Using New York Times Picks to Identify Constructive Comments

Varada Kolhatkar
Discourse Processing Lab
Simon Fraser University
Burnaby, Canada
vkolhatk@sfu.ca

Maite Taboada
Discourse Processing Lab
Simon Fraser University
Burnaby, Canada
mtaboada@sfu.ca

Abstract

We examine the extent to which we are able to automatically identify constructive online comments. We build several classifiers using New York Times Picks as positive examples and non-constructive thread comments from the Yahoo News Annotated Comments Corpus as negative examples of constructive online comments. We evaluate these classifiers on a crowd-annotated corpus containing 1,121 comments. Our best classifier achieves a top F1 score of 0.84.

1 Introduction

Online commenting allows for direct communication among people and organizations from various socioeconomic classes on important issues. Popular news articles receive thousands of comments, but not all of them are constructive. Below we show examples of a constructive and a non-constructive comment on an article about Hillary Clinton's loss in the presidential election in 2016.[1]

(1) There is something inherently sexist about the assumption that women are incorruptible and naturally groomed to be better leaders than their male counterparts by virtue of being female. It is troubling to see intelligent sexist women relay the disturbingly substandard notion that despite Hillary Clinton's deeply flawed and frankly troubling history, this one here must be revered at all cost. Women are equal under the law, and as such should be held to the same legal, ethical, and job performance standards - regardless of their gender or power.

(2) If you think she lost because she was a women then you are really out to lunch. Gender has nothing to do with it.

[1] https://www.theglobeandmail.com/opinion/thank-you-hillary-women-now-know-retreat-is-not-an-option/article32803341/

The first one, which was labelled as constructive by our annotators (see Section 3), presents an argument (that women should be equal in all aspects), a challenge to an assumption (that women are incorruptible), and a protest against overlooking Ms. Clinton's flaws because of her gender. The second comment, labelled as non-constructive, exhibits a dismissive tone (*you are really out to lunch*), and provides no supporting evidence for the claim that gender was not a factor in the election.

There is growing interest in automatically organizing reader comments in a sensible way (Napoles et al., 2017; Llewellyn et al., 2014). One useful way to organize comments is based on their *constructiveness*, i.e., by identifying which comments provide insight and encourage a healthy discussion. For instance, The New York Times manually selects and highlights comments representing a range of diverse views, referred to as *NYT Picks*.

In this paper, we focus on this problem of identifying constructive comments. We define constructive comments as those that contribute to the dialogue, which provide insights relevant to the article, perhaps supported by evidence, and develop computational methods for identifying constructive comments.

The primary challenge in developing a computational system for constructiveness is the lack of annotated data. There is no systematically-annotated training data available for constructiveness of individual comments. So we explore the available resources: a set of NYT Picks as positive examples and non-constructive thread comments from the Yahoo News Annotated Comments Corpus (YNACC) (Napoles et al., 2017) as negative examples for constructiveness. We train support vector machine classifiers and bidirectional long short-term memory networks on this combination dataset and achieve a top F1 score of 0.84 on an

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 100–105
Copenhagen, Denmark, September 7, 2017. ©2017 Association for Computational Linguistics

unseen test dataset containing 1,121 constructive and non-constructive reader comments from the website of a different newspaper, The Globe and Mail.[2]

2 Related work

Napoles et al. (2017) define constructiveness of comment threads in terms of ERICs—Engaging, Respectful, and/or Informative Conversations. They train four machine learning models on 2.1k annotated Yahoo News threads and report an F1 score of 0.73 as their highest when identifying constructive news threads. We deal with a similar problem, but in our case we examine individual comments, rather than threads, as there is value in identifying constructive comments as they come in rather than waiting for a thread to degenerate (Wulczyn et al., 2016). Work closer to ours is that of Park et al. (2016), who explore New York Times comments extracted using the New York Times API to distinguish between NYT Picks and non-picks. They train an SVM classifier on a skewed dataset containing 94 NYT Picks and 5,174 non-picks and achieve a cross-validation precision of 0.13 and recall of 0.60. NYT Picks have also been used to study editorial criteria in comment selection. For instance, Diakopoulos (2015) analyzed 5,174 NYT Picks and found that they show high levels of argument quality, criticality, internal coherence, personal experience, readability, and thoughtfulness.

The data used by Napoles et al. (2017) does not contain constructiveness annotations for individual comments, but only for comment threads. The NYT Picks used by Diakopoulos (2015) and Park et al. (2016) are good representatives of constructive comments, but non-picks are not necessarily non-constructive, as only a few comments among thousands of comments are selected as NYT Picks. We create our training data by combining these two resources: NYT Picks for positive examples and non-constructive comment threads from the YNACC[3] for negative examples.

3 Datasets

Training and validation data We propose to use NYT Picks as representative training data for constructiveness. The New York Times, like many newspaper sites, provides a platform for readers to comment on stories posted on the site. The comments are manually moderated, by a team of only 13 moderators.[4] As a result, only about 10% of the stories published are open for commenting.[5] Comments are classified into three categories: all comments, readers' picks, and NYT Picks. NYT Picks are curated by the team of human moderators, and are chosen because they are interesting and helpful, but also based on the region or the reader.[6] Below we show an example of a NYT pick on an article about a young girl's suicide due to cyber-bullying.[7] The comment urges readers to take an action against cyberbulling, and does so by encouraging others to discuss the hurtful nature of attacks online.

(3) All of us — moms, dads, sisters, brothers, and friends need to talk about how words hurt. We need to take a stronger stance against damaging attacks — Just say no to texting or saying such hurtful comments, racial epithets, etc. We often lament how electronic communication enables uncivil speech, but we need to address the root of the problem here — why 12 year olds (indeed people of any age) are urging another person to kill herself.

Our positive training examples have 15,079 NYT Picks extracted using the NYT API.[8] Our negative training examples consist of 15,950 comments occurring in negative threads in the YNACC (Napoles et al., 2017), which contains thread-level constructiveness annotations for Yahoo News comment threads. Because we are interested in individual comments, rather than threads, we consider all comments from a non-constructive thread to be non-constructive. An example of a comment from a non-constructive thread is shown in (4).

(4) What makes you think that he's not sleeping with the robots already ;).

The training data is split into training set (90%) and validation set (10%).

[2] https://www.theglobeandmail.com
[3] https://webscope.sandbox.yahoo.com
[4] https://www.nytimes.com/times-insider/2014/04/17/a-comments-path-to-publication/?_r=0
[5] https://www.nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html
[6] https://www.nytimes.com/content/help/site/usercontent/usercontent.html/#usercontent-nytpicks
[7] http://www.nytimes.com/2013/09/14/us/suicide-of-girl-after-bullying-raises-worries-on-web-sites.html
[8] https://developer.nytimes.com/

Feature	Description
Length features (4)	Number of tokens in the comment, number of sentences, average word length, average number of words per sentence
Argumentation features (5)	Presence of discourse connectives (*therefore, due to*) Reasoning verbs (*cause, lead*), modals (*may, should*) Abstract nouns (*problem, issue, decision, reason*) Stance adverbials (*undoubtedly, paradoxically)*
Named-entity features (1)	Number of named entities in the comment
Text quality features (2)	Readability score & personal experience description score

Table 1: Constructiveness features.

Test data Our test data consists of 1,121 comments downloaded from the site of The Globe and Mail, a Canadian daily. We conducted an annotation experiment using CrowdFlower,[9] asking annotators to read the article each comment refers to (a total of 10 articles), and to label the comment as constructive or not. For quality control, 100 units were marked as gold: annotators were allowed to continue with the annotation task only when their answers agreed with our answers to the gold questions. As we were interested in the verdict of native speakers of English, we limited the allowed demographic region to English-speaking countries. We asked for three judgments per instance and paid 5 cents per annotation unit. Percentage agreement for the constructiveness question on a random sample of 100 annotations was 87.88%, suggesting that constructiveness can be reliably annotated. Other measures of agreement, such as kappa, are not easily computed with CrowdFlower data, because many different annotators are involved. Constructiveness seemed to be equally distributed in our dataset: Out of the 1,121 comments, 603 comments (53.79%) were classified as constructive, 517 (46.12%) as non-constructive, and the annotators were not sure in only one case.[10] We have made the corpus and annotation guidelines publicly available.[11]

4 Experiments

We present results of three sets of experiments: 1) identifying constructive comments using support vector machine classifiers (SVMs) and constructiveness features, 2) predicting constructive comments using bi-directional long-short term memory neural networks (biLSTMs) and word embeddings, and 3) examining the effectiveness of using NYT picks as representative positive examples for constructiveness.

Measure	Training		Testing	
	C	NC	C	NC
Mean	132.06	46.53	100.19	24.06
SD	71.36	87.52	81.34	19.08

Table 2: The mean length in words and standard deviation (SD) for constructive and non-constructive comments. C = Constructive and NC = Non-constructive.

4.1 SVMs with constructiveness features

We train several SVM classifiers with a number of constructiveness features, shown in Table 1.

Word features We wanted to examine whether certain words or phrases are more common in constructive or non-constructive comments. For that we extracted features representing 1- to 4-gram counts and TF-IDF features.

Length features Constructive comments tend to contain long sentences and long content words. We include four length features, as shown in Table 1. Note that this feature class can also serve as a baseline—if the length alone is sufficient to identify constructiveness, we may not need to explore more sophisticated features for constructiveness. Table 2 shows the mean length in words and standard deviation for constructive and non-constructive comments in our training and test data. In general, constructive texts tend to be longer and in all cases there is great variation in length.

[9] https://www.crowdflower.com/

[10] In our experiments we consider this comment as a non-constructive comment.

[11] https://github.com/sfu-discourse-lab/Constructiveness_Toxicity_Corpus

Argumentation features We postulate a positive correlation between features of argumentative text and news comments. An argumentative text is one that contains argumentation, i.e., a claim supported by evidence, and presented as a coherent whole. The extensive literature on argumentation has identified linguistics aspects that pinpoint to argumentative texts (Biber, 1988; van Eemeren et al., 2007; Moens et al., 2007; Tseronis, 2011; Habernal and Gurevych, 2017). Based on this research, we include argumentation lexical cues, such as discourse connectives and stance adverbials, in our set of features.

Named-entity features Our hypothesis is that comments providing evidence and personal experiences (i.e., constructive comments) tend to contain named entities (e.g., *Hillary Clinton, the Government,* names of public institutions).

Text-quality features We include two features from Park et al. (2016), readability score and personal experience score. Park et al. (2016) also propose a method to identify high quality comments, in their case modelling on NYT Picks and non-picks. Some of their criteria are external to the comment (relevance to the article, whether it was recommended by other readers), but, since we want to rely exclusively on the comment content, we chose the two criteria that do so, both calculated using their tool.

We trained linear SVM classifiers with several feature combinations from the above set of features using sklearn.[12] These models predict constructive comments in our test data. Some of the best validation and prediction results of these classifiers are shown in Table 3.

	Model	Validation			Test		
		P	R	F1	P	R	F1
	Random	.51	.50	.50	.49	.50	.49
SVM Features	wf	.84	.83	.83	.81	.80	.80
	lf	.80	.80	.80	.79	.79	.79
	af	.75	.74	.75	.73	.73	.73
	tqf	.81	.81	.81	.83	.77	.76
	nef	.74	.73	.74	.72	.69	.68
	af+tqf+nef	.80	.78	.79	.84	.84	**.84**
	biLSTM	.86	.86	.86	.82	.81	.81

Table 3: Constructiveness prediction results. P=average precision (for constructive and non-constructive classes), R=average recall, F1=average F1 score, wf=word (features), lf=length, af=argumentative, tqf=text quality, nef=named entity.

4.2 biLSTMs with word embeddings

We wanted to examine to what extent a neural network model is able to learn relevant patterns of constructiveness from NYT Picks. We trained bidirectional long short-term memory networks (biLSTMs) with word embeddings on our training data. We initialized the embedding layer weights with GloVe vectors (Pennington et al., 2014). The biLSTM models are usually used for sequential predictions. Although our task is *not* a sequential prediction task, the primary reason for using biLSTMs is that these models can utilize the expanded paragraph-level context and learn paragraph representations directly. They have recently been used in diverse text classification tasks, such as stance detection (Augenstein et al., 2016), sentiment analysis (Teng et al., 2016), and medical event detection (Jagannatha and Yu, 2016).

We use bidirectional LSTMs as implemented in TensorFlow.[13] We trained with the ADAM stochastic gradient descent for 10 epochs. The important parameter settings are: batch size=512, embedding size=200, drop out=0.5, and learning rate=0.001. Results for the biLSTM classifier are also shown in Table 3. Note that the point of these results is to demonstrate the feasibility of automatically identifying constructive comments and the parameter setting may not be the optimal one.

4.3 Effectiveness of NYT Picks

To examine the effectiveness of using NYT Picks as representative positive training examples for constructiveness, we carried out experiments with training data containing a homogeneous sample from YNACC, in particular, by considering comments from constructive YNACC threads as constructive examples and comments from non-constructive threads as negative examples. When trained on this homogeneous YNACC training data, we observed P, R, and F1 of 0.72, 0.71, and 0.71, respectively. These numbers are markedly lower compared to the numbers we obtained when we used NYT Picks for training (F1 = 0.81), suggesting that using NYT Picks as positive examples for constructiveness does help. NYT Picks

[12] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

[13] https://www.tensorflow.org/

are chosen by human experts and are better representatives of constructiveness. Although the performance numbers with homogeneous YNACC look similar to the numbers reported in Napoles et al. (2017), recall that Napoles et al. (2017) focus on a different problem of identifying constructive conversation threads. A constructive thread may have a non-constructive comment and vice-versa. Moreover, they report cross-validation results, whereas we are reporting results on our test data containing reader comments from a different news paper.

5 Discussion and conclusion

We have explored several approaches to the problem of detecting constructiveness in online comments, focusing specifically on news comments. Constructiveness is a desirable feature in online discussion, and a constructiveness classifier can be useful for moderation tasks, typically performed by humans. Our methods achieve a top F1 score of 0.84, which is probably sufficient to assist news comments moderators.

We used two sets of available data as positive and negative examples for the classifiers: New York Times Picks as positive examples of constructiveness, and comments belonging to non-constructive threads from the Yahoo News Annotated Comments Corpus. Our test data consisted of 1,121 examples annotated for constructiveness through CrowdFlower.

Our methods can be grouped under two main categories: SVMs with various features and bidirectional LSTMs. For SVMs, we considered five classes of features: word, length, argumentation, named entity, and text quality features. Our best F1 score is 0.84 on the test set with the combination of argumentation, text quality, and named entity features. The length features alone give a high F1 score of 0.79. But when we combine them with other features the performance does not increase. On the other hand, argumentation, text quality, and named entity features seem to be complementary and give the best results when combined together.

Our biLSMT model requires only a vector representation of the text. We use an embedding layer initialized with GloVe vectors, and achieved an F1 score of 0.81 with this model. Note the similar performance of SVMs with word features and biLSTMs. We do not conclude from these experiments that either method is superior, since these are preliminary results and many other parameter combinations are possible. The point of these results is just to demonstrate the feasibility of automating the task of identifying constructiveness in news comments. A more rigourous investigation needs to be carried out in order to compare and understand the differences between SVMs and biLSTMs for this problem.

We achieved superior results when we used NYT Picks as positive training examples for constructiveness, suggesting that human-selected NYT Picks are better representatives of constructiveness.

A number of research avenues are planned for this project. First, we are interested in exploring other relevant features for constructiveness, such as the use of emoticons and fine-grained named-entity features (e.g., occurrences of a journalist's name). Second, we are interested in exploring the relation between constructiveness and toxicity. Finally, we are working towards making our computational system for identifying constructive comments robust and easily accessible.

Acknowledgments

We are grateful to Nicholas Diakopoulos for promptly providing us 5,232 NYT Picks, which are part of our training data. Thank you to the reviewers for constructive feedback.

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, TX.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Nicholas Diakopoulos. 2015. Picking the NYT Picks: Editorial criteria and automation in the curation of online news comments. *ISOJ Journal*, 6(1):147–166.

Frans H. van Eemeren, Peter Houtlosser, and A. Francisca Snoeck Henkemans. 2007. *Argumentative Indicators in Discourse: A pragma-dialectical study*. Springer, Berlin.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Abhyuday N. Jagannatha and Hong Yu. 2016. Bidirectional RNN for medical event detection in electronic health records. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, San Diego, CA.

Clare Llewellyn, Claire Grover, and Jon Oberlander. 2014. Summarizing Newspaper Comments. In *Proceedings of ICWSM*, Ann Arbor, MI.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230, Stanford, California. ACM.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop, EACL*, pages 13–23, Valencia.

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1114–1125. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar.

Zhiyang Teng, Duy Tin Vo, and Yue Zhang. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1629–1638, Austin, Texas.

Assimakis Tseronis. 2011. From connectives to argumentative markers: A quest for markers of argumentative moves and of related aspects of argumentative discourse. *Argumentation*, 25(4):427–447.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale. *arXiv:1702.08138v1*.