# Classifying Constructive Comments

Varada Kolhatkar[a,*], Nithum Thain[b,*], Jeffrey Sorensen[b], Lucas Dixon[b], Maite Taboada[c,**]

[a] *University of British Columbia*
[b] *Jigsaw*
[c] *Simon Fraser University*

## Abstract

We introduce the Constructive Comments Corpus (C3), comprised of 12,000 annotated news comments, intended to help build new tools for online communities to improve the quality of their discussions. We define constructive comments as high-quality comments that make a contribution to the conversation. We explain the crowd worker annotation scheme and define a taxonomy of sub-characteristics of constructiveness. The quality of the annotation scheme and the resulting dataset is evaluated using measurements of inter-annotator agreement, expert assessment of a sample, and by the constructiveness sub-characteristics, which we show provide a proxy for the general constructiveness concept. We provide models for constructiveness trained on C3 using both feature-based and a variety of deep learning approaches and demonstrate, through domain adaptation experiments, that these models capture general rather than topic- or domain-specific characteristics of constructiveness. We also examine the role that length plays in our models, as comment length could be easily gamed if models depend heavily upon this feature. By examining the errors made by each model and their distribution by length, we show that the best performing models are effective independently of comment length. The constructiveness corpus and our experiments pave the way for a moderation tool focused on promoting comments that make a meaningful contribution, rather than only filtering out undesirable content.

*Keywords:* Content moderation, online comments, toxicity, constructiveness, annotation, data creation, machine learning, deep learning

## 1. Introduction: Content moderation and constructiveness

One of the key challenges facing online communities, from social networks to the comment sections of news sites, is low-quality discussions. During the print era, with space carrying high cost, publications used strong editorial control in deciding which letters from readers should be published. As news moved online, comments sections were often open and unmoderated, and sometimes outsourced to other companies. One notable exception is *The New York Times* which has, since 2007, employed a staff of full-time moderators to review all comments submitted to their website (Etim, 2017). Exemplary comments representing a range of views are highlighted and tagged as *NYT Picks*. The importance of moderators was unanticipated by many publishers, with many removing comments sections because they could not adequately moderate the comments. With vast numbers of online comments, and growing challenges of how social networks manage toxic language,

Paper accepted for publication in *First Monday*. Draft version: August 4, 2020.

[*] Co-first author

[**] Corresponding author

the role of moderators is becoming much more demanding (Gillespie, 2018; Roberts, 2019). There is thus growing interest in developing automation to help filter and organize online comments for both moderators and readers (Park et al., 2016).

Comment moderation is often a task of filtering out, i.e., deleting toxic and abusive comments, the 'nasty' part of the Internet (Chen, 2017). Research presented at the Abusive Language Workshops[2] often explores methods to automate the task of detecting and filtering abusive and toxic comments, what Seering et al. (2019) define as *reactive* interventions. We propose the flipside of that task, the promotion of constructive comments, a form of *proactive* intervention (see also Jurgens et al., 2019). While filtering will always be necessary, we like to think that what we define as constructive comments are promoted and highlighted, a positive contagion effect will emerge (Meltzer, 2015; West and Stone, 2014). There is, in fact, evidence that nudges and interventions have an impact on the civility of online conversations and that a critical mass effect takes place with enough polite contributors. Stroud (2011) showed that a 'respect' button (instead of 'like' and 'dislike') encouraged commenters to engage with political views they disagreed with. Experiments indicate that having more polite posts highlighted leads to an increased perception of civility (Grevet, 2016) and that commenters exposed to thoughtful comments produce, in turn, higher-quality thoughtful comments (Sukumaran et al., 2011). Evolutionary game models also support the hypothesis that a critical mass of civil users results in the spread of politeness in online interactions (Antoci et al., 2016).

Shanahan (2018) argues that news organizations ought to be engaged in collecting and amplifying news comments, including seeking out diverse participants and including varied perspectives. The identification of constructive comments can become another tool to help news outlets in fostering better conversations online. The dataset and experiments we present in this paper contribute to that effort.

To illustrate the various choices one can make when considering the quality of a comment, we show a potential spectrum of comments and their quality in Table 1. At the bottom of the table, we see both negative and positive comments that are non-constructive, because they do not seem to contribute to the conversation. The middle comment is not necessarily constructive; it provides an only opinion, and no rationale for that opinion. Such non-constructive, but also not toxic comments, are little more than backchannels and do not contribute much to the conversation (Gautam and Taboada, 2019).

The comment about Trump and Clinton may be perceived as constructive, but it contains abusive language that makes it harder to embrace. Finally, we view the top comment as constructive, because it presents a reasoned opinion, supported by personal experience.

Existing approaches to classifying the quality of online content primarily focus on various forms of toxicity in comments (e.g., Davidson et al., 2017; Kwok and Wang, 2013; Mishra et al., 2019; Nobata et al., 2016; Waseem and Hovy, 2016; Wulczyn et al., 2017). Some research has examined the characteristics of constructive comment threads (Napoles et al., 2017) and our previous work (Kolhatkar and Taboada, 2017a) suggests that we need to consider constructiveness along with toxicity when moderating news comments, because some toxic comments may still be constructive, as shown in the second row of Table 1.

In this paper, we first introduce and evaluate a new annotation scheme for crowd workers to rate the *constructiveness* of an individual comment. This is intended to capture readers' ability to assess comments that add value to the article being commented on. We also create a taxonomy of sub-characteristics, attributes related to constructiveness to further evaluate our definition. We

[2] https://sites.google.com/view/alw3

constructiveness ↑

| Label | Example |
|---|---|
| Constructive | Simpson is right: it's a political winner and a policy dud - just political smoke and mirrors. Mulcair wants Canada to adopt a national childcare model so he can hang on to seats in Quebec, that's all. Years ago I worked with a political strategist working to get a Liberal candidate elected in Conservative Calgary. He actually told his client to talk about national daycare - this was in the early 90's. The Liberal candidate said, 'Canada can't afford that!' to which the strategist responded 'Just say the words, you don't have to actually do it. It'll be good for votes.' I could barely believe the cynicism, but over the years I've come to realize that's what it is: vote getting and power politics. Same thing here. http://www.theglobeandmail.com/opinion/daycare-picks-up-the-ndp/article21094039/ |
| Constructive (toxic) | Please stop whining. Trump is a misogynist, racist buffoon and perhaps worse. Clinton is, to put it in the most polite terms possible, ethically challenged and craven in what she will tolerate in her lust for power. Neither of them is a stellar representative of their gender. Next time, put up a female candidate who outshines the male, not one who has sunk to his same level. Simple. https://www.theglobeandmail.com/opinion/thank-you-hillary-women-now-know-retreat-is-not-an-option/article32803341/ |
| Opinion (no justification) | Please do not print anything Dalton writes and do not report anything he says or does until he does his time in prison. http://www.theglobeandmail.com/opinion/being-clean-and-green-comes-with-a-cost/article27730073/ |
| Non-constructive (positive) | Another wonderful read! Thanks Maggie! http://www.theglobeandmail.com/opinion/david-gilmour-an-agent-of-the-patriarchy-oh-please/article14570359/ |
| Non-constructive (insulting) | Another load of tosh from a GTA Liberal. https://www.theglobeandmail.com/opinion/thank-you-hillary-women-now-know-retreat-is-not-an-option/article32803341/ |

Table 1: Constructiveness spectrum. The first row represents the most constructive and the last row represent the most non-constructive ends of the spectrum. Links are to the article that triggered the comment.

then annotate our corpus with both binary constructiveness labels and sub-characteristics through crowdsourcing. Additionally, we record when a crowd worker indicates agreement with the view being expressed in the comment, to test whether crowd workers assign constructiveness more often to comments they agree with.

The annotated corpus constitutes what we term the *Constructive Comments Corpus (C3)*, which consists of 12,000 news comments in English enriched with a crowd-annotated fine-grained taxonomy of constructiveness and toxicity scores.[3] This is the largest corpus of comment constructiveness annotations that we know of, being approximately 10 times larger than the one described in Kolhatkar and Taboada (2017a).

To illustrate the use of this dataset, we develop both classical feature-based and deep learning systems to classify constructive comments. Our methods make the classification decision based solely on the text of the comment, without relying on the commenters' past behaviour. This is

[3] In line with Bender and Friedman (2018) the discussion in Section 3 is aimed at addressing the key points of a data statement. It is incomplete as we do not have access to information about specific speaker or annotator demographics, though trends about the latter can be found in analyses of crowdworker demographics Posch et al. (2018).

useful when such information is limited as well as when one wishes to evaluate a comment on its own merit alone. We also show the transferability of such models across domains, training on one dataset and evaluating on another.

We outline some challenges and new results in modelling constructive comments. Our analysis highlights that naive models of constructive comments will have length as an overwhelmingly important feature. This obviously produces a trivially tricked model which is likely to be of limited practical value. By exploring several contemporary deep learning architectures, we show that some architectures are robust to this effect, such as Convolutional Neural Networks (CNNs) or Transformer-based models, which represent entire comments in ways that do not depend directly on the input length.

Finally, we note that all code and data are released under open-source and public domain licenses. See Section 6 for links.

## 2. Related work: Identifying high-quality comments

Much of the focus on online comments is on their negative characteristics, including toxicity, abusive language, hate speech and polarization (Saleem et al., 2016; Warner and Hirschberg, 2012; Wulczyn et al., 2017). We take an interest in the positive aspects of online comments, those that make a comment worthwhile and help promote engagement. High-quality comments on an online publication foster a sense of community, in particular if comments are perceived to be moderated (Meyer and Carey, 2014).

Research into what constitutes a high-quality comment has shown that they tend to be constructive, i.e., they contribute an opinion or point of view, and provide reasons or background for that opinion. A 2015 study of the New York Times Picks (Diakopoulos, 2015) showed that Pick comments, compared to non-Picks, have higher argument quality, are more critical, show internal coherence (i.e., they do not excessively rely on context), share some personal experience, are thoughtful, and are readable. Automatically computed measures of readability indicate that Picks require a higher reading level. Interestingly, length was a significant factor in this study, with NYT Picks having an average of 127.2 words per comment (as opposed to 81.7 words for non-Picks). Additional considerations in the commercial world of online publishing likely exist, including increasing reader engagement.

We define high-quality comments as comments that are *constructive*. Constructiveness is a subjective term, infused with moral, historical, and political biases. Previous work that has tackled the issue of constructiveness in online comments and discussions offers a range of definitions. Niculae and Danescu-Niculescu-Mizil (2016) define a constructive online discussion as one where the team involved in the discussion improves the potential of the individuals. That is, the individuals are better off (in a game) when their scores are higher than those they started out with. The definition of Napoles et al. (2017) is characterized as more traditional: comments that intend to be useful or helpful. They define constructiveness of online discussion in terms of ERICs—Engaging, Respectful, and/or Informative Conversations. In their annotation experiment, those were positively correlated with informative and persuasive comments, and negatively correlated with negative and mean comments. Loosen et al. (2018) hear from journalists and content moderators that comments to be promoted are those that present new questions, arguments or viewpoints. There may be, however, some disagreement in which comments are rated as constructive by moderators as opposed to readers (Juarez Miro, 2020). Our definition is based on the existing literature and on our previous work. In Kolhatkar and Taboada (2017a,b), we surveyed online users and built a definition based on their answers. Online users characterize constructive comments as posts that intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response. Furthermore, constructive comments are typically targeted to specific points and supported by appropriate evidence. This definition applies narrowly to online news comments, which is why a contribution is seen as providing an opinion with some justification or evidence for the opinion. It also has many points in common with the definitions in Berry and Taylor (2017), based on rater's intuitions about Facebook comment quality.

Given the large volume of comments that a publisher may want to moderate and label so that they can promote constructive comments, the task is a natural candidate for automatic content moderation. Indeed, most of the work on comment identification and moderation tends to take a text classification approach, modelling characteristics of good and bad comments or threads using supervised techniques. We focus here on the task of moderating individual comments using information in the comment itself (i.e., not metadata about the comment or the author). Thus, we rely on textual aspects within the comment, but additional measures of constructiveness are possible, such as degree of connection between comment and article, as a proxy for relevance. See, for instance, research on probabilistic topic models (Hoque and Carenini, 2019) or on word overlap in article and comment (Risch and Krestel, 2018).

While most of the existing work in automated comment moderation focuses on filtering, i.e., blocking or deleting 'bad' comments, some of the techniques can be applied in the complementary task of finding and promoting 'good' comments. Classic approaches include feature-based classifiers, typically using Support Vector Machines (Davidson et al., 2017; Nobata et al., 2016) or logistic regression (Risch and Krestel, 2018; Waseem and Hovy, 2016) with features such as character and word n-grams, average word length of words in a comment, length of comment or linguistic features (modals and hedges, dependency parses).

Word embeddings are popular in the toxicity detection literature, with methods ranging from averaging pre-trained word embeddings (Nobata et al., 2016; Orasan, 2018) to more contextual models using embeddings from paragraph2vec (Djuric et al., 2015). Deep learning approaches employ recurrent neural networks (Pavlopoulos et al., 2017a,b) or various forms of convolutional neural networks (Gambäck and Sikdar, 2017; Zhang et al., 2018). See also Schmidt and Wiegand (2017) for a general survey.

Although many of these approaches are applicable in the complementary task of identifying constructive comments, a slightly different approach is needed. The task is not just one of pinpointing abusive words and expressions, no matter how subtle; it is about detecting that an argument is being made, that evidence is being provided, and that the comment contributes to the conversation. This is also why toxicity detection methods that use sentiment analysis or polarity of the words in the comment are not useful in this context (Orasan, 2018; Sood et al., 2012).

More specifically with the goal of identifying constructive comments, Napoles et al. (2017) use annotated threads (as opposed to individual comments, as we do here) and model constructiveness using different machine learning models, including a linear model with various features (averaged word embeddings, counts of named entities or length) and a neural model (CNN). Their best performance is with a feature-based model, with an F1 score of 0.73. Park et al. (2016) aim to distinguish NYT Picks from non-picks using a Support Vector Machine classifier with features that are a mix of comment-based (relevance to the article via word similarity, length, readabiliy) and user-based (number of comments the user has posted, average length of those, history of recommendation by others). Cross-validation precision for this system was 0.13, with 0.60 recall, both relatively low because of a small dataset.

In this paper, we add to this literature by introducing a new dataset for constructiveness, which served as the basis for several experiments to build an automatic classifier for comments. Our experiments show that this is a rich and very useful dataset, the largest to date with such annotations, and that the task of identifying high-quality comments poses interesting challenges for the research community.

## 3. Data and annotation

We present C3 (Constructive Comments Corpus), a corpus of 12,000 online news comments enriched with constructiveness annotations and toxicity scores. The 12,000 comments were drawn from the SFU Opinion and Comments Corpus (SOCC), a freely available resource,[4] which contains a collection of opinion articles and the comments posted in response to the articles (Kolhatkar et al., in press). The articles include all the opinion pieces published in the Canadian English-language newspaper *The Globe and Mail* in the five-year period between 2012 and 2016, a total of 10,339 articles and 663,173 comments from 303,665 comment threads. The corpus provides a pairing of articles and comments, together with reply structures in the comments and other metadata. The comments are those that were posted on the paper's website, already moderated through a combination of automatic moderation and flagging by other commenters.

SOCC was initially published with a small subset of labeled comments (1,043 comments), which contain constructiveness and toxicity annotations obtained through crowdsourcing. We will refer to this previously annotated subset as SOCC-a (for 'annotated'). SOCC-a contains binary annotations for constructiveness (constructive or non-constructive) and a four-level toxicity classification.

Our contribution, C3, extends constructiveness annotations to a larger subset of 12,000 comments and introduces a refined annotation scheme that captures sub-characteristics of constructiveness. The comments drawn are top-level comments (i.e., head comments and not replies) from comments threads. The dataset helps fill a gap in this research area. As Vidgen et al. (2019) have pointed out, not enough datasets of adequate quality exist for the task of abusive language detection. The situation is even worse for the task of constructive comment classification. One of them, the SENSEI Social Media Annotated Corpus (Barker and Gaizauskas, 2016) contains only 1,845 comments from 18 articles. The Yahoo News Annotated Comments Corpus (YNACC) (Napoles et al., 2017) is much more extensive, at 9,200 comments and 2,400 threads, capturing characteristics such as sentiment, persuasiveness or tone of each comment. Thread-level annotations in YNACC label the quality of the overall thread such as whether the conversation is constructive and whether the conversation is positive/respectful or aggressive. While useful, this corpus does not contain constructiveness levels for each comment, but for the entire thread. C3 contributes annotations for each comment, with constructiveness labels and constructiveness sub-characteristics, as we describe below.

### *3.1. Annotating constructiveness*

We used Figure Eight,[5] formerly known as CrowdFlower, as our crowdsourcing interface. Contributors read the presented comment, read the article the comment refers to, identify constructive and non-constructive characteristics in the comment, and label the comment as *constructive* or *non-constructive*. Crowdsourcing was the natural choice, given the large number of comments that we wanted to annotate.

Inspired by ideas from the literature on comments and constructive conversations (Diakopoulos, 2015; Napoles et al., 2017; Niculae and Danescu-Niculescu-Mizil, 2016; Zhang et al., 2017), from news value theory in journalism (Galtung and Ruge, 1965; Weber, 2014) and from research on civility and incivility online (Coe et al., 2014; Papacharissi, 2002), we operationalize constructiveness in terms

[4] https://github.com/sfu-discourse-lab/SOCC

[5] https://www.figure-eight.com/

of the presence of a number of constructive characteristics and the absence of non-constructive characteristics. In particular, our annotation scheme consists of the following attributes:

- AGREE: whether the contributor agrees with the views expressed in the comment (yes, no, partially, no opinion)
- CONSTRUCTIVE CHARACTERISTICS: characteristics indicating constructiveness in the comment
  - provides a solution (solution)
  - targets specific points (specific_points)
  - provides evidence (evidence)
  - provides a personal story or experience (personal_story)
  - contributes something substantial to the conversation and encourages dialogue (dialogue)
  - does not have any constructive characteristics (no_con)
- NON-CONSTRUCTIVE CHARACTERISTICS:
  - not relevant to the article (non_relevant)
  - does not respect the views and beliefs of others (no_respect)
  - is unsubstantial (unsubstantial)
  - is sarcastic (sarcastic)
  - is provocative (provocative)
  - does not have any non-constructive characteristics (no_non_con)
- CONSTRUCTIVE: overall whether the comment is constructive or not
- COMMENTS: any comments or suggestions by the contributors

We annotated the extracted 12,000 top-level comments in 12 separate batches, each batch containing 1,000 annotation units. Each unit was annotated by three to five experienced, higher accuracy contributors (referred to as *Level 2 contributors* in Figure Eight terminology). We paid 8 cents per judgment and, as we were interested in the verdict of native speakers of English, we limited the allowed demographic region to the following four majority English-speaking countries: Canada, United States, United Kingdom and Australia.

To maintain the annotation quality, Figure Eight uses *gold questions*, which allow it to measure the performance of each contributor on the annotation task and automatically remove contributors who perform poorly. We created a pool of 300 gold questions, making sure to include gold questions with specific constructive and non-constructive characteristics. For each annotation batch we randomly chose between 90 to 130 gold questions. We set the threshold that requires the contributors to maintain the accuracy of 70% on gold questions. We also included secret gold questions, which were used for our internal quality evaluation.

### *3.2. Data quality*

Reliably measuring inter-annotator agreement of crowdsourced data is still an open problem, as different sets of contributors annotate different sets of questions (Card and Smith, 2018; Jagabathula et al., 2017; Kiritchenko and Mohammad, 2016; Mohammad, 2018; Snow et al., 2008). We measure the reliability of our annotated data using four methods: examining the proportion of instances where the contributors agree; calculating chance-corrected inter-annotator agreement; evaluating crowd performance on secret gold questions; and evaluating crowd answers against expert annotations.

First, despite the subjective nature of the phenomenon, we observed that 66.57% instances of the total 12,000 instances had unanimous agreement among the three to five contributors on the constructiveness question and only about 10% of the instances had serious disagreement, suggesting that humans do have common intuitions about constructiveness in news comments.

Second, the average chance-corrected inter-annotator agreement for the binary classification (constructive or not), measured using Krippendorff's $\alpha$, for the 12 annotation batches was 0.71, suggesting that constructiveness can be annotated fairly reliably with our annotation scheme. This number is much better than the Krippendorff's $\alpha$ of 0.49 in Kolhatkar and Taboada (2017a)'s constructiveness corpus[6] or datasets released for other conversational attributes like toxicity (Thain et al., 2017) ($\alpha = 0.59$) and personal attacks (Wulczyn et al., 2017) ($\alpha = 0.45$).

Third, we had kept aside 20 secret gold questions and measured to what extent aggregated crowd answers matched the answers of these gold questions. These secret questions were not labelled as gold questions in Figure Eight, and thus we were confident that the annotators could not guess that they were quality controls. We observed that the crowd agreed with the secret gold questions 90% to 100% of the time.

Finally, we examined the quality of the crowd annotations with expert evaluations. We asked a professional moderator, with experience in creating and evaluating social media content, to annotate a sample of 100 randomly selected instances. We aggregated the crowd annotations and compared them with expert answers. Overall, the expert agreed with the crowd 87% of the time on the constructiveness question. Among the 13% of the cases where the expert did not agree with the crowd, the majority of the cases were marked as constructive by the crowd and non-constructive by the expert. This may be because the expert had a more critical eye, informed by her experience as a moderator. For instance, in Example 1, a comment on an article about the conflict between religious accommodation in separating men from women and the right to gender equality,[7] the crowd workers labelled it as constructive, perhaps because of its strong stance on equality. The expert annotator pointed out that the comment does not open dialogue.

Example 2 is a response to an editorial[8] that criticizes the Alberta provincial government for banning a right-wing media outlet from its news conferences. The assessment of the expert annotator shows that she has read the comment and the editorial very carefully and believes that the commenter did not understand its nuances.

(1) Comment: Any kind of segregation is abhorrent and cannot be accepted.

Expert annotator: Short. No solution provided, nothing shared to encourage active participation. Doesn't leave room to look at views different from commenter's own.

[6] Kolhatkar and Taboada (2017a) do not report chance-corrected agreement. We calculated it for this paper.

[7] http://www.theglobeandmail.com/opinion/my-quarrels-not-with-york-but-ontarios-rights-code/article16350272/

[8] https://www.theglobeandmail.com/opinion/editorials/why-the-premier-of-alberta-shouldnt-get-to-decide-who-is-media/article28775443/

(2) Comment: What credentials and where do they come from? who issues them? You can't possibly accommodate the whole herd of people who call themselves 'media' today, thanks to the Internet. Someone needs to look at this question and come up with an answer for all governments. Sounds like the Alberta NDP plan to do just that: maybe you should all get on board and make a decision that is followed across the country.

Expert annotator: I can now see how this could have been viewed as constructive, because the questions posed by the commenter relate to the article, but at the same time, the questions included in the comment, and the comment itself show that the commenter either didn't actually read the article or understand the points being made.

### *3.3. Corpus analysis*

We aggregated the responses of all annotators for all of the questions in our annotation scheme. For the constructiveness question, we assigned an *aggregation score* in the range $0.0 \leq score \leq 1.0$. Then each instance is assigned a constructiveness label based on a threshold of 0.5; the instances with $score > 0.5$ were labeled as *constructive* and others were labeled as *non-constructive*. The resulting corpus is slightly higher in constructive comments (6,516) than non-constructive comments (5,484). Among all 12,000 instance, 89.7% instances had a clear consensus among annotators. The remaining 10.3% (1,238) comments had aggregation scores in the range $0.4 \leq score \leq 0.6$, suggesting that there was no clear consensus among annotators on these comments.

Since we were interested in examining how constructiveness and toxicity interact with each other, we enriched our corpus with toxicity scores provided by the Perspective system, a proprietary text scoring algorithm by Jigsaw that produces a number of attributes related to toxicity.[9]

| | **Characteristic** | **# of Comments** |
|---|---|---:|
| Constructive | constructive | 6,516 |
| | dialogue | 7,704 |
| | solution | 5,741 |
| | specific_points | 6,897 |
| | personal_story | 4,217 |
| | evidence | 5,551 |
| | con_other | 4 |
| Non-constructive | non_constructive | 5,484 |
| | provocative | 5,557 |
| | sarcastic | 4,685 |
| | no_respect | 3,043 |
| | unsubstantial | 7,532 |
| | non_relevant | 3,833 |
| | noncon_other | 2 |

Table 2: Description of the C3 dataset. The 'constructive'/'non_constructive' numbers are top-level binary labels where a majority of annotators found the comment constructive or not. For the sub-characteristics, we increase the counter when at least one annotator finds the attribute in the comment.

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

### *3.3.1. Constructive and non-constructive characteristics*

Our annotation scheme decomposes the notion of constructiveness into a taxonomy of constructive and non-constructive sub-characteristics. Table 2 provides a breakdown of the distribution of these characteristics in the C3 dataset.

A question that arises is how well these characteristics perform at capturing the broader attributes of constructive and non-constructive comments. Figure 1 breaks down the presence of the sub-characteristics in constructive and non-constructive comments. Constructive characteristics have a higher prevalence among constructive comments and vice versa for non-constructive characteristics. Indeed, there are no constructive comments that lack some constructive characteristic and likewise for non-constructive. While we do include a 'catch-all' bucket of other characteristics, we found its use negligible. This suggests that these attributes form a comprehensive set of necessary conditions for the presence of constructiveness or non-constructiveness.[10] However, we found that they are not sufficient, as a comment can have one or more constructive sub-characteristics without being constructive.

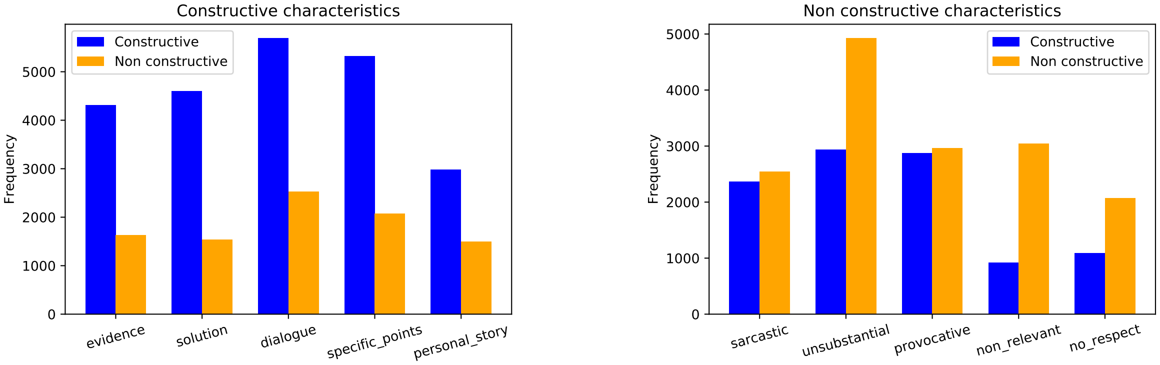

Figure 1: Distribution of constructive and non-constructive characteristics in 6,516 constructive and 5,484 non-constructive comments.

In order to understand the impact of each of the characteristics on whether or not a comment is found to be constructive, we conduct a logistic regression, normalizing for the standard deviation of each of the characteristics. We are able to achieve an F1 value of 0.87 on the logistic regression, again demonstrating that the attributes provide a good breakdown of constructiveness. Table 3 lists the coefficients of each of the characteristics in the logistic regression. The coefficients can be interpreted as the expected change in the log odds of a comment being considered constructive if one more annotator believed it had the corresponding characteristic. From this we see that *dialogue* is the most important predictor of whether a comment is found to be constructive and, for example, that dialogue has 47% more impact on the log-odds of constructiveness than *personal story*. Similarly, we see that a comment being *provocative* does not have much of an impact on its likelihood of being found non-constructive and, indeed, being irrelevant or unsubstantial are much more significant contributors. In summary, the sub-characteristics provide a useful set of criteria for moderation, whether manual or automatic.

[10] Because our survey asked raters multiple questions, it is possible that there was some effect on the raters that would not have happened if they had been asked the questions independently; investigating these potential effects is a possible branch of further work.

| Variable | Coefficient | CI |
|---|---|---|
| dialogue | 0.77 | (0.71, 0.83) |
| solution | 0.74 | (0.67, 0.80) |
| specific_points | 0.59 | (0.53, 0.65) |
| evidence | 0.58 | (0.51, 0.64) |
| personal_story | 0.53 | (0.46, 0.59) |
| provocative | -0.38 | (-0.44, -0.31) |
| no_respect | -0.39 | (-0.45, -0.32) |
| sarcastic | -0.42 | (-0.48, -0.35) |
| non_relevant | -0.84 | (-0.91, -0.78) |
| unsubstantial | -1.17 | (-1.23, -1.10) |

Table 3: Coefficients of normalized logistic regression to predict constructiveness.

#### *3.3.2. Agreeing on the views in the comment*

Intuitively, we might expect annotators to be predisposed to attach the constructive label to comments they agree with. Indeed, when examining the correlation between constructiveness and agreement we get a Pearson correlation coefficient of 0.56 (moderate correlation). This could lead us to believe that constructiveness is really a marker for whether an annotator agreed with the comment rather than a statement about the intrinsic quality of the comment.

In order to differentiate between these two cases, we looked at the 'controversial' comments for which we had at least one annotator agree with the comment and one annotator disagree. For each of these comments, we selected a random agreeing annotator and a random disagreeing annotator. We calculate the inter-annotator agreement between these conflicting annotators on the set of controversial comments. If constructiveness was only a proxy for agreement, we would expect low inter-annotator agreement in this experiment. Instead, we found that the percentage agreement (i.e., the fraction of pairs which gave the same value for whether the comment was considered constructive) was 81.2% (with a Krippendorff's $\alpha$ score of 0.57). Thus, we can establish that the majority of the time, even for comments with disagreement among annotators, constructiveness is measuring something qualitatively different from whether the annotator agrees with the comment.

#### *3.3.3. Constructiveness and toxicity*

We are also interested in understanding the connection between constructiveness and toxicity. As mentioned in Section 2 earlier, toxicity is a negative characteristic of online conversation that has garnered significant research attention. A priori, one might expect there to be a strong negative relationship between constructiveness and toxicity, with constructive comments usually being non-toxic and vice-versa. This would mean that we could rely on existing toxicity detection systems to detect constructiveness.

To understand this relationship, we looked at the correlation coefficient between the aggregated constructiveness scores and the toxicity probabilities given by the Perspective system. We calculated the correlation relationship between the variables (Pearson = $-0.02$, Spearman = 0.04, Kendall $\tau$ = $-0.04$), which is also demonstrated in the scatter plot shown in Figure 2. Additionally, in Section 5 we demonstrate that even a large set of toxic features including toxicity, identity based hate, insults, obscenity and threats achieve relatively low classification scores in modeling constructiveness. We conclude, then, that constructiveness and toxicity are different features. These results are in line with Kolhatkar and Taboada (2017b)'s observation that constructiveness and toxicity are orthogonal.[11]

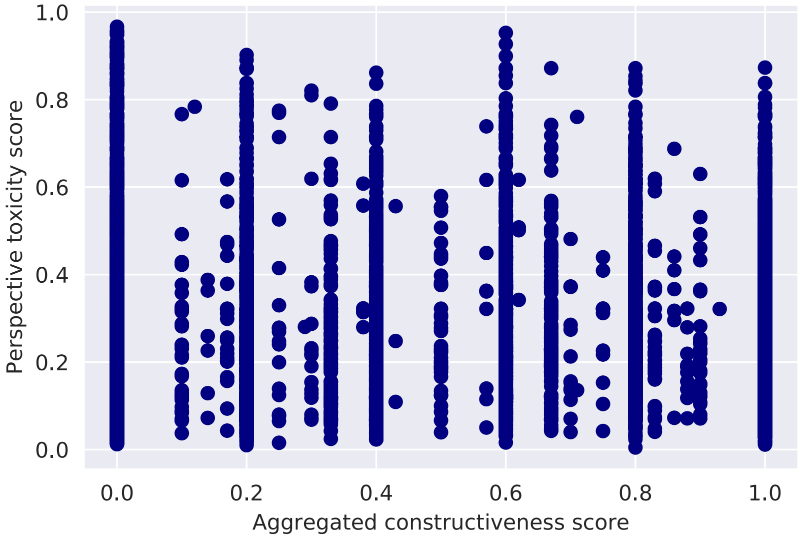

Figure 2: Constructiveness and toxicity.

## 4. Modeling constructiveness

In this section we describe our computational models for identifying constructiveness. We treat the problem of identifying constructive comments as a binary classification problem and investigate the characteristics of constructive comments. We explore classical feature-based models as well as three popular deep learning models: Long Short-Term Memory networks (biLSTMs) and Convolutional Neural Networks (CNNs) with pretrained GloVe embeddings[12] and pretrained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). The implementation details for these architectures can be found in Appendix A.

For the classical feature-based models, we model constructiveness in terms of a number of automatically-extracted features. Our features are primarily derived from two sources: features used by Kolhatkar and Taboada (2017b) and toxicity-related scores given by Jigsaw's Perspective system. We also include some additional linguistic and text quality features that we believe are relevant for the phenomenon. All the features are summarized in Table 4. With these features, we trained sklearn's linear Support Vector Machine (SVM) classifier with Stochastic Gradient Descent learning.[13] We briefly describe the features used in our model as follows.

*Linguistic features.* We incorporate features from Kolhatkar and Taboada (2017a) that capture linguistic aspects of the comment. In particular, we include lexical features, length features, features that are present in argumentative text (e.g., connectives or stance adverbials), named-entity features, and text quality features.

[11] Note that the comments in our dataset are moderated and they rarely contain very toxic comments.

[12] https://nlp.stanford.edu/projects/glove/

[13] We also experimented with other popular classifiers such as Logistic Regression, XGBoost, and Random Forest. We chose SVMs because all of these classifiers performed comparably on our task.

| **Feature class** | **Description** |
|---|---|
| Lexical* (2) | 1- to 3-gram counts and 1- to 3-gram TF-IDF weighted phrases |
| Length* (4) | Length of the comment: number of tokens in the comment, number of sentences, average word length, average number of words per sentence |
| Argumentation* (5) | Presence of discourse connectives (*therefore, due to*)<br>Reasoning verbs (*cause, lead*), modals (*may, should*)<br>Abstract nouns (*problem, issue, decision, reason*)<br>Stance adverbials (*undoubtedly, paradoxically)* |
| Named-entity* (1) | Number of named entities in the comment |
| Text quality* (5) | Readability score, personal experience description score, number of spelling mistakes, number of CAPS words, number of punctuation tokens |
| Content quality† (3) | Text coherence score, unsubstantial probability, and spam probability |
| Aggressiveness† (3) | Aggressiveness expressed in the comment: attack on the author, attack on fellow commenter, attack on the publisher |
| Toxicity† (8) | Toxicity in the comment: severe toxicity, sexually explicit, toxicity, identity hate, insults, obscenity, threats, inflammatory, likely to reject |

Table 4: Automatically-extracted constructiveness features. Features marked with * are from Kolhatkar and Taboada (2017b). Features marked with † are the scores given by Jigsaw's Perspective system.

*Perspective toxicity-related scores.* The Perspective system from Jigsaw is a proprietary text scoring algorithm that produces a number of attributes. It is trained on data similar to that published in the Kaggle Toxic Comment Classification Challenge,[14] where many participants produced high performance models using a variety of neural network techniques. The Perspective system produces a number of scores and we choose 14 of these scores relevant for our task. We organize these 14 features into three groups representing different aspects of constructiveness: content-quality features, aggressiveness features and toxicity features. Content-quality features include probabilities representing text coherence, whether the comment is substantial, and the probability of it being spam. The second group has three features, all related to aggressiveness expressed in the comment: attack on the author, attack on the fellow commenter and attack on the publisher. The third group has eight toxicity related features: severe toxicity, sexually explicit, toxicity, identity hate, insults, obscenity, threats, inflammatory and likely to reject.[15]

*Additional features.* We add three new features in the text quality feature set: number of spelling mistakes, number of capitalized words, and number of punctuation tokens.

## 5. Experiments

We carried out four sets of experiments:

1. Benchmark experiments to compare the models trained on C3 to those of Kolhatkar and Taboada (2017b).

[14] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[15] For more information on these features, please refer to the Perspective documentation: https://github.com/conversationai/perspectiveapi.

2. Feature set experiments to gain insight into the most important features that characterize constructiveness.
3. Domain adaptation experiments to understand how generalizable models trained in one domain are to other domains.
4. Length experiments to investigate to what extent the performance of our models is attributed to length-related features and how we can build more robust models for the phenomenon.

The experiments were conducted using four datasets: C3, SOCC-a, NYT, and YNACC*. We described C3 (12,000 annotated comments) and SOCC-a (1,035 annotated comments) in Section 3 above.[16] C3 was split into 80% train (C3 train) and 20% test (C3 test) portions.

We use two more corpora: the New York Times Picks Corpus (NYT) and a subset of the Yahoo News Annotated Comments Corpus (YNACC*) (Napoles et al., 2017). The NYT corpus contains 15,147 comments chosen as interesting and constructive comments by *The New York Times* moderators before comments were moderated automatically.[17] These comments were extracted by Kolhatkar and Taboada (2017b) using the NYT API.[18] The YNACC* contains 15,178 comments from non-constructive comment threads, a subset of the Yahoo News Annotated Comments Corpus (YNACC),[19] which contains thread-level constructiveness annotations for comment threads posted on Yahoo News articles. The assumption here is that a comment from a non-constructive thread is non-constructive and vice versa, although this assumption may introduce some noise in the data.

### *5.1. Benchmark experiments*

The previous best performance on the SOCC-a dataset was achieved by Kolhatkar and Taboada (2017b) who trained a linear SVM on a combination of data from the NYT and YNACC* datasets and achieved an F1-score of 0.84. In Table 5, we benchmark the linear SVM model trained on C3 against this previous work and show that, despite having about one third of the data used in Kolhatkar and Taboada (2017b), a model trained on C3 achieves a better model performance on the SOCC-a test set.

| **Train dataset** | **Size** | **F1** |
|---|---|---|
| C3 train | 9,600 | 0.87 |
| NYT + YNACC* | 30,325 | 0.84 |

Table 5: Comparison with Kolhatkar and Taboada (2017b). The F1 column shows the F1 score on SOCC-a (size = 1,035) with a linear SVM.

### *5.2. Feature sets experiments*

The goal of these experiments was to gain insight into the most important features that characterize constructiveness. To that end, we examined how individual feature sets contribute to predicting constructiveness.

Table 6 shows the results of these experiments. The first results column shows F1 scores for different models when we trained on C3 train and tested on C3 test. Among the feature-based

[16] Although the original SOCC-a contains 1,043 instances, we had to eliminate 7 instances because they did not correspond smoothly to the comment identifiers in raw SOCC. Our experiments are conducted on 1,035 instances of SOCC-a.

[17] https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html

[18] https://developer.nytimes.com/

[19] https://github.com/cnap/ynacc

models, our four length features are the best predictors of constructiveness. This is not a surprise given the skewed distribution of length in constructive and non-constructive comments (see Figure 3 below). The text-quality and all features also demonstrate comparable performance. Perspective's aggressiveness and toxicity features do not seem to be adequate predictors for constructiveness, confirming our results from Section 3.3.3 that toxicity is orthogonal to constructiveness and presence or absence of toxicity is not a strong indicator of constructiveness.

| **Model** | **TRAIN** | **C3 train** | **C3 train** | **NYT +YNACC*** |
|---|---|---|---|---|
| | **TEST** | **C3 test** | **NYT +YNACC*** | **C3 test** |
| Lexical | | 0.82 | 0.81 | **0.84** |
| Length | | **0.93** | 0.82 | 0.76 |
| Argumentation | | 0.76 | 0.69 | 0.75 |
| Named-entity | | 0.73 | 0.72 | 0.73 |
| Text quality | | 0.90 | 0.82 | 0.81 |
| Content quality | | 0.88 | 0.79 | 0.78 |
| Aggressiveness | | 0.60 | 0.75 | 0.61 |
| Toxicity | | 0.67 | 0.66 | 0.67 |
| All features | | 0.91 | 0.82 | 0.81 |
| biLSTM | | **0.93** | 0.83 | 0.80 |
| CNN | | 0.92 | 0.83 | 0.72 |
| BERT | | **0.93** | **0.84** | 0.78 |

Table 6: Feature and domain adaptation results. Each cell shows the F1 score with the given model and train/test setting.

### *5.3. Domain adaptation experiments*

The second set of experiments was conducted to examine how transferable the learned models are on different datasets and whether the models are capturing general characteristics of constructiveness or topic-specific characteristics. C3 contains comments posted on articles from a Canadian national newspaper, which are likely to discuss Canadian issues and politics, whereas NYT and YNACC* contain comments posted on *The New York Times* and Yahoo News articles, respectively, which are more likely to address American issues and politics. We examined how features learned on one dataset perform on the other. The second results column (C3 train and NYT + YNACC* test) and the third results column (NYT + YNACC* train and C3 test) in Table 6 show the outcome of these experiments. While length is the most important feature in the single domain context, when we move to a new context where the training and test sets differ, its relative importance decreases. In the domain transfer context, text quality and lexical features play an important role.

In case of the deep models, the performance drops markedly when we move to a new context, but they are still the best performing models when trained on C3 train and tested on NYT + YNACC* (second results column of Table 6). When trained on NYT + YNACC* and tested on C3 test, CNNs and BERT perform poorly. This is perhaps because the training data itself is noisy. Recall that NYT contributed the positive cases (constructive) and YNACC* the negative instances. Drawing positive and negative samples from different sources is never ideal, and likely explains the modeling mismatch.

### *5.4. Length experiments*

Longer length is a characteristic feature of constructive comments; people tend to write elaborate and substantial comments when they intend to contribute something to the conversation. Figure 3 shows how this manifests for the C3 corpus where we observed a high correlation of 0.65 between the length of the comment in words and the constructiveness label. We can observe this length correlation in the examples in Table 1. We chose those examples before we realized the importance of length for constructiveness. It seems, however, that we were intuitively inclined to correlate length and constructiveness.

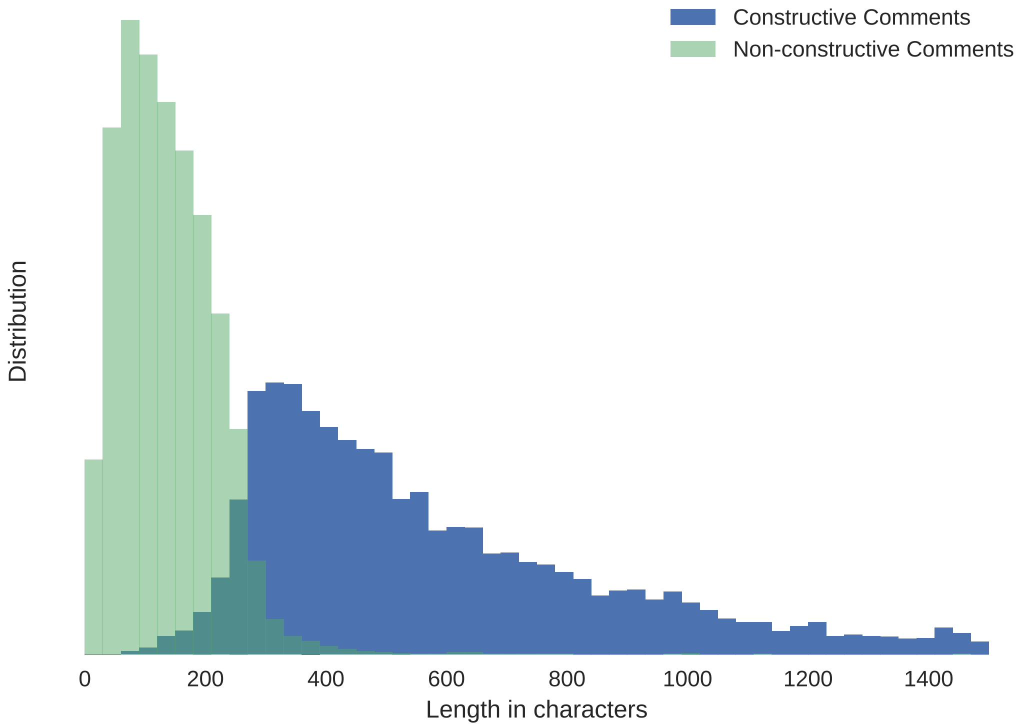

Figure 3: Distribution of comment lengths.

It is unsurprising, then, that many of the best performing features from Table 6 have some relationship with length, though this fact has not been examined in previous work. For instance, the readability score from the text quality features is a function of length of the comment, calculated using the SMOG index formula (Mc Laughlin, 1969), as shown in Equation (1).

$$1.043 \times \sqrt{\text{words with polysyllables} \times \frac{30}{\#\ \text{sentences}}} + 3.1291 \tag{1}$$

Length, however, is not necessarily a generalizable feature for constructiveness, as seen in the domain adaptation experiments in Table 6. It is also vulnerable to adversaries who could attempt to write long low-quality comments in order to fool the models into classifying their comments as constructive.

In this section, we evaluate our models' dependence on length as a feature. It is insufficient to simply compute the correlation of each model's predictions with comment length as we've seen that the true labels already exhibit strong correlation with length. The predictions of any well-performing model, therefore, should have some correlation with comment length. Instead, we will investigate the distribution by length of the errors that each model makes.

For each model, Figure 4 plots a histogram by comment length of the False Negative (FN) and False Positive (FP) errors exhibited by the model on the C3 test set. It can be seen that, for every model, the false positives are distributed over comments of higher length than the false negatives, indicating an over-dependence on length as a signal. Table 7 confirms this insight by exhibiting that the average length of a false negative is less than that of a false positive. This gap is smallest for CNN models, demonstrating how the the max-pooling layer in this architecture reduces its ability to overfit on comment length.

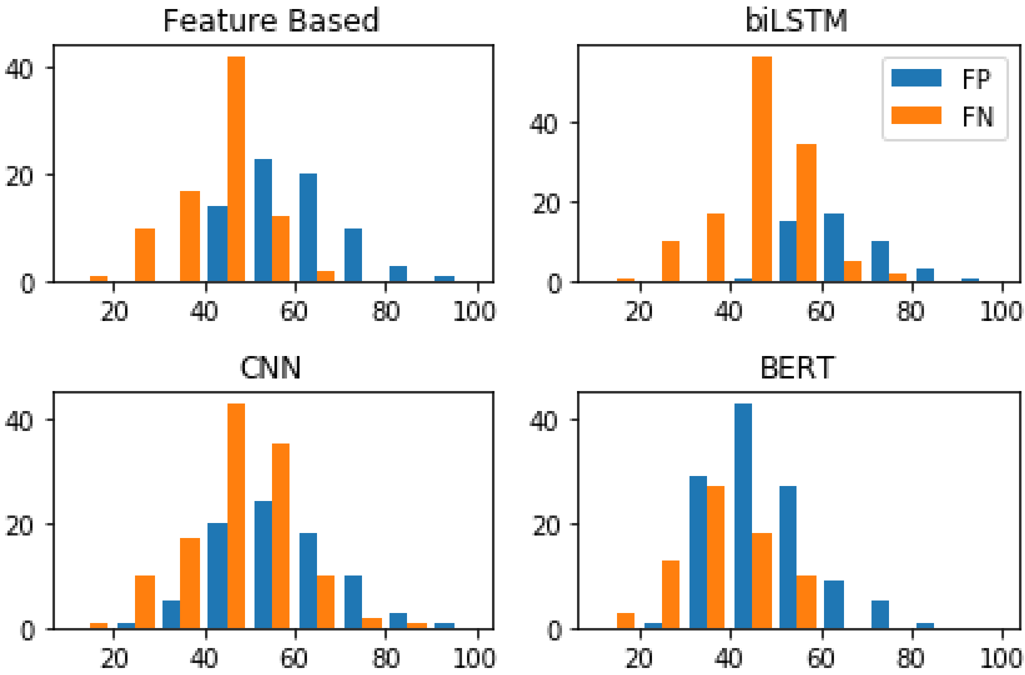

Figure 4: Histogram of model errors by length. FP = False positives. FN = False negatives.

| **Model** | **F1 on C3 test** | **Length Corr.**<br>(on errors) | **FN Length**<br>(mean) | **FP Length**<br>(mean) |
|---|---|---|---|---|
| Feature based | 0.91 | 0.75 | 42 | 63 |
| biLSTM | 0.93 | 0.76 | 45 | 69 |
| CNN | 0.92 | 0.32 | 49 | 59 |
| BERT | 0.93 | 0.56 | 37 | 49 |

Table 7: Results of the length experiments. Length Corr = Length Correlation. FN Length = mean comment length of false negatives. FP Length = mean comment length of false positives.

Table 7 also compares the performance of the models based on their F1 score and their overall dependence on length as a feature. The latter is measured by computing the correlation of the constructiveness score of each model's errors with comment length. Each model shows a positive correlation, though the effect size is dramatically smaller for CNNs than the other model types due to their architectural constraints.

These experiments confirm that length insensitive deep models like CNNs are strong choices to overcome the challenge of length imbalance in constructive data. They are sufficiently flexible to benefit from being trained on the entire dataset but have inbuilt resistance to overfitting to length as a feature.

## 6. Conclusions and further work

This paper contributes: (1) a definition and operationalization of the concept of constructiveness in online comments; (2) a set of linguistic features that can be used to identify constructiveness; (3) an annotated dataset of 12,000 comments; and (4) a deep learning architecture for comment moderation.

We first conduct a thorough exploration of the problem of identifying constructive comments. This is intended to capture when a comment is seen as adding value, i.e., when it is constructive. We explored this using both crowd worker annotations and expert judgments as well as with machine learning methods. The purpose of our work is to contribute to automatic moderation, by identifying high-quality comments, rather than just filtering out undesirable comments. Constructive comments can thus be promoted or highlighted, contributing to better conversations online. Many news organizations try to moderate their comments not only by filtering, but also by promoting constructiveness. The New York Times, for instance, promotes constructive comments as New York Times Picks (Diakopoulos, 2015; Etim, 2017). The approach we present here, which is firmly text-based, can also be combined with metadata such as users' history, to provide a more comprehensive approach to moderation. While our work focuses on online news comments, and on comments in response to opinion pieces, we believe that the overall constructiveness labels and the constructiveness sub-characteristics are applicable to many online conversations and commenting platforms.

The Constructive Comments Corpus (C3) is a set of 12,000 comments annotated by crowd workers, including constructive and non-constructive labels, a breakdown of which constructive and non-constructive characteristics are present in the comment, and also labels for when the human rater agreed with content of the comment. We use this to show that, even on the subset of annotations where one rater agreed with the comment and one disagreed, the two raters agree on the constructiveness label in 81% of the cases. We show that, using our instructions, crowd workers can annotate constructiveness quite reliably (Krippendorff's $\alpha = 0.71$, an improvement from 0.49 in earlier work).

Additionally, our corpus reveals the surprising effectiveness of text-length alone as a feature in the prediction of constructiveness. While this feature would not produce a useful classifier—it is trivially gamed—it leads us to another fascinating result: While many machine learning models 'accidentally' learn to depend on length, CNNs do not, and, moreover, they produce an effective model of constructiveness. This suggests that CNNs are a pragmatic choice of model architecture to support products that highlight constructive comments in a similar style to the New York Times' Editor's Picks.

The high quality of the C3 annotations opens many new avenues of possible research. Further questions may be asked of the data. For instance, we could examine the distribution of constructiveness by topic (see Gautam and Taboada, 2019, for an initial exploration of topics). One could also explore whether some words or parts of speech seem to be predictive of constructive comments. Another direction is to explore the role of a conversation's context. While context is both difficult to make crowd-workers take account of, and rarely affects toxicity judgements (Pavlopoulos et al., 2020), the relationship of constructiveness and context is still unexplored. With more data, other possibilities open up. C3 is still fairly small, especially for data-hungry deep learning models. Expanding this dataset would allow us to explore more complex modelling methods. Attention models could be used to investigate what specific aspects make a comment constructive. The unintended biases in these models should be further probed prior to any sensitive applications. Moreover, much additional exploration is needed to understand the key relationship between agreement with a point made within a comment and the constructiveness of the comment, to disentangle the potential effects of prior human biases. Finally, there are likely to be other potential indicators, beyond constructiveness, of high-quality contributions in comments, such as contributing diverse points of view or healthy levels of disagreement (Muddiman and Stroud, 2017; Shanahan, 2018). The success in modelling constructiveness suggests that there may also be significant opportunities for modelling other kinds of positive contributions to online discussion.

The Constructive Comments Corpus (C3) is available both on Kaggle and on Simon Fraser University's online repository[20,21] and can be cited as Kolhatkar et al. (2020). The code for the experiments conducted in this paper is available on GitHub.[22] We have also created an interactive demo to classify constructive comments.[23]

### About the authors

**Varada Kolhatkar** is a Postdoctoral Teaching and Research Fellow at the University of British Columbia. E-mail: varada.kolhatkar@gmail.com
**Nithum Thain** is a Software Engineer at Google Jigsaw. E-mail: nthain@google.com
**Jeffrey Sorensen** is a Software Engineer working on the Perspective API project for Jigsaw. E-mail: sorenj@google.com
**Lucas Dixon** is a Research Scientist at Google Research. E-mail: ldixon@google.com
**Maite Taboada** is Professor of Linguistics and Director of the Discourse Processing Lab at Simon Fraser University. E-mail: mtaboada@sfu.ca

### Acknowledgements

We thank members of the Discourse Processing Lab at Simon Fraser University for their help in testing the crowdsourcing interface and for lots of feedback. This work was supported by the Social Sciences and Humanities Research Council of Canada and by NVIDIA Corporation, with the donation of a Titan Xp GPU.

### Appendix A. Deep Model Architectures

In this appendix we outline the implementation details of the deep models discussed in Sections 4 and 5.

Our convolutional neural network (CNN) model has a single embedding layer, a single convolution and pooling layer, and a single fully connected layer for the classification head. The embedding layers use pretrained GloVe embeddings of dimension 300 to represent the input word tokens. The convolution layer uses 128 filters each of size 3, 4, and 5 and the pooling layer performs a global max-pooling across the length of the sentence. The fully connected layer produces two values, one for each class.

The bidirectional LSTM (biLSTM) model has a single embedding layer, a single recurrent layer, and a fully connected layer for the classification head. Again, pretrained GloVe embeddings of

[20] https://www.kaggle.com/mtaboada/c3-constructive-comments-corpus
[21] https://dx.doi.org/10.25314/ea49062a-5cf6-4403-9918-539e15fd7b52
[22] https://github.com/kvarada/constructiveness
[23] http://moderation.research.sfu.ca/

dimension 300 are used by the embedding layer. The recurrent layer is a bidirectional LSTM with cells of size 128 for each direction. The fully connected layer produces two values.

The BERT model is built on top of the uncased variant of the pretrained $\text{BERT}_{\text{BASE}}$ model available on TF-Hub.[24] The output sentence representation is then fed into a 3-layer fully connected neural network with layer sizes 256, 128, and 64 and a final classification head. In addition to learning the parameters of the fully connected layers, the model also tuned the top 6 layers of the BERT model, but left the remainder fixed.

At training time, all models use a dropout of 0.5 before the fully connected layers. They are trained with the adam optimizer using a learning rate of 0.001.

## References

## References

Antoci, A., Delfino, A., Paglieri, F., Panebianco, F., Sabatini, F., 2016. Civility vs. incivility in online social interactions: An evolutionary approach. PLoS ONE 11 (11), e0164286.

Barker, E., Gaizauskas, R., 2016. Summarizing multi-party argumentative conversations in reader comment on news. In: Proceedings of ACL 2016. Berlin, pp. 12–20.

Bender, E. M., Friedman, B., 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics 6, 587–604.

Berry, G., Taylor, S. J., 2017. Discussion quality diffuses in the digital public square. In: Proceedings of the 26th International Conference on World Wide Web. Perth, Australia, pp. 1371–1380.

Card, D., Smith, N. A., 2018. The importance of calibration for estimating proportions from annotations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, pp. 1636–1646.

Chen, G. M., 2017. Online incivility and public debate: Nasty talk. Palgrave Macmillan, New York.

Coe, K., Kenski, K., Rains, S. A., 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. Journal of Communication 64 (4), 658–679.

Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017. Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International Conference on Web and Social Media. Montréal, pp. 512–515.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., Jun. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, pp. 4171–4186.

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

[24] https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N., 2015. Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. ACM, Florence, Italy, pp. 29–30.

Etim, B., 2017. The Times sharply increases articles open for comments, using Google's technology. New York Times, June 13.

Galtung, J., Ruge, M. H., 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. Journal of Peace Research 2 (1), 64–90.

Gambäck, B., Sikdar, U. K., 2017. Using Convolutional Neural Networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, pp. 85–90.

Gautam, V., Taboada, M., 2019. Hey, Tyee commenters! Scholars studied you. Here's what they found. The Tyee, November 6.

Gillespie, T., 2018. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, New Haven.

Grevet, C., 2016. Being nice on the internet: Designing for the coexistence of diverse opinions online. PhD dissertation, Georgia Institute of Technology.

Hoque, E., Carenini, G., 2019. Interactive topic hierarchy revision for exploring a collection of online conversations. Information Visualization 18 (3), 318–338.

Jagabathula, S., Subramanian, L., Venkataraman, A., 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. The Journal of Machine Learning Research 18 (1), 3233–3299.

Juarez Miro, C., 2020. The comment gap: Affective publics and gatekeeping in The New York Times comment sections. Journalism, 1464884920933754.

Jurgens, D., Hemphill, L., Chandrasekharan, E., 2019. A just and comprehensive strategy for using NLP to address online abuse. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, pp. 3658–3666.

Kiritchenko, S., Mohammad, S., 2016. Capturing reliable fine-frained sentiment associations by crowdsourcing and bestworst scaling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 811–817.

Kolhatkar, V., Taboada, M., 2017a. Constructive language in news comments. In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, pp. 11–17.

Kolhatkar, V., Taboada, M., 2017b. Using New York Times Picks to identify constructive comments. In: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism. Copenhagen, pp. 100–105.

Kolhatkar, V., Thain, N., Sorensen, J., Dixon, L., Taboada, M., 2020. C3: The Constructive Comments Corpus. Jigsaw and Simon Fraser University, DOI: 10.25314/ea49062a-5cf6-4403-9918-539e15fd7b52.
URL https://dx.doi.org/10.25314/ea49062a-5cf6-4403-9918-539e15fd7b52

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., Taboada, M., in press. The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. Corpus Pragmatics.
URL http://link.springer.com/article/10.1007/s41701-019-00065-w

Kwok, I., Wang, Y., 2013. Locate the hate: Detecting tweets against blacks. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. Bellevue, Washington, pp. 1621–1622.

Loosen, W., Häring, M., Kurtanović, Z., Merten, L., Reimer, J., van Roessel, L., Maalej, W., 2018. Making sense of user comments: Identifying journalists' requirements for a comment analysis framework. SCM Studies in Communication and Media 6 (4), 333–364.

Mc Laughlin, G. H., 1969. SMOG grading: A new readability formula. Journal of Reading 12 (8), 639–646.

Meltzer, K., 2015. Journalistic concern about uncivil political talk in digital news media: Responsibility, credibility, and academic influence. The International Journal of Press/Politics 20 (1), 85–107.

Meyer, H. K., Carey, M. C., 2014. In moderation: Examining how journalists' attitudes toward online comments affect the creation of community. Journalism Practice 8 (2), 213–228.

Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E., 2019. Abusive language detection with graph convolutional networks. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, pp. 2145–2150.

Mohammad, S., 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, pp. 174–184.

Muddiman, A., Stroud, N. J., 2017. News values, cognitive biases, and partisan incivility in comment sections. Journal of Communication 67 (4), 586–609.

Napoles, C., Tetreault, J., Pappu, A., Rosato, E., Provenzale, B., 2017. Finding good conversations online: The Yahoo News Annotated Comments Corpus. In: Proceedings of the 11th Linguistic Annotation Workshop, EACL. Valencia, Spain, pp. 13–23.

Niculae, V., Danescu-Niculescu-Mizil, C., 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, pp. 568–578.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web. Montréal, pp. 145–153.

Orasan, C., 2018. Aggressive language identification using word embeddings and sentiment features. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Santa Fe, pp. 113–119.

Papacharissi, Z., 2002. The virtual sphere: The internet as a public sphere. New Media & Society 4 (1), 9–27.

Park, D., Sachar, S., Diakopoulos, N., Elmqvist, N., 2016. Supporting comment moderators in identifying high quality online news comments. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Diego, California, pp. 1114–1125.

Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I., 2017a. Deep learning for user comment moderation. In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, pp. 25–35.

Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I., 2017b. Deeper attention to abusive user content moderation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Vancouver, pp. 1125–1135.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., Androutsopoulos, I., Jul. 2020. Toxicity detection: Does context really matter? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 4296–4305. URL https://www.aclweb.org/anthology/2020.acl-main.396

Posch, L., Bleier, A., Flöck, F., Strohmaier, M., 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. arXiv preprint arXiv:1812.05948.

Risch, J., Krestel, R., 2018. Delete or not delete? Semi-automatic comment moderation for the newsroom. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Santa Fe, pp. 166–176.

Roberts, S. T., 2019. Behind the Screen: Content moderation in the shadows of social media. Yale University Press, New Haven.

Saleem, H. M., Dillon, K. P., Benesch, S., Ruths, D., 2016. A web of hate: Tackling hateful speech in online social spaces. In: Proceedings of Text Analytics for Cybersecurity and Online Safety (TA-COS), LREC. Portoroz, Slovenia, pp. 1–9.

Schmidt, A., Wiegand, M., 2017. A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia, Spain, pp. 1–10.

Seering, J., Fang, T., Damasco, L., Chen, M., Sun, L., Kaufman, G., 2019. Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland, p. 606.

Shanahan, M. K., 2018. Journalism, Online Comments, and the Future of Public Discourse. Routledge, New York.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y., 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in NLP (EMNLP). Waikiki, HI, pp. 254–263.

Sood, S. O., Churchill, E. F., Antin, J., 2012. Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology 63 (2), 270–285.

Stroud, N. J., 2011. Niche News: The politics of news choice. Oxford University Press, Oxford.

Sukumaran, A., Vezich, S., McHugh, M., Nass, C., 2011. Normative influences on thoughtful online participation. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. Vancouver, pp. 3401–3410.

Thain, N., Dixon, L., Wulczyn, E., 2 2017. Wikipedia Talk Labels: Toxicity. Figshare, 10.6084/m9.figshare.4563973.v2.
URL https://figshare.com/articles/Wikipedia_Talk_Labels_Toxicity/4563973

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., Margetts, H., 2019. Challenges and frontiers in abusive content detection. In: Proceedings of the Third Abusive Language Conference. Florence, Italy, p. 8093.

Warner, W., Hirschberg, J., 2012. Detecting hate speech on the World Wide Web. In: Proceedings of the Second Workshop on Language in Social Media. Montréal, pp. 19–26.

Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of NAACL-HLT. San Diego, CA, pp. 88–93.

Weber, P., 2014. Discussions in the comments section: Factors influencing participation and interactivity in online newspapers reader comments. New Media & Society 16 (6), 941–957.

West, D. M., Stone, B., 2014. Nudging news producers and consumers toward more thoughtful, less polarized discourse. Technical report, Brookings Institution.

Wulczyn, E., Thain, N., Dixon, L., 2017. Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on the World Wide Web. Perth, Australia, pp. 1391–1399.

Zhang, A. X., Culbertson, B., Paritosh, P., 2017. Characterizing online discussion using coarse discourse sequences. In: Proceedings of the Eleventh International Conference on Web and Social Media. Montréal, pp. 357–366.

Zhang, Z., Robinson, D., Tepper, J., 2018. Detecting hate speech on Twitter using a Convolution-GRU based deep neural network. In: European Semantic Web Conference. Heraklion, Greece, pp. 745–760.