Constructive Language in News 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 discuss the characteristics of constructive news comments, and present methods to identify them. First, we define the notion of *constructiveness*. Second, we annotate a corpus for constructiveness. Third, we explore whether available argumentation corpora can be useful to identify constructiveness in news comments. Our model trained on argumentation corpora achieves a top accuracy of 72.59% (baseline=49.44%) on our crowd-annotated test data. Finally, we examine the relation between constructiveness and toxicity. In our crowd-annotated data, 21.42% of the non-constructive comments and 17.89% of the constructive comments are toxic, suggesting that non-constructive comments are not much more toxic than constructive comments.

1 Introduction

The goal of online news comments is to provide constructive, intelligent and informed remarks that are relevant to the article, often in the form of an exchange with other readers. Many comments, however, do not contribute to achieving this goal. Online comments have a broad range: they can be vacuous, dismissive, abusive, hateful, but also constructive. Below we show two comments on an article about Hillary Clinton's loss in the presidential election in 2016.[1]

(1) I have 3 daughters, and I told them that Mrs. Clinton lost because she did not have a platform. The only message that I got from her was that Mr. Trump is not fit to be in office and that she wanted to be the first female President. I honestly believe that she lost because she offered no hope, or direction, to the average American. Mr. Trump, with all his shortcomings, at least offered change and some hope.

(2) This article was a big disappointment. Thank you Ms Henein. Now women know that wasting their time reading your emotion-based opinion is not an option.

Both comments disagree with the author, but one does it constructively and the other dismissively. Comment (1) treats the article as a genuine starting point for discussion and presents disagreement without denigrating, with reasons for the disagreement. On the other hand, comment (2) is dismissive and probably sarcastic.

Our goal is to understand constructiveness in news comments, which may help in filtering and organizing many kinds of online comments. News comments may be filtered according to different criteria, for example, based on their toxicity and/or constructiveness. Toxic comments may be filtered negatively, i.e., they can be blocked, deleted, or demoted. Constructive comments may be filtered positively, i.e., they can be promoted, as it is done manually for the New York Times Picks (Diakopoulos, 2015). A number of approaches have been proposed for toxicity (e.g., Kwok and Wang, 2013; Waseem and Hovy, 2016; Wulczyn et al., 2016; Nobata et al., 2016; Davidson et al., 2017). A recent example is the effort by Google to identify abusive or toxic comments through the Perspective API.[2] There is, however, not as much research on the constructiveness of individual comments. Niculae and Danescu-Niculescu-Mizil (2016) and Napoles et al. (2017) study constructiveness at the comment thread-level, but not at the comment level.

In this paper, we focus on the constructiveness of individual news comments. First, we define the notion of *constructiveness*. Second, we de-

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

[2] https://www.perspectiveapi.com/

scribe our annotated corpus of online comments labelled for constructiveness. Third, we explore deep learning approaches for identifying constructive comments. Fourth, we discuss the association between constructiveness and a number of argumentation features. Finally, we examine the relationship between toxicity and constructiveness.

2 Constructiveness: Definition and corpus

We are interested in comments that contribute to the conversation, which construct, build and promote a dialogue. Napoles et al. (2017) define constructive conversations in terms of ERICs—Engaging, Respectful, and/or Informative Conversations. Rather than relying on our intuitions, we posted a survey asking what a constructive comment is. We opened a survey on SurveyMonkey[3], requesting 100 answers. A composite of the answers is: *Constructive comments intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response. They are typically targeted to specific points and supported by appropriate evidence*.

In order to study constructiveness in news comments, we crawled 1,121 comments from 10 articles of the Globe and Mail news website[4] covering a variety of subjects: technology, immigration, terrorism, politics, budget, social issues, religion, property, and refugees. We used CrowdFlower[5] as our crowdsourcing annotation platform and annotated the comments for constructiveness. We asked the annotators to first read the relevant article, and then to tell us whether the displayed comment was 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 was 87.88%, suggesting that constructiveness can be reliably annotated. Agreement numbers are provided by CrowdFlower, and are calculated on a random sample of 100 annotations. 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. We use this annotated corpus as the test data in our experiments. We have also made the corpus publicly available.[6]

3 Identifying constructive comments

We take the view that constructiveness is closely related to argumentation. Argumentative texts usually establish a position on a topic and provide reasoning for that particular position. Similarly, a constructive comment provides reasoning for the commenter's point of view. We exploit argumentation-related datasets to train a bidirectional Long Short-Term Memory (biLSTM) model (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) to identify constructive comments. We also explore the association between constructiveness and argumentation features.

3.1 Building a constructiveness classifier

Constructiveness is an interplay between different kinds of linguistic knowledge: lexical, syntactic, semantic and pragmatic knowledge. Lexical and syntactic features, such as use of hedges and modals, sentence structure, readability or text complexity; semantic features, such as the use of personal and emotion words or the sentiment score for the comment; and discourse features, such as cohesion, discourse relations, the comment's topic, or the topic distance from the article, have shown to help in identifying similar phenomena, such as quality of student essays or constructiveness of a comment thread (Pitler and Nenkova, 2008; Brand and Van Der Merwe, 2014; Diakopoulos, 2015; Momeni et al., 2015; Niculae and Danescu-Niculescu-Mizil, 2016). The primary challenge in developing a computational system for constructiveness is the lack of training data from which we can learn about these different aspects of constructiveness.

Training data Since there is no training data available for constructiveness at the comment

[3] https://www.surveymonkey.com/
[4] http://www.theglobeandmail.com/
[5] https://www.crowdflower.com/
[6] https://github.com/sfu-discourse-lab/Constructiveness_Toxicity_Corpus

level, we gathered annotated data from similar tasks. In particular, we exploit two annotated corpora. The first corpus is the Yahoo News Annotated Corpus (YNC)[7] (Napoles et al., 2017), which contains thread-level constructiveness annotations for Yahoo News comment threads. We are interested in comment-level annotations, and thus assume that a comment from a constructive thread is constructive and vice versa for non-constructive threads. We extracted 33,957 comments from constructive conversations and 26,821 comments from non-constructive conversations from this dataset. Other than constructiveness annotations, the YNC corpus also contains annotations for sub-dialogue type (argumentative, flame-war, off topic, personal stories, positive, respectful, snarky or humorous). We concatenate these annotations to the comments when training.

The second corpus is the Argument Extraction Corpus (AEC)[8] (Swanson et al., 2015). The corpus includes annotations for argument quality on sentences extracted from the topics of gun control, gay marriage, evolution, and death penalty. Our intuition is that sentences with high argument quality are constructive and low argument quality are non-constructive. We extract 2,613 examples with high argumentation quality and 2,761 examples with low argumentation quality. In total, we had 36,570 constructive and 29,582 non-constructive training examples.

Test data Our test data is our crowd-sourced constructiveness corpus containing 1,121 instances marked for constructiveness. As news comments are not always well written, we carried out some preprocessing of the data, such as word segmentation and spelling correction. For example, in *Climate change has always been a **hoax,as*** ..., our preprocessing will add a space between *hoax,* and *as*.

Model and results We carry out preliminary experiments to assess whether argumentative comment representations are useful to identify constructive comments. We train biLSTM models with the annotated argumentation corpora. These models are usually used for sequential predictions. The models have *memory* in the sense that the results from the previous predictions can inform future predictions. The model learns what kind of

[7] https://webscope.sandbox.yahoo.com
[8] https://nlds.soe.ucsc.edu/node/29

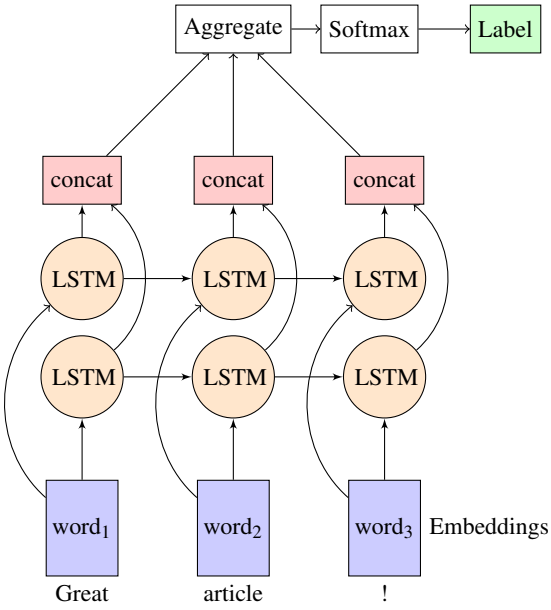

Figure 1: Bidirectional LSTM architecture. Label values: constructive, non-constructive.

memories are important in predicting the output.

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 contexts and learn paragraph representations directly. In our case, the memory is used not to remember the previous comments' predictions, but to remember the long-distance context within the same comment. Moreover, biLSTMs have been shown to learn better representations of sequences by processing them from left to right and from right to left. They have recently been used in diverse tasks, such as stance detection (Augenstein et al., 2016), sentiment analysis (Teng et al., 2016), and medical event detection (Jagannatha and Yu, 2016).

Figure 1 outlines the general architecture of our model. The words in each comment are mapped to their corresponding word representation using the embedding layer. The embedding layer contains the word vector mapping from words to dense n-dimensional vector representations. We initialize the embedding layer weights with GloVe vectors (Pennington et al., 2014). The word embeddings are fed into the LSTM layer. The LSTM layer has two LSTM chains: one propagating in the forward direction and one propagating in the backward direction. The representations are combined by taking linear combinations of the LSTM outputs. The output is then passed through the Softmax acti-

Training	Validation accuracy (%)	Test accuracy (%)
YNC + AEC	68.43	68.45
YNC	72.76	**72.59**
AEC	69.30	52.54

Table 1: Constructiveness prediction results using argumentation corpora. The test data was our annotated constructiveness data in all cases. Random baseline accuracy = 49.44%.

Feature	OR
Argumentative discourse relations	3.49
Stance adverbials	2.52
Reasoning verbs & modals	2.02
Root clauses	1.37
Conjunctions & connectives	0.82
Abstract nouns	0.51

Table 2: Association of constructiveness with linguistic features in terms of OR (odds ratio).

vation function, which produces a probability-like output for each label type, in our case for the labels constructive and non-constructive.

The network is trained with backpropagation. The embedding vectors are also updated based on the backpropagated errors. We use bidirectional LSTMs as implemented in TensorFlow[9]. 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.

We wanted to examine which argumentation dataset is more effective in identifying constructiveness. So we carried out experiments with different train and test combinations. In each experiment, 1% of the training data was used as the validation set.

Table 1 shows the average validation and test accuracies for three runs with the same parameter settings. Below we note a few observations. First, we achieved the best result when YNC was included in the training set. Second, AEC seems not to have much effect on the test accuracy but YNC does; when we do not have YNC in the training data, the results drop markedly. This might be because the size of the AEC corpus is relatively small and the model was not able to learn any relevant patterns from this data. Finally, the validation and test accuracy is more or less same for the first two rows, when YNC is included in the training data.

3.2 Association with argumentation features

In addition to the classifier described above, we also examine the association between constructiveness and a number of linguistic and discourse features typically found in argumentative texts, based on the extensive literature on argumentation (Biber, 1988; van Eemeren et al., 2007; Moens et al., 2007; Tseronis, 2011; Becker et al., 2016; Habernal and Gurevych, 2017; Azar, 1999; Peldszus and Stede, 2016). We calculate association in terms of odds ratio (Horwitz, 1979), which tells us the odds of a comment being constructive in the presence of a feature. Results are shown in Table 2. We observed a strong association between constructiveness and occurrence of argumentative discourse relations (Cause, Comparison, Condition, Contrast, Evaluation and Explanation).[10] The odds ratio for argumentative discourse relations is 3.49, which means that constructive texts are 3.49 times more likely to have this feature than non-constructive texts. Other features with strong association with constructiveness are stance adverbials (e.g., *undoubtedly, paradoxically, of course*), and reasoning verbs (e.g., *cause, lead*) and modals. Root clauses (clauses with a matrix verb and an embedded clause, such as *I think that* ...) show a medium association with constructiveness. On the other hand, abstract nouns (e.g., *issue, reason*) and, surprisingly, conjunctions and connectives are not associated with constructive texts. The latter is surprising because many discourse relations contain a connnective.

4 Toxicity in news comments

In the context of filtering news comments, we are also interested in the relationship between constructiveness and toxicity. We propose the label *toxicity* for a range of phenomena, including verbal abuse, offensive comments and hate speech. To better understand the nature of toxicity and its relationship with constructiveness, we extended our CrowdFlower annotation. For the 1,121 comments described in Section 2, we also asked anno-

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

[10] For this analysis we used the discourse relations given by the discourse parser described in Joty et al. (2015).

tators to identify toxicity. The question posed was: How toxic is the comment? We established four classes: *Very toxic*, *Toxic*, *Mildly toxic* and *Not toxic*. The definition for Very toxic included comments which use harsh, offensive or abusive language; comments which include personal attacks or insults; or which are derogatory or demeaning. Toxic comments were sarcastic, containing ridicule or aggressive disagreement. Mildly toxic comments were described as those which may be considered toxic only by some people, or which express anger and frustration.

The distribution of toxicity levels by constructiveness label is shown in Table 3. The percentage agreement provided by CrowdFlower for this task was 81.82%. The most important result of this annotation experiment is that there were no significant differences in toxicity levels between constructive and non-constructive comments, i.e., constructive comments were as likely to be toxic (in its three categories) as non-constructive comments. For instance, consider Example (3) below. It was labelled as constructive by two out of three annotators, and toxic by all three (two as Toxic, and one as Very toxic). It could be the case, in some situations, that a moderator may allow a somewhat toxic comment if it contributes to the conversation, i.e., if it is constructive.

(3) If it's wrong to vote AGAINST someone based on their gender,Then surely it is also wrong to vote FOR someone based on their gender.Yet there were many people advocating openly for people to to do just that.I wonder how many votes Clinton got just because she was a woman.

We conclude, then, that constructiveness and toxicity are orthogonal categories. The results also suggest that it is important to consider constructiveness of comments along with toxicity when filtering comments, as aggressive constructive debate might be a good feature of online discussion. Given these results, the classification of constructiveness and toxicity should probably be treated as separate problems.

5 Discussion and conclusion

We have proposed a definition of constructiveness that hinges on argumentative aspects of news comments. We have shown that well-known linguistic indicators of argumentation, such as adverbials and rhetorical relations show an association with constructive comments. Our definition of constructiveness is at the comment level, because it is important to identify comments as they come in, rather than waiting for a thread to degenerate (Wulczyn et al., 2016), and because many comments are top-level, i.e., not part of a thread.

	C ($n = 603$)	**Non-C** ($n = 518$)
Not toxic	82.09%	78.57%
Mildly toxic	16.08%	15.44%
Toxic	1.33%	5.21%
Very toxic	0.50%	0.77%
Total	100%	100%

Table 3: Percent distribution of constructive and toxic comments in CrowdFlower annotation. C = Constructive.

We assume that constructive comments contain good argumentation and explored argumentation datasets to train a bidirectional LSTM to identify constructive comments. The highest accuracy of our model was 72.59% (random baseline=49.44%).

Through an annotation experiment, we studied the relationship between constructiveness and toxicity, and found that constructive comments are just as likely to be toxic (or not toxic) as non-constructive comments. In terms of filtering, this poses an interesting question, since some of our toxic comments were also deemed to be constructive by the annotators.

As for future work, our long-term goal is to build a robust system for identifying constructive news comments. We also plan to investigate the relation between toxicity and constructiveness more deeply. We plan to train on more relevant and directly related training data, such as the New York Times Picks, and systematically explore different argumentation features for constructiveness (e.g., readability, cohesion, coherence).

Acknowledgments

We are grateful to Shafiq Joty for running his discourse parser on our data. We would like to thank the members of the Discourse Processing Lab: Hanhan Wu for data collection; Luca Cavasso, Jennifer Fest, Emilie Francis, Emma Mileva and Kazuki Yabe for testing CrowdFlower questions and providing feedback. Thank you also to the reviewers for very constructive suggestions.

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*. Austin, TX, pages 876–885.

Moshe Azar. 1999. Argumentative text as rhetorical structure: An application of Rhetorical Structure Theory. *Argumentation* 13(1):97–144.

Maria Becker, Alexis Palmer, and Anette Frank. 2016. Clause types and modality in argumentative microtexts. In *Proceedings of the Workshop on Foundations of the Language of Argumentation (in conjunction with COMMA 2016)*. Postdam, pages 1–9.

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

Dirk Brand and Brink Van Der Merwe. 2014. Comment classification for an online news domain. In *Proceedings of the First International Conference on the use of Mobile Informations and Communication Technology in Africa UMICTA*. Stellenbosch, South Africa, pages 50–56.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media*. Montréal.

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

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks, IJCNN*. volume 4, pages 2047–2052.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9(8):1735–1780.

Ralph I. Horwitz. 1979. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix: Cornfield J: J Nat Cancer Inst 11: 12691275, 1951. *Journal of Chronic Diseases* 32(1-2):i.

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*. San Diego, CA, pages 473–482.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* 41(3):385–435.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI'13, pages 1621–1622.

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*. ACM, Stanford, California, pages 225–230.

Elaheh Momeni, Claire Cardie, and Nicholas Diakopoulos. 2015. A survey on assessment and ranking methodologies for user-generated content on the web. *ACM Computing Surveys* 48(3):1–49.

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*. Valencia, pages 13–23.

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*. San Diego, pages 568–578.

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 World Wide Web Conference*. Montréal, pages 145–153.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the 3rd Workshop on Argument Mining, ACL*. Berlin, pages 103–112.

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*. Doha, Qatar, pages 1532–1543.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, HI, pages 186–195.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic, pages 217–226.

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*. Association for Computational Linguistics, Austin, Texas, pages 1629–1638.

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.

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

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, pages 88–93.

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