Constructiveness and Toxicity in Online News Comments

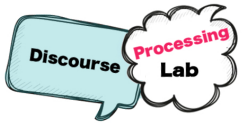
Vagrant Gautam and Maite Taboada

Discourse Processing Lab

Simon Fraser University

mtaboada@sfu.ca

November 2019



1 What is good and what is not about online news comments

When was the last time you saw a thought-provoking and constructive comment posted on a news article? Many of us think that they do not exist and that online comments usually do more harm than good. Online comments, especially those responding to articles about sensitive issues such as climate change, immigration or Indigenous people, often lead to news organizations closing down their comments sections.

Are there no readers left who have something interesting and informative to contribute? Is it the case that there are some constructive comments but they are so few that they are hidden in the ocean of non-constructive and non-informative comments?

Our research group at Simon Fraser University is trying to answer these questions. Over the last few years, we have studied online news comments from several organizations. We were pleasantly surprised to see people expressing different viewpoints, providing evidence to support their opinion, sharing personal stories and experience, attempting to inform, convince or better understand the other side, and overall engaging in a meaningful conversation. We have also, sadly, read many vile and insulting comments.

Our in-depth analysis of more than 1.5 million comments gives us hope, while at the same time providing insights on how to make comment sections better.

2 Constructiveness and toxicity

Constructiveness and toxicity are two axes along which we can evaluate a comment. Intuitively, we expect constructive comments to make well-reasoned arguments and add meaningful discussion. On the opposite end of the spectrum, toxic comments contain insults, profanity or attacks on the authors or people mentioned in an article.

Using automatic text analysis methods, we have classified online news comments along those two categories, constructiveness and toxicity, as a way to help content moderators, journalists and readers make sense and organize comments. We assume that content moderators will want to promote constructive and non-toxic comments. In our data analysis, comments that are constructive and non-toxic read like an essay. They tend to be longer and their sentences are linked cohesively, providing convincing arguments for the point of view expressed. Nobody is insulted, regardless of whether the comment writer agrees or disagrees with the content of the article or with the viewpoints of other commenters.

A non-constructive and non-toxic comment does not add meaningful discussion but does not insult either. Examples of this type of comment tend to be characterized either by polite and unjustified agreement or disagreement with the article.

A non-constructive and toxic comment, on the other hand, consists of directed malice without providing good reasoning. This type of comment often contains colourful language, but this category also includes sarcasm and more subtle insults. Despite the fact that agreement or disagreement with the article's content has nothing to do with a comment being in this category, we found a pattern in the data we considered. Non-constructive toxic comments tended to disagree with the article's content or show disapproval of the views of the author or others mentioned in the article.

Finally, perhaps the most interesting category is the constructive and toxic comment. Almost paradoxically, they contain well-reasoned arguments as well as hateful language. They sound condescending and often follow a format where they open with a toxic insult and continue with good reasoning. The shift is often so dramatic that if the first sentence was ignored, the comment would land squarely in the constructive and non-toxic category. These comments are more common than one would expect and are particularly intriguing because

they mix reasoned argument with insults.

Below is a visual depiction of the four categories with illustrative examples. The non-constructive and toxic category consists of several examples stitched together where words considered insulting are replaced with red asterisks.

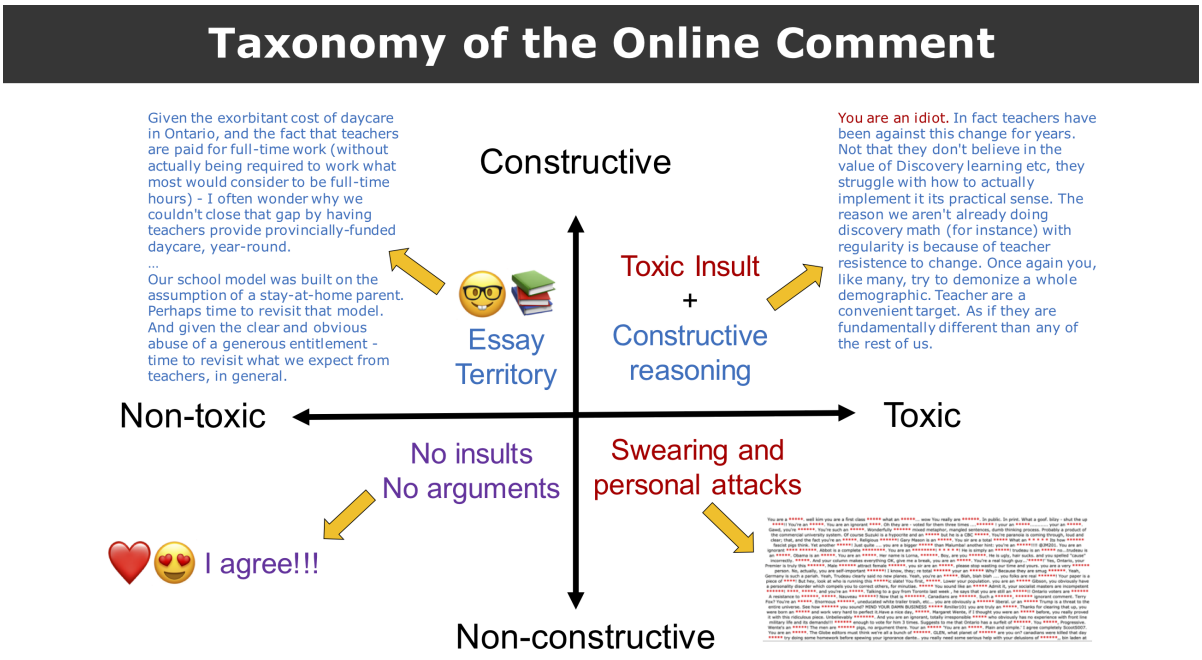


Figure 1. Online comments along the axes of constructiveness and toxicity

3 One and a half million comments

With this taxonomy of the online comment in mind, we examine some of the trends in more than 1.5 million comments collected from three Canadian publications (numbers are rounded):

- 660,000 comments from *The Globe and Mail*, posted on opinion articles published between 2012 and 2016.
- 730,000 comments posted in response to all articles published by *The Tyee* between 2003 and 2018.
- 110,000 comments from all articles published by *The Conversation Canada* between 2017 and 2019.

We analyzed this data with algorithms designed to automatically identify constructiveness and toxicity. We show the results first altogether and then broken up by publication. These graphs all show toxicity on the horizontal axis, the number of comments on the vertical axis, and one line each for constructive and non-constructive comments.

In the visualization of all comments, the highest point in the graph at over 200,000 comments is a blue dot, i.e., non-constructive comments. These comments are low in toxicity—around 0-15% toxicity. Non-constructive

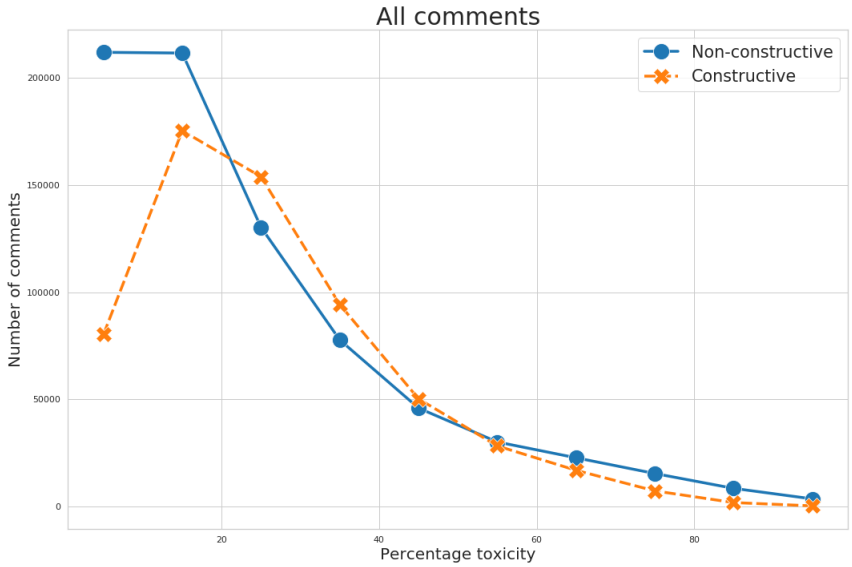


Figure 2. 1.5 million comments - constructiveness and toxicity

and non-toxic comments like "I agree" or "I disagree" fall into this category and make up the majority of online comments in these three publications.

Through most of the graph, non-constructive comments are more common than constructive comments, which also aligns with our intuitions about seeing few 'good' comments on the internet. However, in the central part of the graph with middling levels of toxicity, constructive comments tend to be slightly more common. Additionally, the peak of the constructive (orange) line is at about 15% toxicity. Both these findings imply that **constructive comments tend to contain a small amount of toxicity**. The explanation we suggest for this is that constructive comments provide qualified rather than total agreement or disagreement, unlike non-constructive comments. This qualification gives room for some name-calling in many constructive comments.

The tail end of the graph is towards the right, i.e., high toxicity, where we see proportionately few comments. This is somewhat unusual because our perception of comments online tends to be that many are nasty and hurtful. We propose that this impression can be attributed to the negativity bias (Rozin and Royzman, 2001), a human tendency to pay more attention to negative stimuli. Readers will probably identify with the enduring gnawing feeling of the one instance of negative feedback, even when in a sea of positive praise. There is, undoubtedly, lots of toxicity online. We simply suggest that these comments are not necessarily the place to find it. We should also add, of course, that all three publications moderate their comments.

Another point to notice when looking at the leftmost part of the graph is that at a very low level of toxicity, there seem to be may more non-constructive comments than constructive comments—more than twice as many. This means that there are more than twice as many non-toxic comments like "I agree" or "I disagree" than well-reasoned comments without insults.

Now we will consider each publication individually to see how closely their comments match the average case. Fist of all, we examine comments from *The Globe and Mail*. We have extensively studied these comments, which are part of SOCC, the SFU Opinion and Comments Corpus (Kolhatkar et al., 2019).

The main difference between SOCC comments (comments from *The Globe and Mail*) and the overall pattern is that non-constructive comments are consistently far more frequent than constructive comments at all levels

of toxicity. Additionally, the drop in the number of comments with increasing toxicity is less steep than the overall pattern, indicating that *Globe* comments are more toxic on average.

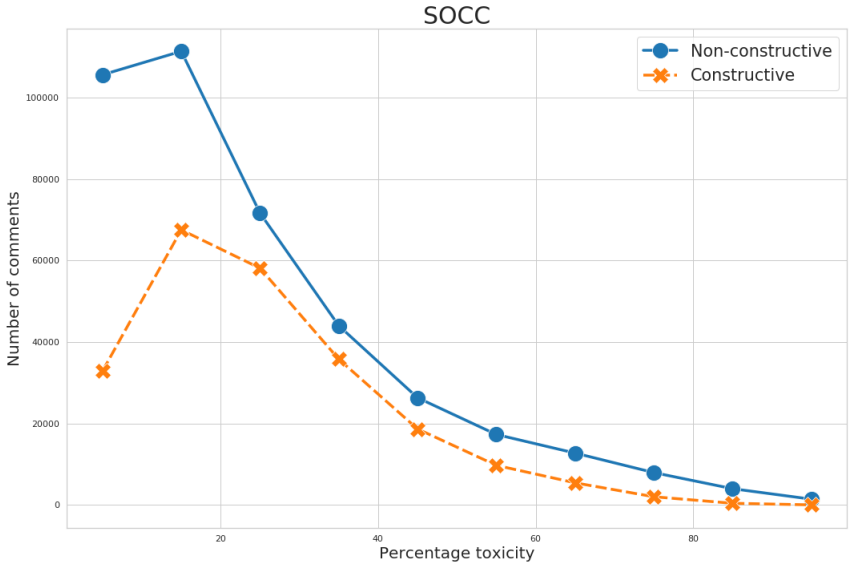


Figure 3. Constructiveness and toxicity in *Globe and Mail* comments

The Tyee, in contrast, has many more constructive comments with the orange line consistently above the blue one starting at about 15% toxicity. The original pattern at a low level of toxicity of more non-constructive than constructive comments is consistent with the overall pattern.

The Conversation seems to have much fewer toxic comments than the average, as the drop in the lines from the left (low toxicity) to right (high toxicity) is very steep. Most interestingly, the low level toxicity pattern is not seen in this publication, i.e., at 0-15% toxicity, there is a very small difference between the numbers of constructive and non-constructive comments. We attribute this to the nature of the publication. With content from academics and researchers, we assume that the audience interested in reading these articles is self-selected to be more cautious in their comments.

4 Which topics trigger the most comments?

We used topic modelling, a technique to discover the themes (henceforth 'topics') in large bodies of data (Blei, 2012). After running a topic modelling algorithm on the large *Globe and Mail* dataset, we identified 15 distinct topics. Below we display word clouds of the 10 most frequent words for each topic.

It is difficult to assign labels to topic classifications created with topic modelling, and one can see from the word clouds that there are some mixed topics. It seems that politics is present in most of the topics, but with some differences between local politics (topic 1, about Ontario; topic 2, about Toronto when Rob Ford was mayor), national issues (topic 4) and international politics (topics 5 and 11). Topics 8 and 12 seem to be about policy (Aboriginal issues, natural resources, education), with topic 9 more specifically about internet policy.

With this topic classification in mind, we first checked which topics receive the most articles, that is, which topics the publications cover the most. We performed this analysis only for *Globe and Mail* data, as it is the only dataset for which we have articles as well as comments. We extracted topics for all articles and assigned

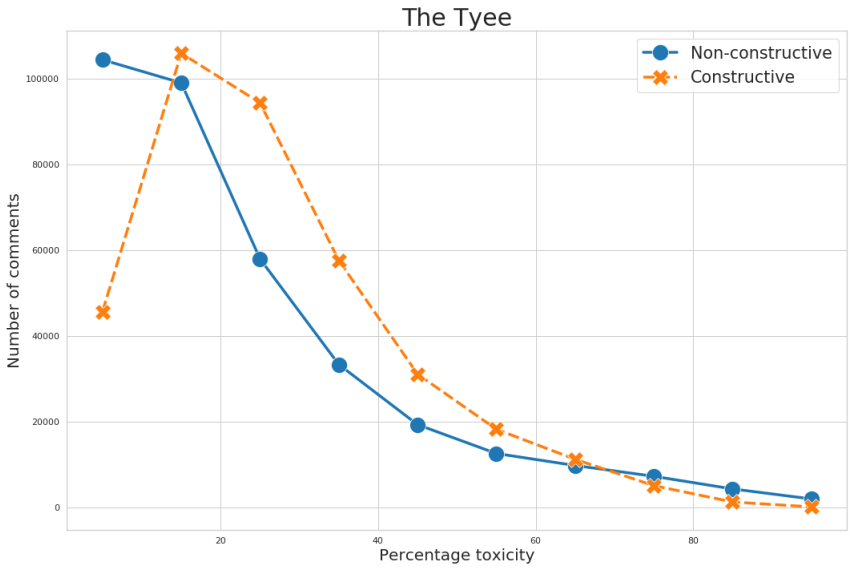


Figure 4. Constructiveness and toxicity in *Tyee* comments

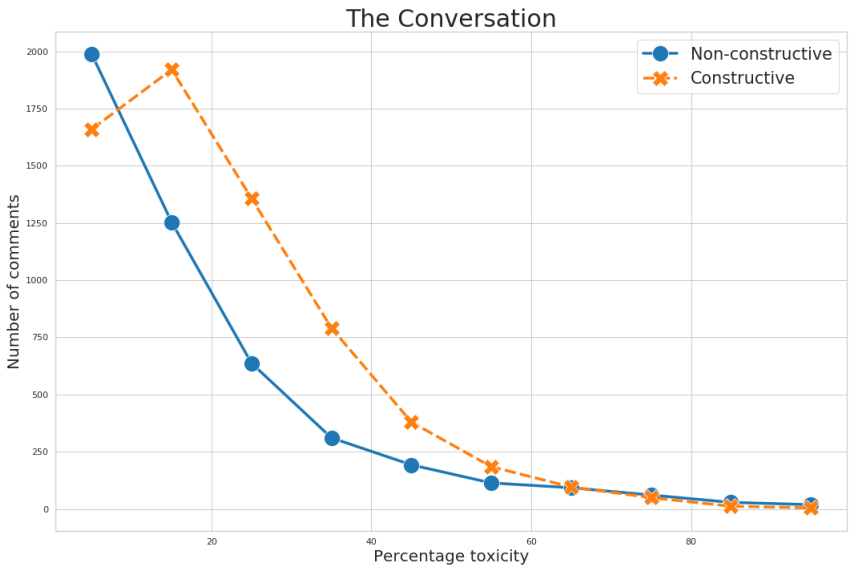


Figure 5. Constructiveness and toxicity in *Conversation* comments

Topic 1 | Topic 2 | Topic 3

Topic 4 | Topic 5 | Topic 6

Topic 7 | Topic 8 | Topic 9

Topic 10 | Topic 11 | Topic 12

Topic 13 | Topic 14 | Topic 15

Figure 6. Topics in *Globe* opinion articles

an article to a topic if the topic model predicted a probability of more than 10% for the topic in that article.

The next figure (Figure 7) shows which topics had the most articles. We clearly see that the three most popular topics of *Globe and Mail* articles are topics 14, 10 and 4.

These topics appear political but are not limited to Canada. The most common one, topic 14, contains the words "British" and "Chinese", suggesting international politics and economics. Topic 4 is clearly about Canadian politics at the national level, with the words "government", "NDP", "Liberals", "Harper" and "party". (Recall that *Globe* articles are from the period between 2012 and 2016.) Topic 10 is about "people", "public" and "system[s]", suggesting local news, further emphasized by the appearance of "Vancouver".

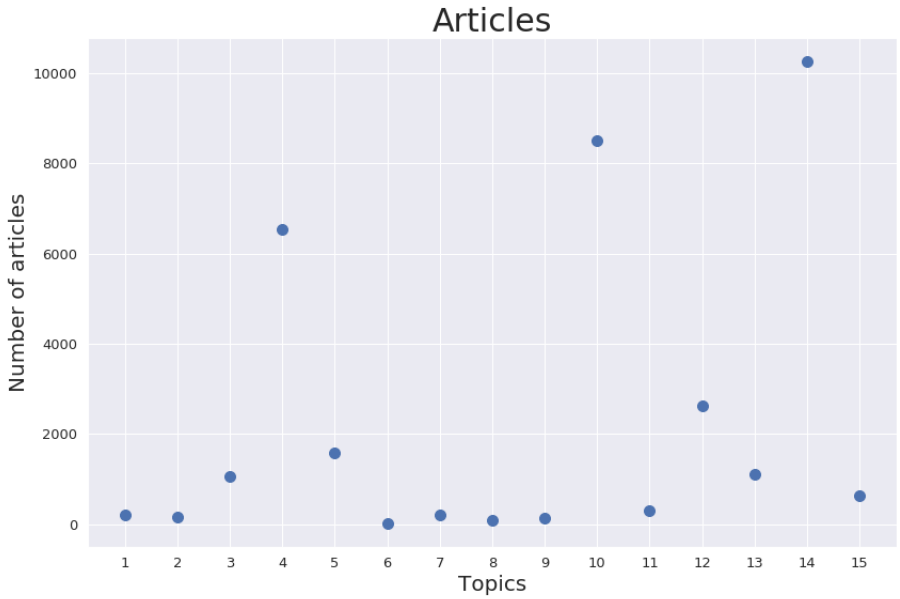


Figure 7. Topics in *Globe and Mail* articles (2012-2016)

Our next question is which topics received the most comments. When predictions were made on the three comment datasets with the topic model, similar patterns emerged (see Figure 8). The top three topics are the same for both articles and comments and they remain the top three by a large margin. The graph in Figure 8 ('Topics in comments') shows the results for all comments combined, but even when split up by publication, the results are not far from each other.

These graphs lead us to our first conclusion, which is that **the proportions of topics discussed in comments seem to correlate directly with those of articles**. This suggests that what people talk about in the comment section is associated with topics that the publications cover the most.

We cannot conclude from these graphs that people comment more about politics than any other topic, however. This is because if there is a correlation between topics in articles and topics in comments, then they are not independent from each other. To find out what people comment most about, we need to account for and normalize the number of articles in any given topic. We do this by dividing the number of comments on a certain topic by the number of articles on that same topic. Interesting results emerge, as shown in the next graph (Figure 9, 'Topics in comments, normalized').

This graphs shows the difference in the proportion of comments about a topic when compared to articles about that same topic. What we see here is a different set of top three topics—topics 6, 7 and 2.

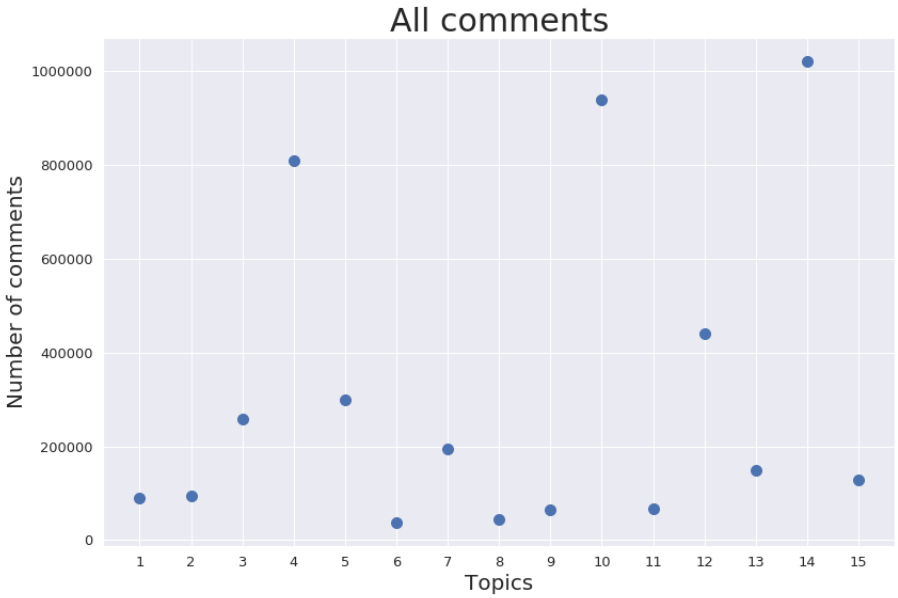


Figure 8. Topics in comments

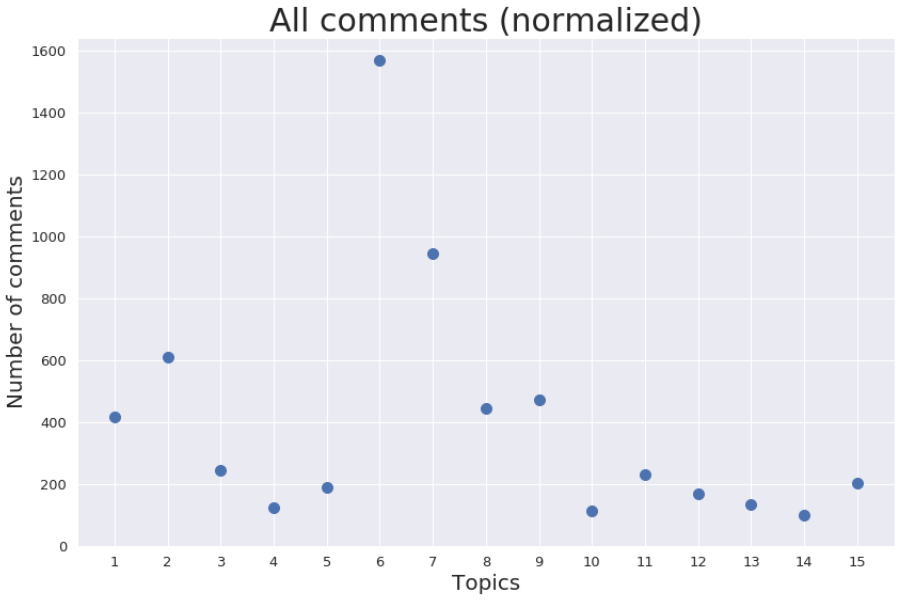


Figure 9. Topics in comments, normalized

In these topics, we see more people mentioned—"Charest", "Redford" and "Romney". We also see more personal, abstract concepts such as "talk", "problem", "matter", "home", "community", "life" and "death". In contrast to these words, articles tend to be more factual and have more concrete nouns. At the same time, judging from the words "legislation", "program", "Trade", "Commons", "candidates", "global" and "local", it seems that comments do talk considerably about politics at different levels. This leads to our second conclusion, that **people comment more about politics than other topics, but that they bring in personal experience and anecdotes when they do so**.

5 Topics and constructiveness

We combined our results to see if there were differences in the constructiveness of comments by the topics being discussed. It turns out that the answers to this question are quite interesting. We began by counting the proportion of comments by topic that were constructive, over all corpora. Since we are calculating the proportion and not the raw numbers of comments, these results are unaffected by how popular a topic is for discussion. Figure 10 shows the average constructiveness score (which was assigned a value of 0 for non-constructive and 1 for constructive) for all 15 topics.

On average, the constructiveness across all topics appears to be between 0.30 and 0.35, indicating that roughly one in three comments is constructive. The fluctuations in constructiveness bear a striking resemblance to the distribution of articles in each topic, though the changes are less dramatic. As in the number of articles by topic, the top three topics are topics 4, 10 and 14. This suggests that there is a higher degree of constructiveness in the comments relating to topics about which more articles are written. **Regular discussion in the media of an issue seems to promote better discussions about it**.

It is also worth pointing out that there is no clear relationship between constructiveness and the topics most often discussed in comments (topics 2, 6 and 7).

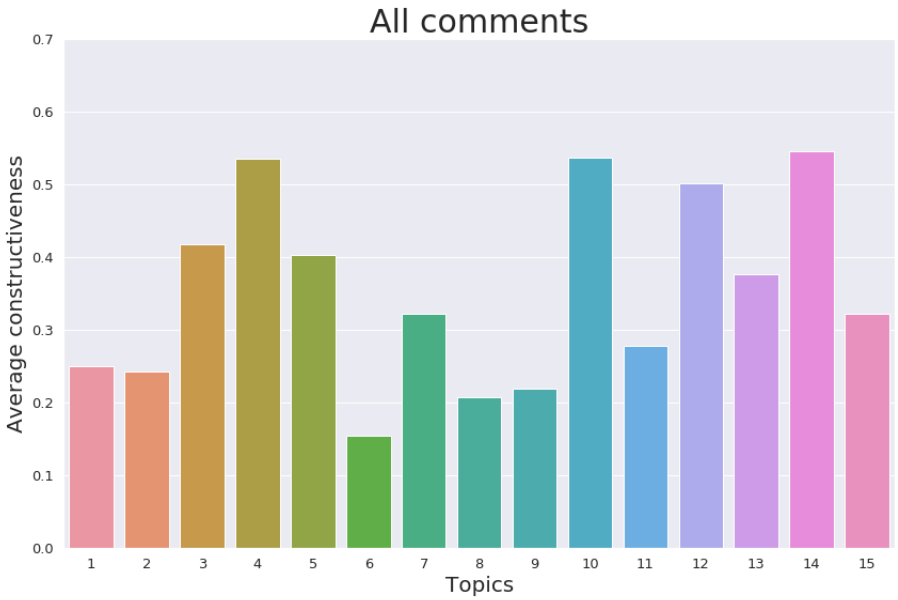


Figure 10. Topics and constructiveness, overall

We now look at the patterns within each dataset as there were interesting differences.

Globe and Mail comments (Figure 11) are **below the average in their constructiveness** overall, but the pattern is remarkably consistent with the pattern seen across the board. The only difference is that the separation is larger between topics with less constructive comments and those with more constructive comments. Topic 6 (top words include "program", "global", "Charest", "death" and "trade") is the least constructive, whereas topics 4, 10 and 14 have the highest constructiveness ratings.

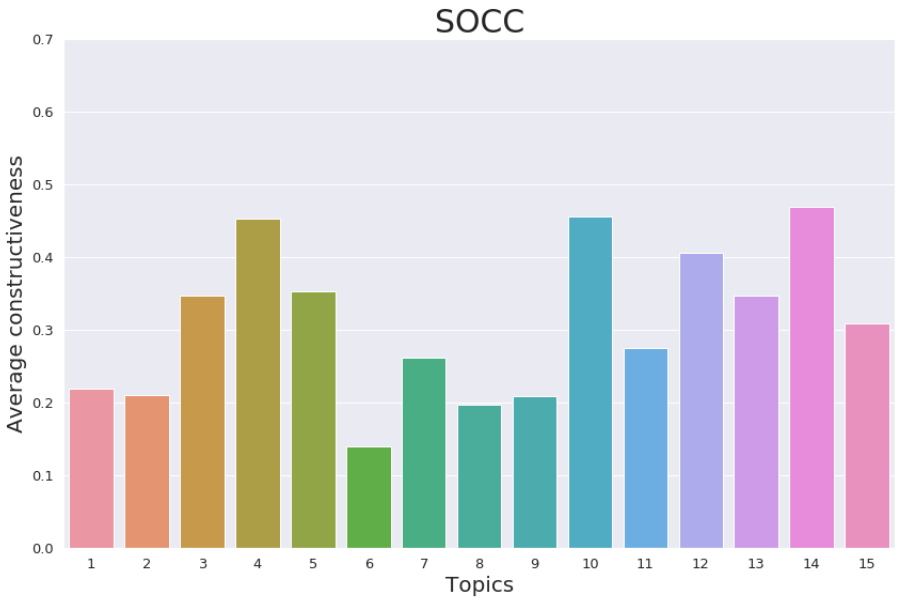


Figure 11. Topics and constructiveness, *Globe and Mail*

The overall pattern of constructiveness of comments made on *The Tyee's* website (Figure 12) is consistent with the average case, except that across all topics, the **constructiveness comes in higher than the average**. The higher constructiveness could be attributed to better moderation of comments, or to efforts by the editorial to encourage civil discourse.

Finally, *The Conversation* (Figure 13) has the same pattern as the distribution of all comments put together, with the exaggerated differences as in *Globe and Mail* comments. One point to make is that **the average constructiveness in *The Conversation* is much higher** than the other publications, with comments on the top three topics (4, 10 and 14) reaching close to a constructiveness value of 0.66. This means that two in three comments on these topics are constructive, twice the average. As before, we would attribute this to the nature of the publication and audience. The Conversation publishes articles by the academic and research community in language and format that are accessible to the general public.

6 Topics and toxicity

The next question is whether there are differences in the toxicity of comments by the topics discussed. Our assumption was that certain topics might be more controversial and lead to more toxic comments.

To our surprise, this is not the case. Unlike constructiveness, toxicity seems to pattern in roughly the same proportion across topics. The graph below shows the average toxicity score (a decimal prediction between 0.0

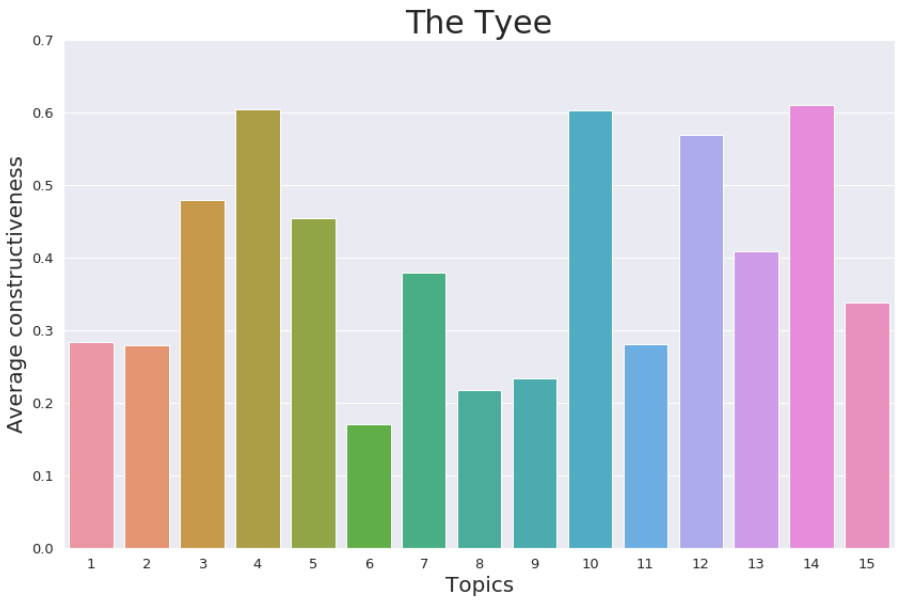


Figure 12. Topics and constructiveness, *Tyee*

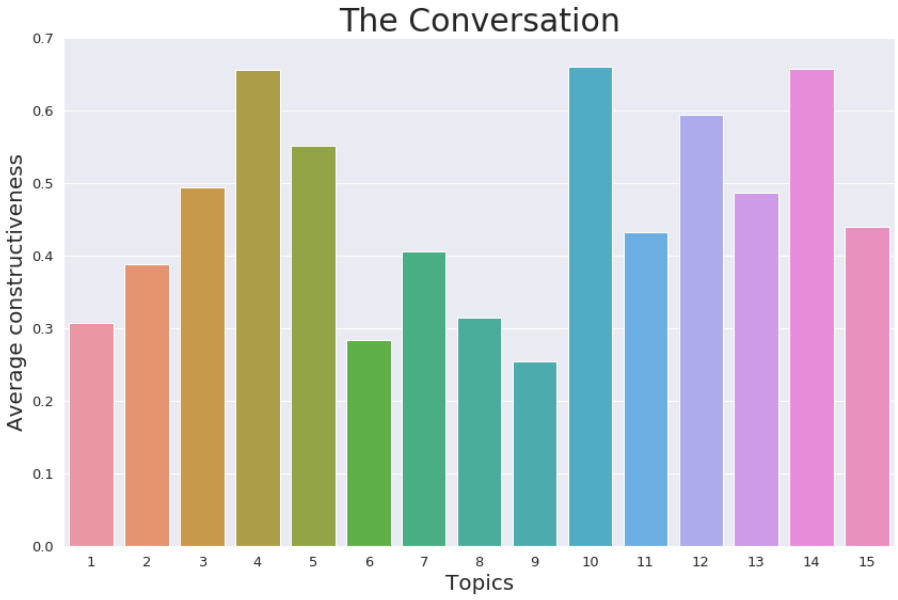


Figure 13. Topics and constructiveness, *Conversation*

and 1.0), over all datasets. As before, we are calculating the proportion and not the raw numbers of comments, so that these results are unaffected by how popular a topic is for discussion.

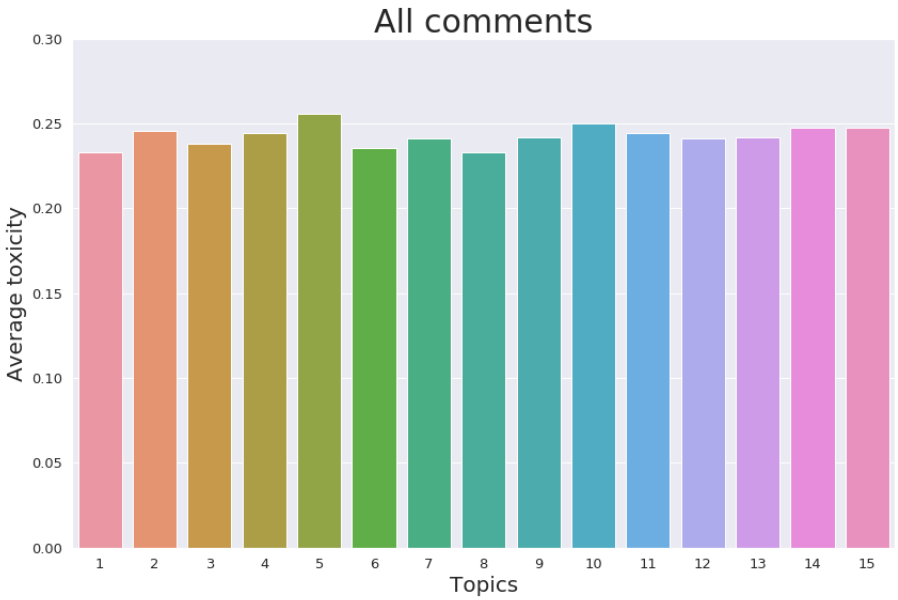


Figure 14. Topics and toxicity, overall

On average, the toxicity across all topics appears to be about 0.23, meaning that topics, on average, have one fourth toxicity. There is not much fluctuation between topics, a pattern that is maintained across the three corpora. This lack of variation suggests that toxicity in comments is not exacerbated by certain topics over others, but is more likely just a feature of online language. For any given article, regardless of what it is about or how often the issue is discussed, **there seems to be a fixed proportion of comments on it that are toxic**.

By publication, *Globe and Mail* comments show a flat, unchanging pattern in their toxicity by topic (Figure 15). This provides support for the claim that there is a fixed proportion of toxic comments on every topic.

The Tyee shows a little more varied distribution in the toxicity of its comments by topic, but these differences are not significant. The average toxicity is the same as the overall average of 0.23.

The Conversation has the same overall pattern, but like its better performance on constructiveness, these comments are also slightly less toxic than online comments in general. The average comment toxicity is 0.18.

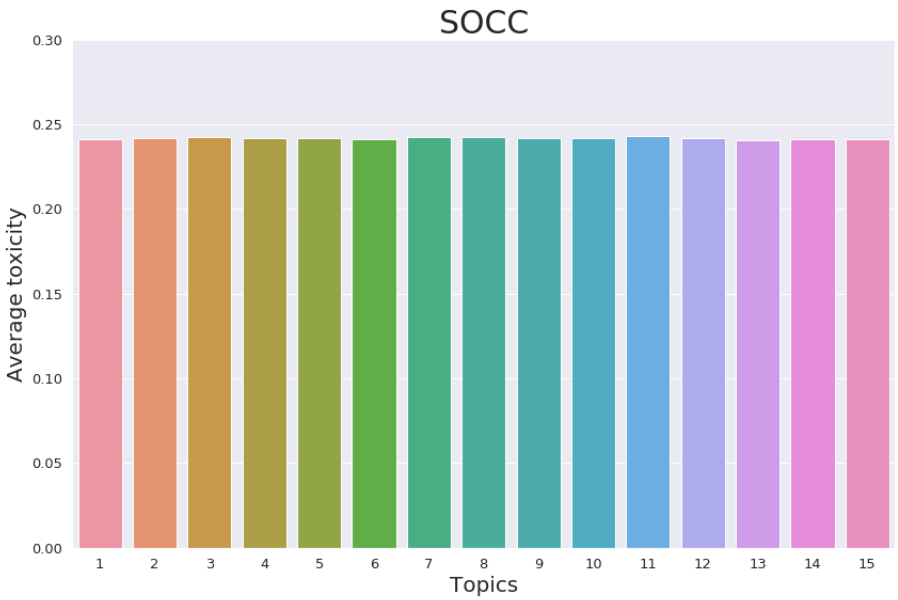


Figure 15. Topics and toxicity, *Globe and Mail*

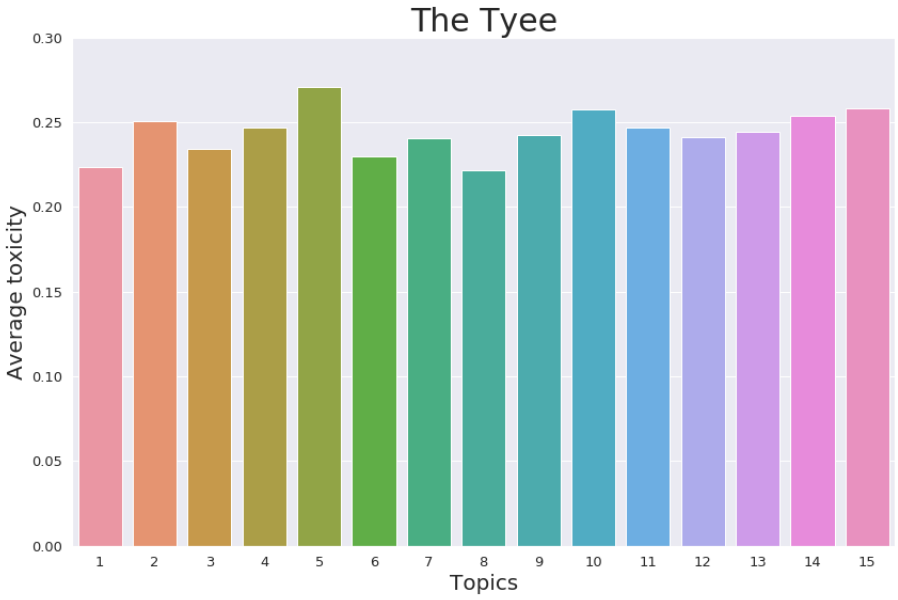


Figure 16. Topics and toxicity, *Tyee*

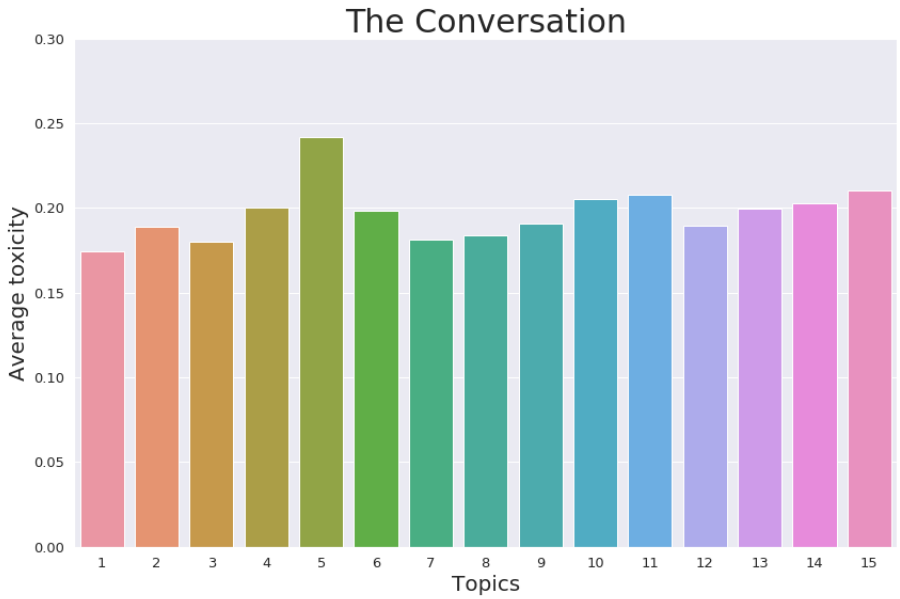


Figure 17. Topics and toxicity, *Conversation*

7 In conclusion

Online news comments exhibit a mix of constructiveness and toxicity. In our study of 1.5 million comments from three Canadian publications (*The Globe and Mail, The Tyee* and *The Conversation*), we have found that most online comments tend to be non-constructive, but also not toxic. A large proportion of comments do contribute to the conversation, providing insight or backing opinion with evidence. At the same time, it is somewhat encouraging to see that there are small numbers of toxic comments (with the caveat that all three publications moderate their comments). We do see differences across publications, with *The Globe* displaying a lower proportion of constructive comments than the other two, which have a more selective audience.

When examining comments by topic being discussed, we see that constructiveness varies across topics, whereas toxicity seems to be equally distributed. It seems that we can expect toxic comments to appear on any topic.

Many more analyses are possible with such a rich dataset. One interesting question is whether we can find a relationship between anonymity and toxicity. Is it the case that commenters with usernames of the type 'ilikecats123' post more toxic comments than people with what appear to be real names? Another interesting question is whether the gender, racial background or ideological position of the article writer influence the toxicity in comments. A large-scale analysis by *The Guardian* found that this was, indeed, the case.

With these and similar analyses, we hope to contribute to the robust debate about the role of human and automatic moderation of online content (Gillespie, 2018; Roberts, 2019; Shanahan, 2018).

Our own work explores how we can best detect constructiveness, so that we can promote or showcase constructive comments. A demo of our current moderation system is available at moderation.research.sfu.ca.

Notes on methodology

Data for all three publications was collected at different times. We thank *The Tyee* and *The Conversation* for making their data available to us. Articles and comments from *The Globe and Mail* were scraped and are available for research purposes.

Topic analysis was conducted with a publicly-available implementation of LDA trained on SOCC articles. Constructiveness analyses are our own, built in collaboration with Jigsaw. Toxicity analyses were carried out with the publicly-available Perspective system. Scripts and code to carry out the analysis are available from our lab's GitHub project page.

The automatic methods used for all the analyses do have some margin of error. Our conclusions are based on the best available evidence provided by those methods.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada (Insight Grant 435-2014-0171 to M. Taboada), an Undergraduate Research Award from Simon Fraser University to V. Gautam and by NVIDIA Corporation, with the donation of a Titan Xp GPU. We thank current and former members of the Discourse Processing Lab, and especially Varada Kolhatkar, for contributions to this research and for many fruitful discussions.

References

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

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

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 2019. URL http://link.springer.com/article/10.1007/s41701-019-00065-w.

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

Paul Rozin and Edward B. Royzman. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296–320, 2001.

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