forums.

Threads from online news articles YNACC contains threads from the "comments section" of Yahoo News articles from April 2016.[2] Yahoo filters comments containing hate speech (Nobata et al., 2016) and abusive language using a combination of manual review and automatic algorithms, and these comments are not included in our corpus. From the remaining comments, we identified threads, which contain an initial comment and at least one comment posted in reply. Yahoo threads have a single-level of embedding, meaning that users can only post replies under a top-level comment. In total, we collected 521,608 comments in 137,620 threads on 4,714 articles on topics including finance, sports, entertainment, and lifestyle. We also collected the following metadata for each comment: unique user ID, time posted, headline, URL, category, and the number of *thumbs up* and *thumbs down* received. We included comments posted on a thread regardless of how much time had elapsed since the initial comment because the vast majority of comments were posted in close sequence: 48% in the first hour after an initial comment, 67% within the first three hours, and 92% within the first 24 hours.

We randomly selected 2,300 threads to annotate, oversampling longer threads since the aver-

[2]Excluding comments labeled *non-English* by LangID, a high-accuracy tool for identifying languages in multiple domains (Lui and Baldwin, 2012)

	IAC	Yahoo
# Threads	1,000	2,400
# Comments	16,555	9,160
Thread length	29 ± 55	4 ± 3
Comment length	568 ± 583	232 ± 538
Trained	0	1,400 threads 9,160 comments
Untrained	1,000 threads	1,300 threads

Table 1: Description of the threads and comments annotated in this work and and the number coded by trained and untrained annotators. Thread length is in comments, comment length in characters.

age Yahoo thread has only 3.8 comments. The distribution of thread lengths is 20% with 2–4 comments, 60% 5–8, and 20% 9–15. For a held-out test set, we collected an additional 100 threads from Yahoo articles posted in July 2016, with the same length distribution. Those threads are not included in the analysis performed herein.

Threads from web debate forums To test this annotation scheme on a different domain, we also code online debates from the IAC 2.0 (Abbott et al., 2016). IAC threads are categorically different from Yahoo ones in terms of their stated purpose (debate on a particular topic) and length. The mean IAC thread has 29 comments and each comment has 102 tokens, compared to Yahoo threads which have 4 comments with 51 tokens each. Because significant attention is demanded to code the numerous attributes, we only consider IAC threads with 15 comments or fewer for annotation, but do not limit the comment length. In total, we selected 1,000 IAC thread to annotate, specifically: 474 threads from 4forums that were coded in the IAC, all 23 threads from CreateDebate, and 503 randomly selected threads from ConvinceMe.

4.1 Annotation

The corpus was coded by two groups of annotators: professional trained editors and untrained crowdsourced workers. Three separate annotators coded each thread. The trained editors were paid contractors who received two 30–45-minute training sessions, editorial guidelines (2,000-word document), and two sample annotated threads. The training sessions were recorded and available to the annotators during annotation, as were the guidelines. They could communicate their questions to the trainers, who were two authors of this paper, and receive feedback during the training and annotation phases.

Because training is expensive and time consuming, we also collected annotations from untrained coders on Amazon Mechanical Turk (AMT). To simplify the task for AMT, we only solicited thread-level labels, paying $0.75 per thread. For quality assurance, only workers located in the United States or Canada with a minimum HIT acceptance rate of 95% could participate, and the annotations were spot-checked by the authors. Trained annotators coded 1,300 Yahoo threads and the 100-thread test set on the comment- and thread-levels; untrained annotators coded thread-level labels of 1,300 Yahoo threads (300 of which overlapped with the trained annotations) and 1,000 IAC threads (Table 1). In total, 26 trained and 495 untrained annotators worked on this task.

4.2 Confidence

To assess the difficulty of the task, we also collected a rating for each thread from the trained annotators describing how confident they were with their judgments of each thread and the comments it comprises. Ratings were made on a 5-level Likert scale, with 1 being not at all confident and 5 fully confident. The levels of confidence were high (3.9 ± 0.7), indicating that coders were able to distinguish the thread and comment codes with relative ease.

4.3 Agreement levels

We measure inter-annotator agreement with Krippendorff's alpha (Krippendorff, 2004) and find that, over all labels, there are substantial levels of agreement within groups of annotators: $\alpha = 0.79$ for trained annotators and $\alpha = 0.71$ and 72 for untrained annotators on the Yahoo and IAC threads, respectively. However, there is lower agreement on thread labels than comment labels (Table 2). The agreement of *thread type* is 25% higher for the Yahoo threads than the IAC (0.62–0.64 compared to 0.48). The less subjective comment labels (i.e., agreement, audience, and topic) have higher agreement than persuasiveness, sentiment, and tone. While some of the labels have only moderate agreement ($0.5 < \alpha < 0.6$), we find these results satisfactory as the agreement levels are higher than those reported for similarly subjective discourse annotation tasks (e.g., Walker et al. (2012)).

To evaluate the untrained annotators, we compare the thread-level annotations made on 300 Yahoo threads by both trained and untrained coders,

	Yahoo		IAC
Thread label	**Trained**	**Untrained**	**Untrained**
Agreement	0.52	0.50	0.53
Constructive	0.48	0.52	0.63
Type	0.62	0.64	0.48
Comment label			
Agreement	0.80	–	–
Audience	0.74	–	–
Persuasiveness	0.48	–	–
Sentiment	0.50	–	–
Tone	0.63	–	–
Topic	0.82	–	–

Table 2: Agreement levels found for each label category within trained and untrained groups of annotators, measured by Krippendorff's alpha.

Category	Label	Matches
Constructive class	–	0.61
Agreement	–	0.62
Thread type	*Overall*	0.81
	Argumentative	0.72
	Flamewar	0.80
	Off-topic	0.82
	Personal stories	0.94
	Respectful	0.81
	Snarky/humorous	0.85

Table 3: Percentage of threads (out of 300) for which the majority label of the trained annotators matched that of the untrained annotators.

by taking the majority label per item from each group of annotators and calculating the percent of exact matches (Table 3). When classifying the *thread type*, multiple labels are allowed for each thread, so we convert each option into a boolean and analyze them separately. Only 8% of the threads have no majority *constructive* label in the trained and/or untrained annotations, and 20% have no majority *agreement* label. Within both annotation groups, there are majority labels on all of the *thread type* labels. The category with the lowest agreement is *constructive class* with only 61% of the majority labels matching, followed closely by *agreement* (only 62% matching). A very high percent of the thread type labels (81%). The strong agreement levels between trained and untrained annotators suggest that crowdsourcing is reliable for coding thread-level characteristics.

5 Annotation analysis

To understand what makes a thread constructive, we explore the following research questions:

1. How does the overall thread categorization differ between ERICs and non-ERICs? (§5.1)
2. What types of comments make up ERICs compared to non-ERICs? (§5.2)
3. Are social signals related to whether a thread is an ERIC? (§5.3)

5.1 Thread-level annotations

Before examining what types of threads are ERICs, we first compare the threads coded by different sets of annotators (trained or untrained) and from different sources (IAC or Yahoo). We measure the significance of annotation group for each label with a test of equal proportions for binary categories (*constructiveness* and each *thread type*) and a chi-squared test of independence for the *agreement* label. Overall, annotations by the trained and untrained annotators on Yahoo threads are very similar, with significant differences only between some of the *thread type* labels (Figure 3). We posit that the discrepancies between the trained and untrained annotators is due to the former's training sessions and ability to communicate with the authors, which could have swayed annotators to make inferences into the coding scheme that were not overtly stated in the instructions.

The differences between Yahoo and IAC threads are more pronounced. The only label for which there is no significant difference is *personal stories* ($p = 0.41$, between the IAC and trained Yahoo labels). All other IAC labels are significantly different from both trained and untrained Yahoo labels ($p < 0.001$). ERICs are more prevalent in the IAC, with 70% of threads labeled *constructive*, compared to roughly half of Yahoo threads. On the whole, threads from the IAC are more concordant and positive than from Yahoo: they have more agreement and less disagreement, more than twice as many positive/respectful threads, and fewer than half the flamewars.

For Yahoo threads, there is no significant difference between trained and untrained coders for *constructiveness* ($p = 0.11$) and the *argumentative* thread type ($p = 0.07$; all other thread types are significant with $p < 10^{-5}$). There is no significant difference between the *agreement* labels, either ($p = 1.00$). Untrained coders are more likely than trained to classify threads using emotional labels like *snarky*, *flamewar*, and *positive/respectful*, while trained annotators more frequently recognize *off-topic* threads. These differences should be taken into consideration for evaluating the IAC codes, and for future efforts collecting subjective annotations through crowdsourcing.

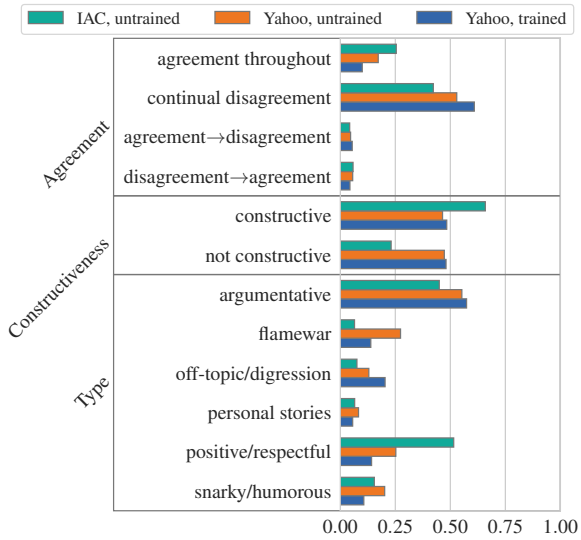

Figure 3: % threads assigned labels by annotator type (trained, untrained) and source (Yahoo, IAC).

We measure the strength of relationships between labels with the phi coefficient (Figure 4). There is a positive association between ERICs and all *agreement* labels in both Yahoo (trained) and IAC threads, which indicates that concord is not necessary for threads to be constructive. The example in Figure 1 is a *constructive* thread that is *argumentative* and contains *disagreement*. Thread types associated with non-ERICs are *flamewars*, *off-topic digressions*, and *snarky/humorous* exchanges, which is consistent across data sources. The labels from untrained annotators show a stronger correlation between *flamewars* and *not constructive* compared to the trained annotators, but the former also identified more *flamewars*. Some correlations are expected: across all annotating groups, there is a positive correlation between threads labeled with *agreement throughout* and *positive/respectful*, and *disagreement throughout* is correlated with *argumentative* (Figures 1 and 2) and, to a lesser degree, *flamewar*.

The greatest difference between the IAC and Yahoo are the *thread types* associated with ERICs. In the IAC, the *positive/respectful* label has a much stronger positive relationship with *constructive* than the trained Yahoo labels, but this could be due to the difference between trained and untrained coders. *Argumentative* has a positive correlation with *constructive* in the Yahoo threads, but a weak negative relationship is found in the IAC. In both domains, threads characterized as *off-topic*, *snarky*, or *flamewars* are more likely to be non-ERICs. Threads with some level of *agreement* characterized as *positive/respectful* are commonly ERICs. A two-tailed z-test shows a significant difference between the number of ERICs and non-ERICs in Yahoo articles in the Arts & Entertainment, Finance, and Lifestyle categories ($p < 0.005$; Figure 5).

5.2 Comment annotations

We next consider the codes assigned by trained annotators to Yahoo comments (Figure 6). The majority of comments are *not persuasive*, *reply to a previous comment*, express *disagreement*, or have *negative sentiment*. More than three times as many comments express disagreement than agreement, and comments are labeled *negative* seven times as frequently as *positive*. Approximately half of the comments express disagreement or a negative sentiment. Very few comments are *funny*, *positive*, *sympathetic*, or contain a *personal story* ($< 10\%$). Encouragingly, only 6% of comments are *off-topic with the conversation*, suggesting that participants are attuned to and respectful of the topic. Only 20% of comments are *informative*, indicating that participants infrequently introduce new information to complement the article or discussion.

The only strong correlations are between the binary labels, but the moderate correlations provide insight into the Yahoo threads (Figure 7). Some relationships accord with intuition. For instance, participants tend to go off-topic with the article when they are responding to others and not during broadcast messages; comments expressing disagreement with a commenter are frequently posted in a reply to a commenter; comments expressing agreement tend to be sympathetic and have positive sentiment; and mean comments correlate with negative sentiment. Commenters in this domain also express disagreement without particular nastiness, since there is no correlation between *disagreement* and *mean* or *sarcastic* comments. The *informative* label is moderately correlated with *persuasiveness*, suggesting that comments containing facts and new information are more convincing than those without.

The correlation between comment and thread labels is shown in Figure 7. Many of the relationships are unsurprising, like *off-topic* threads tend to have *off-topic* comments, *personal-story* threads have *personal-story* comments; thread agreement levels correlate with comment-level

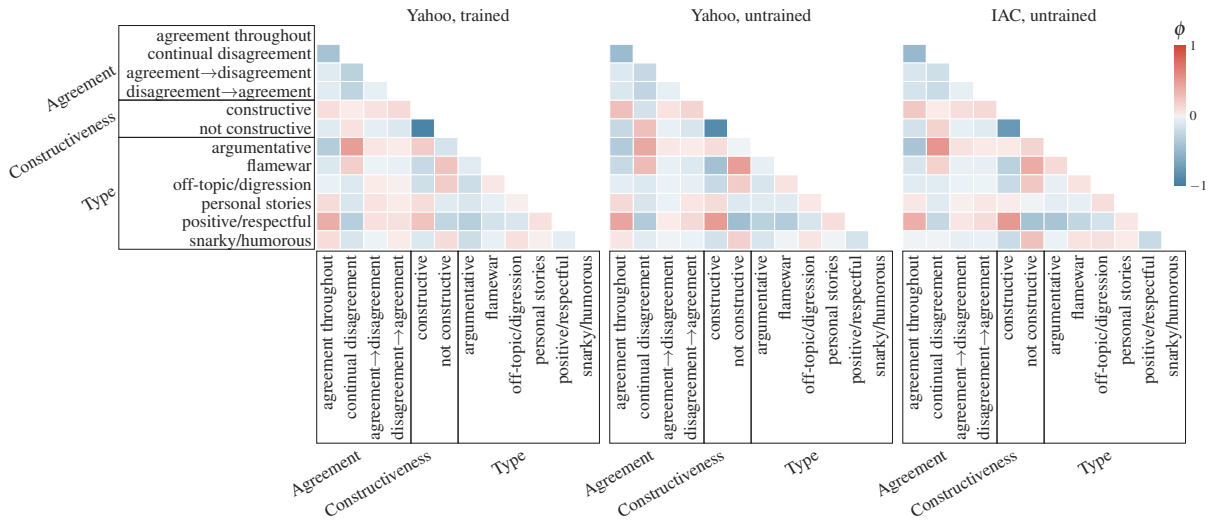

Figure 4: Correlation between thread labels, measured by the phi coefficient (ϕ).

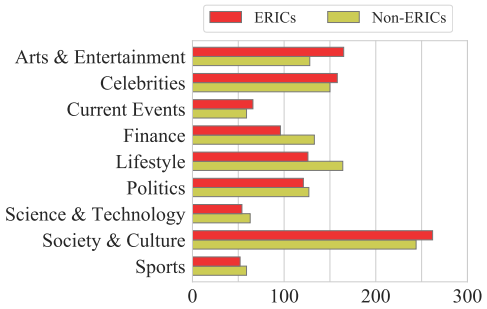

Figure 5: Number of threads by article category.

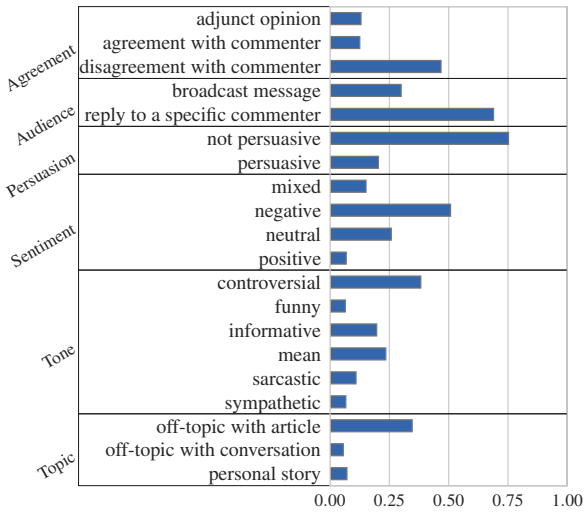

Figure 6: % Yahoo comments assigned each label.

agreements; and *flamewars* are correlated with *mean* comments.

In accord with our definition of ERICs, *constructiveness* is positively correlated with *informative* and *persuasive* comments and negatively correlated with *negative* and *mean* comments. From these correlations one can infer that *argumentative* threads are generally respectful because, while they are strongly correlated with comments that are *controversial* or express *disagreement* or a *mixed* sentiment, there is no correlation with *mean* and very little with *negative* sentiment. More surprising is the positive correlation between *controversial* comments and *constructive* threads. Controversial comments are more associated with ERICs, not non-ERICs, even though the *controversial* label also positively correlates with *flamewars*, which are negatively correlated with *constructiveness*. The examples in Figures 1–2 both have controversial comments expressing disagreement, but comments in the second half of the non-ERIC veer off-topic and are not persuasive, where the ERIC stays on-topic and persuasive.

5.3 The relationship with social signals

Previous work has taken social signals to be a proxy for thread quality, using some function of the total number of votes received by comments within a thread (e.g., Lee et al. (2014)). Because earlier research has indicated that user votes are not completely independent or objective (Sipos et al., 2014; Danescu-Niculescu-Mizil et al., 2009), we take the use of votes as a proxy for quality skeptically ad perform our own exploration of the relationship between social signals and the presence of ERICs. On Yahoo, users reacted to comments with a *thumbs up* or *thumbs down* and we collected the total number of such reactions for each comment in our corpus. First, we compare the total number of thumbs up (TU) and thumbs down (TD) received by comments in a

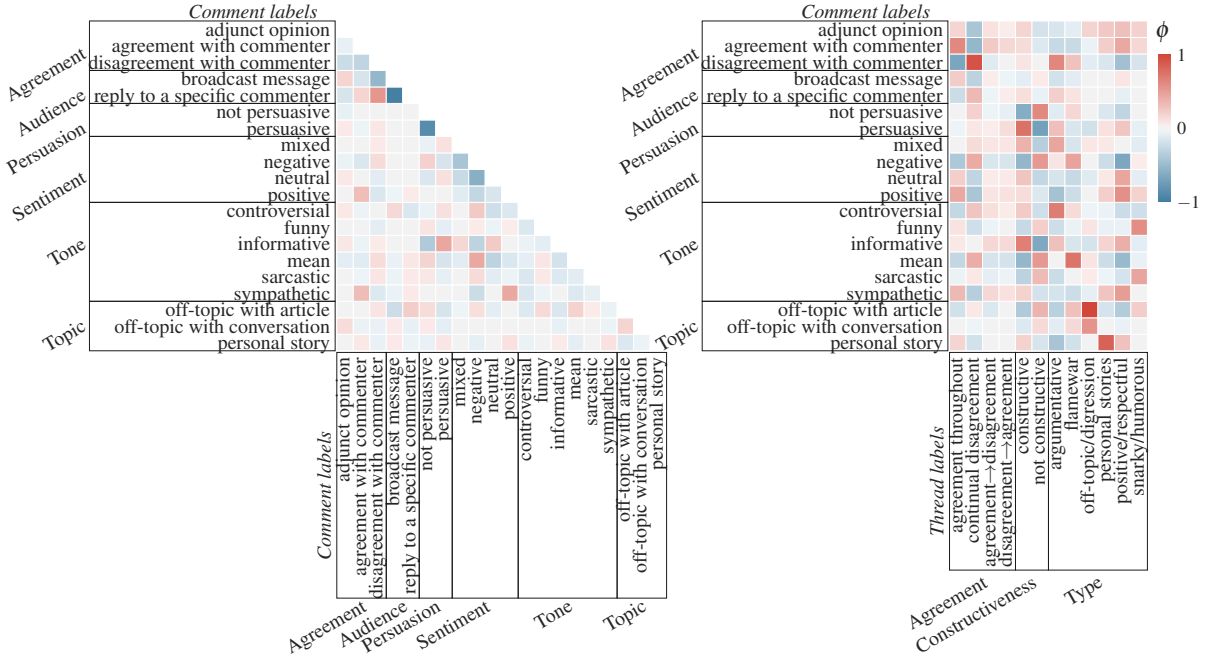

Figure 7: Correlation between comment labels (left) and comment labels and thread labels (right).

thread to the coded labels to determine whether there are any relationships between social signals and threads qualities. We calculate the relationship between labels in each category with TU and TD with Pearson's coefficient for the binary labels and a one-way ANOVA for the *agreement* category. The strongest correlation is between TD and untrained annotators' perception of *flamewars* ($r = 0.21$), and there is a very weak to no correlation (positive or negative) between the other labels and TU, TD, or TU−TD. There is moderate correlation between TU and TD ($r = 0.46$), suggesting that threads that elicit reactions tend to receive both thumbs up and down.

The correlation between TU and TD received by each comment is weaker ($r = 0.23$). Comparing the comment labels to the TU and TD received by each comment also show little correlation. Comments that reply to a specific commenter are negatively correlated with TU, TD, and TU−TD ($r = 0.30$, -0.25, and -0.22, respectively). The only other label with a non-negligible correlation is *disagreement with a commenter*, which negatively correlates with TU ($r = -0.21$). There is no correlation between social signal and the presence of ERICs or non-ERICs. These results support the findings of previous work and indicate that thumbs up or thumbs down alone (and, presumably, up/down votes) are inappropriate proxies for quality measurements of comments or threads in this domain.

6 Conclusion

We have developed a coding scheme for labeling "good" online conversations (ERICs) and created the Yahoo News Annotated Comments Corpus, a new corpus of $2.4k$ coded comment threads posted in response to Yahoo News articles. Additionally, we have annotated $1k$ debate threads from the IAC. These annotations reflect several different characteristics of comments and threads, and we have explored their relationships with each other. ERICs are characterized by argumentative, respectful exchanges containing persuasive, informative, and/or sympathetic comments. They tend to stay on topic with the original article and not to contain funny, mean, or sarcastic comments. We found differences between the distribution of annotations made by trained and untrained annotators, but high levels of agreement within each group, suggesting that crowdsourcing annotations for this task is reliable. YNACC will be a valuable resource for researchers in multiple areas of discourse analysis.

Acknowledgments

We are grateful to Danielle Lottridge, Smaranda Muresan, and Amanda Stent for their valuable input. We also wish to thank the anonymous reviewers for their feedback.

References

Rob 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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4445–4452, Paris, France, May. European Language Resources Association (ELRA).

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 818–822, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert Gaizauskas. 2016. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–52, Los Angeles, September. Association for Computational Linguistics.

Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. 2012. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45, Montréal, Canada, June. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 141–150, New York. ACM.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 133–142, New York. ACM.

Nicholas Diakopoulos. 2015. Picking the NYT picks: Editorial criteria and automation in the curation of online news comments. *ISOJ Journal*, 5(1):147–166.

Ethan Fast and Eric Horvitz. 2016. Identifying dogmatism in social media: Signals and models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 690–699, Austin, Texas, November. Association for Computational Linguistics.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June. Association for Computational Linguistics.

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.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, August. Association for Computational Linguistics.

Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. 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*, pages 1245–1255, Cambridge, MA, October. Association for Computational Linguistics.

Hyeju Jang, Mario Piergallini, Miaomiao Wen, and Carolyn Rose. 2014. Conversational metaphors in use: Exploring the contrast between technical and everyday notions of metaphor. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 1–10, Baltimore, MD, June. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage Publications, Thousand Oaks, CA, 2nd edition.

Jung-Tae Lee, Min-Chul Yang, and Hae-Chang Rim. 2014. Discovering high-quality threaded discussions in online forums. *Journal of Computer Science and Technology*, 29(3):519–531.

Macmillan Publishers Ltd. 2009. The online English dictionary: Definition of constructive. `http://www.macmillandictionary.com/dictionary/american/constructive`. Accessed January 20, 2017.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July. Association for Computational Linguistics.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Sharon Meraz. 2013. Public dialogue: Analysis of tolerance in online discussions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1680–1690, Sofia, Bulgaria, August. Association for Computational Linguistics.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California, June. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles, September. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. 2014. Was this review helpful to you?: It depends! Context and voting patterns in online content. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 337–348. ACM.

Maria Skeppstedt, Magnus Kerren, Carita Sahlgren, and Andreas Paradis. 2016. Unshared task: (Dis) agreement in online debates. In *3rd Workshop on Argument Mining (ArgMining'16), Berlin, Germany, August 7-12, 2016*, pages 154–159. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Charlie Warzel. 2012. Everything in moderation. *Adweek*, June 18. `http://www.adweek.com/digital/everything-moderation-141163/`. Accessed February 20, 2017.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany, August. Association for Computational Linguistics.