The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments

Varada Kolhatkar · Hanhan Wu · Luca Cavasso · Emilie Francis · Kavan Shukla · Maite Taboada

January 26, 2018

Abstract We present the SFU Opinion and Comments Corpus (SOCC), a collection of opinion articles and the comments posted in response to the articles. The articles include all the opinion pieces published in the Canadian newspaper *The Globe and Mail* in the five-year period between 2012 and 2016, a total of 10,339 articles and 663,173 comments. The corpus is part of a project that investigates the linguistic characteristics of online comments. The corpus can be used to study, among other aspects, the connections between articles and comments; the connections of comments to each other; the types of topics discussed in comments; the nice (constructive) or mean (toxic) ways in which commenters respond to each other; and how language is used to convey very specific types of evaluation. Our current focus is the study of constructiveness and evaluation in the comments. To that end, we have annotated a subset of the large corpus (1,043 comments) with three layers of annotations: constructiveness, negation, and Appraisal (Martin and White, 2005). This paper details our corpus, the data collection process, the characteristics of the corpus, and describes the annotations. The corpus presented here constitutes an invaluable resource for the

Manuscript under review. Current version: January 26, 2018.

Discourse Processing Lab, Simon Fraser University, Canada.

Varada Kolhatkar
E-mail: vkolhatk@sfu.ca

Hanhan Wu
E-mail: wuhanhan999@gmail.com

Luca Cavasso
E-mail: lcavasso@sfu.ca

Emilie Francis
E-mail: emilief@sfu.ca

Kavan Shukla
E-mail: kavans@sfu.ca

Maite Taboada
E-mail: mtaboada@sfu.ca

study of online comments. While our focus is comments posted in response to opinion news articles, the phenomena in this corpus are likely to be present in many commenting platforms: other news comments, comments and replies in fora such as Reddit, feedback on blogs, or YouTube comments.

Keywords constructiveness · toxicity · sentiment and opinion · negation · Appraisal · news discourse · online comments

1 Introduction

Online commenting allows for direct communication among people and organizations from diverse socioeconomic classes and backgrounds on important issues. Popular news articles receive thousands of comments. These comments create a rich resource for computational linguists, as they are an excellent source of evaluative, abusive and argumentative language; sarcasm; dialogic structure; and occasionally well-informed constructive language. They contain information about people's opinion or stance on important issues, policies, popular topics, and public figures. A number of interesting research questions about journalism, online language, and human conversation can be explored from such a resource. Some example questions are:

1. How can we best organize comments to encourage constructive and civil conversations online?
2. Do the comments express varying views on issues and policies? What is the most popular view? What is the most informative and constructive view supported by evidence?
3. How can we create a succinct summary of different views on an article or issues or policies in the article?
4. Do people engage more in emotionally-driven conversation or in discussion based on facts and evidence?
5. How often do commenters target authors and other commenters, rather than the arguments or issues in the article?
6. How common is sarcasm in online comments?

In order to study these questions systematically, we need a large, well-curated corpus of reader comments. Currently, only a few such corpora are available. One of them is the Yahoo News Annotated Comments Corpus (YNACC) (Napoles et al., 2017).[1] This corpus contains 522,000 comments from 140,000 threads posted in response to Yahoo News articles. Among these, 9,200 comments and 2,400 threads have been annotated at the comment level and the thread level. The comment-level annotations capture characteristics such as sentiment, persuasiveness or tone of each comment; whereas thread-level annotations label the quality of the overall thread such as whether the conversation is constructive and whether the conversation is positive/respectful or argumentative. The other prominent comments corpus is the SENSEI Social Media Annotated Corpus[2] (Barker and Gaizauskas, 2016). The goal of

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=83

[2] http://sensei.group.shef.ac.uk/sensei/corpus.html

this work is to create summaries of reader comments, and accordingly, they have created an annotated corpus of 1,845 comments posted on 18 articles from the British newspaper *The Guardian*. The annotations are done in four stages. First, each comment is labeled with a summary of the main points and the arguments or propositions expressed in the comment. The label is a short, free text annotation, capturing its essential content. A couple of examples are shown in (1).

(1) **Comment:** OK so the lady made a mistake , calm down for Heavens sake.
Label: non-topic; calm down

Comment: That would be fairly pointless, given that Network Rail don't operate passenger trains!
Label: Network Rail do not set fares / operate trains

Second, comments with similar or related labels are grouped together. Each group is assigned a label, which describes the common theme of the group, for example, in terms of topic, propositions, or contradicting viewpoints. Third, annotators write unconstrained and constrained summaries based on the group labels and their analysis of the groups. Finally, the annotators link back the sentences in their constrained summaries to the groups that supported that sentence.

The available corpora contain rich sources of information. We were interested, however, in the link between articles and comments, and in particular in how evaluative language varies between articles that contain opinion and their comments. The SFU Opinion and Comments Corpus (SOCC) contributes to these efforts by providing a pairing of articles and comments, and introducing the largest dataset of this kind to date. Our corpus not only contains comments, but also the articles from which the comments originated. Furthermore, the articles are all opinion articles, not hard news articles. This is important, because it allows for comparisons of evaluative language in both text types, opinion articles and reader comments. Opinion articles are generally subjective and evaluative, but their language tends to be more formal and argumentative. The comments are also subjective; they, however, tend to be more informal and personal in nature. The corpus is larger than any other currently available comments corpora, and has been collected with attention to preserving reply structures and other metadata. In addition to the raw corpus, we also present annotations for four different phenomena: constructiveness, toxicity, negation and its scope, and Appraisal (Martin and White, 2005). We believe the corpus will be an invaluable resource for those interested in the language of evaluation, a host of linguistic phenomena, and how public opinion is expressed through comments.

In the next sections, we describe our raw and annotated corpora in detail, explaining the data collection and annotation methods, and the structure of the corpus. The corpus description and download links are publicly available.[3] We are also publishing all the code and scripts that we used to find articles, scrape them, and clean up the data.[4]

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

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

2 The SFU Opinion and Comments Corpus (SOCC)

In this section, we describe our raw corpus of opinion articles and their corresponding comments.

2.1 Overview

The corpus contains 10,339 opinion articles (editorials, columns, and op-eds) together with their 663,173 comments from 303,665 comment threads, from the main Canadian daily newspaper in English, *The Globe and Mail*, for a five-year period (from January 2012 to December 2016). We organize our corpus into three sub-corpora: the articles corpus, the comments corpus, and the comment-threads corpus.

The articles corpus. This corpus has 10,339 opinion articles, among which 7,797 articles have at least one comment. We have included the remaining 2,542 articles, which did not receive any comments, in the corpus because they can be useful in studying what kind of articles get commenters' attention.[5] The articles were written by 1,628 different authors, and cover a variety of topics from politics and social issues to policies and technology. The articles corpus has 6,666,012 words. The corpus is organized as a CSV. For each article, we provide: date, author, title, URL, the number of comments, the number of top-level comment, and the article text of that article. The detailed information of each field can be found on our project GitHub page.[6]

The comments corpus. The second sub-corpus of SOCC is the comments corpus. This corpus contains the reader comments posted in response to the opinion articles from the articles corpus.[7] The corpus is organized as a CSV containing individual comments and their metadata with minimal duplicate information. The corpus contains all unique comments after removing duplicates and comments with large overlap. The corpus is useful to study individual comments, i.e., without considering their location in the comment thread structure. In our data, we observed that some commenters tend to copy-paste their replies in multiple threads. For instance, a commenter posted the following comment[8] in eight different threads on an article[9] on Canada's vanishing from the United Nations.

(2) Any organisation with 'permanent seats' on its security council is a farce. The UN is that.

[5] Unlike other popular newspapers such as *The New York Times*, *The Globe and Mail* allows comments for all articles.

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

[7] Unfortunately, these comments are not visible on the *Globe and Mail* online interface anymore, as they are in the process of changing their commenting system.

[8] From now on, all examples are from our corpus, and are reproduced verbatim.

[9] `https://www.theglobeandmail.com/opinion/its-not-just-the-drought-treaty-canada-is-vanishing-from-the-united-nations/article10600939/`

If we want to study language used in comments, such repetition is problematic. So we have created the comments sub-corpus by ruthlessly deleting duplicate comments or comments with large overlap and keeping only the most representative ones. The most representative comment is selected based on its source and its length. In Example (3), we show overlap across the three comments in red. We choose to keep (a) and discard the other two, as it is the longest among all the comments and it comes from a more reliable source.[10]

(3) a. Why are you obsessing over our added CO2 when it is net beneficial? Here are some quotes from last week: Prominent Scientists Declare Climate Claims Ahead of UN Summit 'Irrational' — 'Based On Nonsense' — 'Leading us down a false path'. Princeton Physicist Dr. Will Happer: 'Policies to slow CO2 emissions are really based on nonsense. We are being led down a false path. To call carbon dioxide a pollutant is really Orwellian. You are calling something a pollutant that we all produce. Where does that lead us eventually? Greenpeace Co-Founder Dr. Patrick Moore: 'We are dealing with pure political propaganda that has nothing to do with science.'

b. Prominent Scientists Declare Climate Claims Ahead of UN Summit 'Irrational' — 'Based On Nonsense' — 'Leading us down a false path'. Princeton Physicist Dr. Will Happer: 'Policies to slow CO2 emissions are really based on nonsense. We are being led down a false path. To call carbon dioxide a pollutant is really Orwellian. You are calling something a pollutant that we all produce. Where does that lead us eventually? Greenpeace Co-Founder Dr. Patrick Moore: 'We are dealing with pure political propaganda that has nothing to do with science.'

c. Science shows clearly that we are impacting climate Bull. Prominent Scientists Declare Climate Claims Ahead of UN Summit 'Irrational' — 'Based On Nonsense' — 'Leading us down a false path'. Princeton Physicist Dr. Will Happer: 'Policies to slow CO2 emissions are really based on nonsense. We are being led down a false path. To call carbon dioxide a pollutant is really Orwellian. You are calling something a pollutant that we all produce. Where does that lead us eventually? Greenpeace Co-Founder Dr. Patrick Moore: 'We are dealing with pure political propaganda that has nothing to do with science.'

The comments corpus has 37,609,691 words. In the comments corpus, we provide article identifier, thread information, date, the commenter username, comment text, and the popularity of the comment.

The comment-threads corpus. The third sub-corpus of SOCC is the comment-threads corpus. This corpus contains all unique comment threads. The corpus can be used to study online conversations. The number of comments from this corpus is different from the comments corpus because we keep all comments in a conversation intact.

[10] The notion of the source of a comment will be made clear in the *Scraping The Globe and Mail* section.

For example, the repeated comment in example (2), will be kept in all distinct threads, as it plays part in each conversation. This corpus is also organized as a CSV, and the conversation structure is encoded in a field called *comment_counter*. The position of a comment in a comment thread is encoded with numbers separated by underscores, depending upon the level of the comment. For instance:

- First top-level comment: source1_article-id_0
- First child of the top-level comment: source1_article-id_0_0
- Second child of the top-level comment: source1_article-id_0_1
- Grandchildren: source1_article-id_0_0_0, source1_article-id_0_0_1

The comment-threads corpus contain 303,665 threads and 773,716 comments. On average, there were 3 comments per thread. Table 1 shows some statistics of SOCC.

Table 1 Statistics of the SFU Opinion and Comments Corpus

	Item	Frequency
Articles corpus	Number of articles	10,339
	Number of words in articles	6,666,012
	Number of unique article authors	1,628
	Number of articles with comments	7,797
	Average number of comments per article	85
	Average number of threads per article	39
	Average number of top-level comments per article	35
Comments corpus	Number of comments	663,173
	Number of words in comments	37,609,691
	Number of unique commenters	34,472
	Number of top-level comments	272,787
	Average number of comments per commenter	19
Threads corpus	Number of threads	303,665
	Average number of comments per thread	3

2.2 Data collection process

We focused on opinion articles (editorials, columns and op-eds) and their comments. The reason for choosing opinion articles is that these articles tend to receive interesting comments, as the articles themselves are more subjective than hard news articles. We are interested in the difference in subjectivity and evaluative language between articles and comments. The reason for choosing *The Globe and Mail* is that it is the main nationally distributed newspaper in Canada with 6.5 million weekly readers in print and digital.[11] Unfortunately, *The Globe and Mail* does not have an API, so we wrote a scraper to get opinion articles and their comments.[12] Our data is made

[11] `https://www.theglobeandmail.com/report-on-business/globe-has-countrys-largest-weekly-readership-survey/article34119464/`

[12] We contacted *The Globe and Mail*, but were not able to obtain their help in the data collection process.

publicly available under the fair dealing provision in Canada's *Copyright Act*, which permits the use of copyrighted protected works for research purposes.

Figure 1 shows our corpus construction process and below we describe important steps in this process in detail.

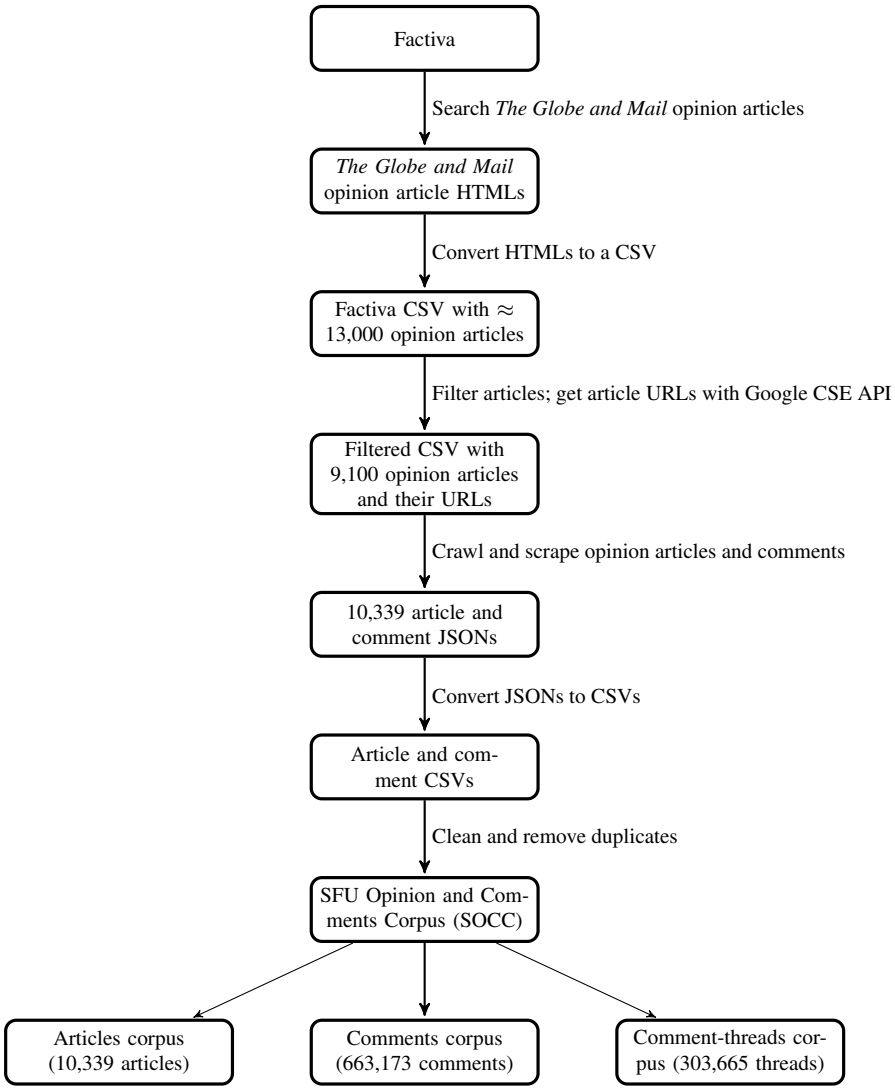

Fig. 1 SOCC construction process

Collecting only opinion articles. We were interested only in opinion articles and did not want to crawl and scrape all articles on the site. We also wanted to restrict our

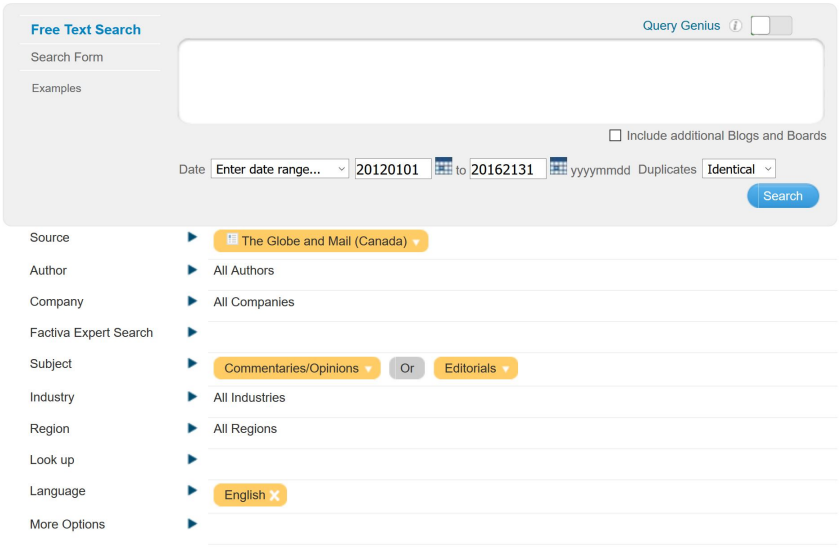

Fig. 2 An example of the Factiva search interface and our search parameters

searches to a specific time period. To get a list of all opinion articles within the last five years, we use Factiva,[13] a business information and research tool that aggregates content from both licensed and free sources, and provides organizations with search, alerting, dissemination, and other information management capabilities. It provides a wide range of information from newspapers, newswires, industry publications, websites, company reports, and more, allowing users to conduct in-depth research on news, companies, industries, and regional affairs. It includes *The Globe and Mail* with same-day and archival coverage. Factiva does not offer an API but it has a search engine where certain parameters can be specified. The search interface is shown in Figure 2. We searched Factiva for *Globe and Mail* opinion articles with the following search parameters.

- **Date:** We selected the date range to be 20120101 (January 01, 2012) to 20161231 (December 31, 2016).
- **Duplicates:** We selected *Identical* from the drop down menu, so that Factiva removes identical duplicates.
- **Source:** We selected *The Globe and Mail (Canada)*.
- **Subject:** We selected *Editorials* and *Commentaries/Opinions* under *Content Types* option.
- **Language:** We selected *English*.

When searched with these parameters, Factiva returned $\approx$ 13,000 results. The interface shows 100 results per page, with various sorting and display options. We sorted the results where the oldest results show up first. For results display, we selected Full Article/Report plus Indexing option. We manually saved each result page as an

[13] `https://global.factiva.com/sb/default.aspx?lnep=hp`

HTML. An example of a Factiva article with indexing is shown in Appendix A. Later we converted the saved HTML files into a comma-separated values (CSV) file, where each metadata field represents a column (e.g., the metadata field LP (lead paragraph) is a column in this CSV). The CSV contains all metadata from the Factiva data. The search result initially yielded over 13,000 articles. Upon closer inspection, we discovered that many "columns" were not really opinion articles, but rather advice columns, listings for events, or articles in specific sections such as health.[14] We identified the labels for those, and removed them from consideration, as they do not express opinion in the traditional sense (editorials or op-eds).

Finding opinion article URLs from opinion articles in Factiva. Factiva contains article text and its metadata. However, it neither includes reader comments nor the article URL to scrape article comments. We found URLs by creating search queries from the Factiva CSV and then using these search queries with two different search engines. We created search queries using the first few sentences of the article text or the lead paragraph text, and then we searched for the appropriate URLs with the Google Custom Search Engine (CSE) API. We restricted the search parameters so that we only looked for the text on the *Globe and Mail* website. The CSE API places a restriction on the number of searches per day and thus this process of gathering URLs for opinion articles took a few weeks. Once we had article URLs, we added them to our Factiva CSV file. In a few cases (less than 200 cases), the Google CSE API did not return any results, and then we used the same search queries manually on the Bing search engine.[15] In the end, we had 9,100 opinion article URLs.

Scraping The Globe and Mail. We considered the set of URLs collected in the previous step as seed URLs for our crawler. The crawler crawled each of these seed URLs and looked for other opinion articles[16] satisfying our year-range criteria. We started with 9,100 distinct seed URLs, and ended up with 10,339 opinion articles between January 1, 2012 and December 31, 2016. We scraped these articles and their corresponding comments using scrapy.[17] The scraped output is stored into JavaScript Object Notation (JSON) files.

Our comments are scraped from two different sources and with two different methods because, during our comment extraction process, *The Globe and Mail* changed their commenting structure. In particular, they added a *reactions* option for the commenters. Before this change, the commenters could only *like* or *dislike* a comment. With the new interface, commenters had a variety of options, such as *Like*, *Funny*, *Wow*, *Sad*, and *Disagree*. Because of this change, the comments posted before the date of the change (2016/11/28) disappeared from the website. We contacted *The Globe and Mail*, and they could not guarantee that the disappeared old comments

[14] For example, `https://www.theglobeandmail.com/life/health-and-fitness/ask-a-health-expert/which-granola-bars-are-the-healthiest-to-eat/article11703335`

[15] `https://www.bing.com/`

[16] We consider a scraped article as an opinion article if its URL starts with `http://www.theglobeandmail.com/opinion/`.

[17] `https://doc.scrapy.org/en/latest/`

would be recovered in future. Therefore, in our code, we have two separate methods to extract comments: the method to extract the old comments that were present before the comment structure change and the method to extract new comments with commenter reactions. We refer to the first as source1 comments and the second as source2 comments. There was some overlap between the old and new comments, and thus we preprocessed the corpus to remove the duplicate comments. That said, if a thread from source1 is slightly different from a similar thread from source2 (only one comment is different), we have kept both of these threads by considering them as two separate conversations.

Organizing data into CSV files. To distribute the data, we have organized the three sub-corpora into three comma-separated values (CSV) files.

1. **The articles corpus CSV (gnm_articles.csv)**
   This CSV contains opinion articles from *The Globe and Mail*, with the following fields.
   - **article_id:** A unique article identifier, which is also used in the comments CSV.
   - **title:** The headline of the article.
   - **article_url:** *The Globe and Mail* URL or the article.
   - **author:** The author of the opinion article.
   - **published_date:** The date of publication of the article.
   - **ntop_level_comments:** The number of top-level comments for this article.
   - **ncomments:** The number of all comments for this article.
   - **article_text:** The article text with preserved paragraph structure.
2. **The comments corpus CSV (gnm_comments.csv)**
   A CSV containing individual comments and their metadata with minimal duplicate information. This CSV can be used to study individual comments in isolation, i.e., without considering their location in the comment thread structure.
   Below we list the most relevant columns of this CSV. Our GitHub page contains an exhaustive list of all columns.
   - **article_id:** The article identifier, also used in the article CSV.
   - **comment_counter:** The comment counter, which is a unique comment identifier and encodes the location of the comment in the associated comment thread.
   - **comment_text:** The comment text. We have carried out minimal preprocessing on this text, where we have deleted HTML characters and added missing spaces after punctuation.
   - **comment_author:** The author of the comment.
   - **time_posted:** The time when the comment was posted.
3. **The comment-threads corpus CSV (gnm_comment_threads.csv)**
   A CSV containing comment threads and their metadata. In this CSV, we retain all comments in threads, and you may find duplicate comments. The columns of this CSV are same as the comments corpus CSV.

3 Constructiveness and toxicity annotations on SOCC

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

We annotate a small sample of our SOCC corpus for constructiveness and toxicity level of individual comments. The goal of creating this corpus is twofold: first, to examine to what extent people agree on these notions and second, to examine the relationship between toxicity and constructiveness. Note that this corpus and the analysis is also described in our previous papers (Kolhatkar and Taboada, 2017a,b). That said, the constructiveness and toxicity corpus we are releasing with this article[18] is a carefully curated corpus and contains an extra layer of expert annotations. The goal of the annotations is to develop methods to detect constructiveness, with both supervised learning and deep learning approaches, which we have explored in preliminary work (Kolhatkar and Taboada, 2017a,b).

3.1 Definitions

Rather than providing dictionary definitions, or rely on our intuitions, we decided to post an online survey, asking people what they thought a constructive comment was. This is a form of crowdsourcing a definition. We are interested in crowd definitions, because presumably it is that population that posts comments on news sites. We posted a survey through SurveyMonkey,[19] requesting 100 answers to the question 'What does *constructive* mean in the context of news comments?'. Representative samples of the answers are in Table 2.

Constructive	Non-constructive
provides evidence-based information	opinions without support
offers an alternative viewpoint	merely assigns a blame
builds up and does not tear down	dismisses the terms of debate
asks an informed question	emotional reactions
provides a well-researched answers	excessively flattering
adds new information or provides a new perspective	personal or derogatory
is specific and references facts	irrelevant or too general

Table 2 Sample answers to the question 'What does *constructive* mean?'

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

[19] https://www.surveymonkey.com/

Previous papers that have tackled the issue of constructiveness in online comments and discussions offer different 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. In our annotation experiment, we used the following definition of constructiveness which was inspired by our survey answers: *Constructive comments intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response; they are typically targeted to specific points and supported by appropriate evidence.*

We propose the label *toxicity* for a range of phenomena, including verbal abuse, offensive comments and hate speech. A *toxic* comment is one that is likely to offend or cause distress. Previous definitions have included personal attacks (Wulczyn et al., 2017), abuse (Nobata et al., 2016), harassment (Bretschneider et al., 2014), threats (Spitzberg and Gawron, 2016), use of profane, obscene or derogatory language (Sood et al., 2012; Wang et al., 2014; Davidson et al., 2017), inflammatory language (Wiebe et al., 2001), hate speech (Warner and Hirschberg, 2012; Djuric et al., 2015; Waseem and Hovy, 2016), or apply the more general term cyberbullying, which may be found not only in the language used, but also in other disturbing online behaviour such as repeated messages or threat of public exposure (Reynolds et al., 2011; Pieschl et al., 2015). For our annotations, we define toxicity on a 4-point-scale (very toxic, toxic, mildly toxic, not toxic), where each point on the scale is defined in terms of the following characteristics. The definition for *very toxic* included comments which use harsh, offensive or abusive language; comments which include personal attacks or insults; or which are derogatory or demeaning. *Toxic* comments were sarcastic, containing ridicule or aggressive disagreement. *Mildly toxic* comments were described as those which may be considered toxic only by some people, or which express anger and frustration.

3.2 Constructiveness annotations

We carried out a preliminary crowdsourcing experiment, where we annotated 1,121 comments for constructiveness and toxicity from SOCC.[20]

Interface and settings. We used CrowdFlower[21] as our crowdsourcing interface. We asked annotators to read the article each comment refers to and to label the comment as constructive or not. For quality control, 100 units were marked as gold: Annotators

[20] This dataset is available at: `https://github.com/sfu-discourse-lab/Constructiveness_Toxicity_Corpus`

[21] `http://www.crowdflower.com/`

were allowed to continue with the annotation task only when their answers agreed with our answers to the gold questions. As we were interested in the verdict of native speakers of English, we limited the allowed demographic region to English-speaking countries. We asked for three judgments per instance and paid 5 cents per annotation. Figure 3 shows our annotation interface.

Agreement and results. Percentage agreement for the constructiveness question on a random sample of 100 annotations was 87.88%, suggesting that constructiveness can be reliably annotated. In our dataset, constructiveness is more or less equally distributed: Out of the 1,121 comments, 603 comments (53.79%) were classified as constructive, 517 (46.12%) as non-constructive, and the annotators were not sure in only one case. Below we show examples of constructive and non-constructive comments from our corpus on an article on a newly-proposed national daycare plan.[22] The comment shown in Example (4) is clearly a constructive comment. It is relevant to the article; it addresses a specific point (the cost of such a program); and it proposes an alternative solution, with a detailed description. The comment shown in Example (5), on the other hand, is not constructive. It dismisses the idea of a national daycare plan and criticizes the NDP's approach without much explanation or evidence.[23] Note that neither the constructive nor the non-constructive comment is particularly well-written. For example, there are minor punctuation problems, such as missing punctuation or missing space after punctuation, in both comments.

(4) While I support the notion of subsidized daycare, a national daycare program is an expensive boondoggle waiting to happen.A means tested subsidy paid directly to parents who use qualified facilities would create opportunities to increase available spaces. I would support a subsidy using the following guidelines:First, no family with income over $100,000 per year should require a subsidy. Subsidies could be scaled according to income levels from 25% to 75%Second, people should expect to pay at least $10/day ($200/month) at any income level.Third, facilities that qualify for subsidies should be required to offer a minimum standard of service.

(5) Pay for your own kids' babysitting.The last thing that any problem needs is an NDP style, big goverment, one-size fits all approach.

3.3 Toxicity annotations

In the context of filtering news comments, we are also interested in the relationship between constructiveness and toxicity. To better understand the nature of toxicity and its relationship with constructiveness, we included toxicity annotations in our CrowdFlower annotation. For the 1,121 comments, we also asked annotators to identify toxicity. The question posed was: How toxic is the comment? As explained in Section 3.1, We established four classes: *Very toxic*, *Toxic*, *Mildly toxic* and *Not toxic*.

[22] `https://www.theglobeandmail.com/opinion/daycare-picks-up-the-ndp/article21094039/`

[23] NDP=National Democratic Party, one of the main political parties in Canada.

Figure 3 shows our annotation interface for annotating toxicity level. The annotation parameters are already explained in Section 3.2. The percentage agreement for the toxicity question on a random sample of 100 annotations provided by CrowdFlower was 81.82%.

The distribution of toxicity levels by constructiveness label is shown in Table 3. The most important result of this annotation experiment is that there were no significant differences in toxicity levels between constructive and non-constructive comments, i.e., constructive comments were as likely to be toxic (in its three categories) as non-constructive comments. For instance, consider Example (6) below. It was labelled as constructive by two out of three annotators and our expert, and toxic by all three and the expert, as it includes personal attacks on Trump and Clinton. It could be the case, in some situations, that a moderator may allow a somewhat toxic comment if it adds value to the conversation, i.e., if it is constructive.

(6) 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.

In our corpus, constructiveness and toxicity are orthogonal categories. The results also suggest that it is important to consider constructiveness of comments along with toxicity when filtering comments, as aggressive constructive debate might be a good feature of online discussion. Given these results, the classification of constructiveness and toxicity should probably be treated as separate problems.[24]

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

Table 3 Percent distribution of constructive and toxic comments in CrowdFlower annotation

3.4 Expert evaluation

To examine the quality of the crowd annotations we asked a professional moderator, with experience in creating and evaluating social media content, to evaluate the acceptability of the crowd's answers. For that, we randomly selected 222 instances

[24] Take these results with a grain of salt, as our corpus has only moderated comments and the really toxic comments identified by *The Globe and Mail* comment moderation system are not present in our corpus.

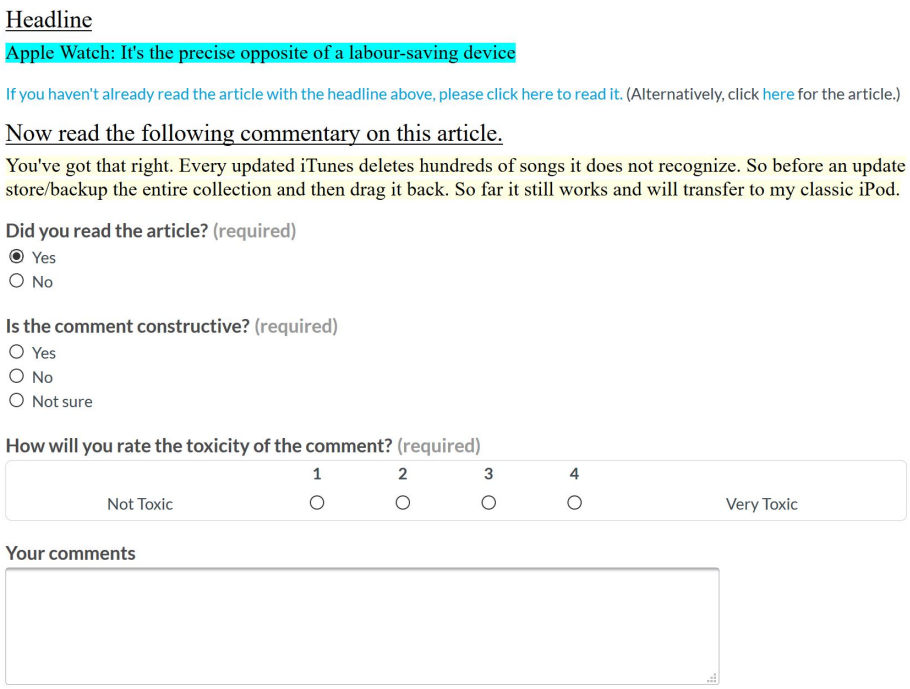

Fig. 3 CrowdFlower interface to annotate constructiveness and toxicity

from the crowd-annotated data. We made sure to choose instances with medium confidence ($0.6 \leq$ confidence < 0.9) and high confidence ($0.9 \leq$ confidence ≤ 1.0).[25] We asked the expert whether they agree with the crowd's answer on constructiveness or not. We also asked them to rate the toxicity of the given comments on a scale of four toxicity levels. We add this layer of expert annotations in our constructiveness and toxicity corpus.

Overall the expert agreed with the crowd 77.93% of the time on the constructiveness question. Among 22.07% of the cases where the expert did not agree with the crowd, 20.27% of the cases were marked as constructive by the crowd and the expert disagreed with these annotations. Here is the list of some of the prominent reasons why the expert thought the comments were non-constructive in these cases.

1. Not enough content
2. Not offering any real solutions or insights
3. Assigns blame and is insulting and disrespectful
4. Not relevant to the article
5. Sarcastic and lacks evidence
6. Intended to provoke an emotional response

[25] In CrowdFlower terminology, each annotator has a trust level based on how they perform on the gold examples, and each answer has a *confidence*, which is a normalized score of the summation of the trusts associated with annotators.

Interestingly, there were four comments where the crowd thought the comments were not constructive, but the expert disagreed. An example is shown in (7). This comment was marked as non-constructive by our annotators, but the expert thought that the claims in the comment were supported by evidence.

(7) The numbers show that Democrats stayed at home and didn't vote. She lost the support of her own party. In 2008 almost 80 million voted D and 50+ million R. That dwindled to about 65 million D and 50 million R in 2012. 2016 saw an even 50+ million to each party. Who's fault is that?

To tackle the issues our expert pointed out, we are designing an annotation experiment where we ask specific questions on important aspects of constructiveness, in addition to asking the binary question of whether the given comment is constructive or not.

3.5 SFU constructiveness and toxicity corpus

Our annotators found a few duplicate instances in our corpus. In particular, they noted some instances of equivalent comments, some with the text of the parent comment included in them and others without this text, as shown in Example 8.

(8) a. Wow, Seiko's cost that much? For half that you can buy a 75 year old Bulova, which will run another 75 years.

b. (In reply to:PRINCIPLE: The purpose of technology is to serve humankind, not the other way around. APPLICATION: Instead of buying a $500 Apple Watch, buy a $500 Seiko and enjoy a real watch.– Excimer) Wow, Seiko's cost that much? For half that you can buy a 75 year old Bulova, which will run another 75 years.

We carefully curated the crowd annotated corpus and removed the instances containing the text of the parent comment. Note that our comments corpus and comment-threads corpus do not have such instances, as our preprocessing takes care of them. Since constructiveness and toxicity annotations were carried out before we cleaned SOCC, we carried out the duplicate removal process after the annotation process for this annotated corpus. Our curated corpus contains 1,043 comments, which are organized into a CSV file. Below we are describing the most relevant fields from this CSV. For information about the other fields and the corpus download link, please refer to out project GitHub page.[26]

- **article_id:** The article identifier of the article, which can be used to link the comment to the article corpus.
- **comment_counter:** The comment counter, which can be used to link the comment to the comments corpus or the comment-threads corpus.
- **comment_text:** The comment text which was shown to the annotators
- **is_constructive:** Crowd annotation for constructiveness (yes/no/not sure)

[26] https://github.com/sfu-discourse-lab/SOCC#constructiveness

- **toxicity_level:** The crowd annotation for toxicity level
- **expert_is_constructive:** The expert annotation for constructiveness
- **expert_toxicity_level:** The expert annotation ob toxicity level
- **expert_comments:** The expert comments on crowd annotation

4 Negation annotations on SOCC

The automatic identification and detection of negation has been a significant topic in biomedical research where its usage is extensive (Aronow et al., 1999; Chapman et al., 2013; Mutalik et al., 2001). In the biomedical sphere, negation is often discussed in tandem with the concept of speculation (Vincze et al., 2008; Cruz Díaz et al., 2012). Negation and speculation are similar in their tendency to influence a sentence with the projection of a scope. Automatic methods for negation detection have also attracted much attention in the domain of sentiment analysis. Previous research has focused on classifying the scope of negation in relation to opinion in on-line reviews of products and film (Mittal et al., 2013; Dadvar et al., 2011; Councill et al., 2010). While earlier studies on negation have suggested some correlation between negation and negative affect (Potts, 2010), it is not always the case that negation indicates negativity (Blanco and Moldovan, 2014).

The primary intention of this research and annotation is to examine the relationship between negation, negativity, and Appraisal. We expand upon previous research in the field by considering the interaction between toxicity, constructiveness, and negation in online comments. For this purpose, we create a unique and practical dataset containing comments annotated for negation. In doing so, we devise a formal strategy for annotating the focus of negation, a topic which has proven quite challenging in the past.

4.1 Definitions

Before discussing the annotations, there are some essential terms which must be clearly defined and understood. The three core concepts customarily employed in negation analysis and annotation efforts are as follows.

Keyword. A keyword, or a negator, is the element which triggers the negation. Keywords are a closed class of words, such as *no* or *not*, which project a scope and specify a focus. A keyword, as the name suggests, tends to be at most one or two tokens. They are generally unambiguous and easily identified. Keywords include the negator *not*, whether by itself or attached to a verb (*the solution doesn't lie in decolonizing*), other words such as *never*, and negative polarity items such as *nothing, nobody* or *nowhere* (*nobody is suggesting that*) (Huddleston and Pullum, 2002; Horn, 1989).

Scope. The scope of negation has been comprehensively researched in previous literature (Vincze et al., 2008; Jiménez-Zafra et al., 2017). It is defined as the part of the

meaning that is being negated. In order to ensure that all elements which may plausibly fall within the scope are included, a maximal approach is implemented. Unlike the keyword and the focus, scope spans over the largest possible syntactic unit.

Focus. Focus is the most divisive in its definition. One definition is that the focus is the element which is intended to be false and is crucial for the interpretation of the negation (Vincze et al., 2008). Another definition of focus assumes that focus correlates with the answer to a wh-question (Rooth, 1985). A third definition interprets focus based on a *question under discussion*, or QUD (Anand and Martell, 2012). For this project, we adopt the definition in Blanco and Moldovan (2014). Focus is determined as the part of the meaning most explicitly negated. To keep focus concise, a minimal approach is adopted.

Example (9) shows a sample annotation. The keyword is *cannot*, with the negation attached to modal verb. The scope is the entire VP, and the focus, i.e., the item most directly negated, is *believe*.

(9) I (cannot)$_{keyword}$ ((believe)$_{focus}$ that one of the suicide bombers was deported back to Belgium.)$_{scope}$

4.2 The annotation process

We annotated a total of 1,121 comments for negation, using Webanno, (de Castilho et al., 2016). After duplicate removal, the final annotations contain 1,043 comments (see Section 3.5). These comments were annotated by up to two individuals, evaluated for agreement, then curated to acquire the most precise annotation for each comment. Specific guidelines were developed to assist the annotators throughout the annotation process, and to ensure that annotations are standardized. These guidelines were developed based on previous annotation projects, namely Vincze et al. (2008); Jiménez-Zafra et al. (2017); Martín Valdivia et al. (2017), and improved upon to provide a thorough analysis of negation. Several versions of the guidelines were tested for efficiency before deciding on the final version, which is publicly available through the GitHub page for the corpus.[27]

There are four annotation labels associated with the negation project in WebAnno, these are *focus*, *scope*, *xscope*, and *neg* for keyword. Unlike previous annotation strategies which have included the keyword in the scope of the negation, the annotation system applied to this dataset uniquely excludes the keyword in the scope. This decision is motivated by the definition of scope, that it is the part of the meaning being negated. A keyword cannot negate itself, therefore it is more logical to conceive of the keyword as a flashlight which projects light, the scope, on other elements in the sentence.

In cases of elision or question and response, a special annotation label, *xscope*, has been created to indicate the implied content of an inexplicit scope. Rather than considering instances of elision as negation without scope, it is assumed that the scope has simply been omitted and that it can be extrapolated from previous elements in the

[27] https://github.com/sfu-discourse-lab/SOCC/tree/master/guidelines

discourse. To indicate the context derived scope of a negation involving omission, *xscope* is used, as shown in Example (10). Aside from the two previously discussed deviations, the rest of the annotation approach follows prior conventions.

(10) He said he would (change the world)$_{xscope}$, but he obviously (won't)$_{keyword}$

4.3 Overview of negation in the corpus

After processing the corpus to remove duplicates, 1,043 comments remained for preliminary analysis. Counting the spans for each label revealed that there were 1,397 instances of *keyword*, 1,349 instances of *scope*, 34 instances of *xscope*, and 1,480 instances of *focus*. There are fewer instances of scope than of focus because, in some cases, both focus and scope are the same element, and in some cases the keyword has no scope.

Generally, the annotation process was uncomplicated. Most cases of negation followed a familiar pattern, and contained an easily identifiable focus. Since determining the focus of annotation is heavily dependent on context, the general lack of contextual information in comments on-line is problematic. Although annotators read the article before annotating, they did not always have the full context of the discussion thread. This issue is particularly exacerbated with sentences of considerable length. As a sentence containing negation became longer, the more challenging it was to confidently ascertain the most viable focus. In Example (11), there is a particularly long clause which has multiple candidates for focus. Given the limited context, it is quite difficult to make a decision as any of the noun clauses within the scope are acceptable options. Ultimately, it was decided that the final element in the scope should bear the focus.

(11) So why (don't)$_{keyword}$ (moderate Muslims head to places like Iraq and Syria and to other countries where Muslim extremists and terrorists exist to eradicate (those Muslim radicals.)$_{focus}$)$_{scope}$

Although we achieved respectable results from these annotations, the process was not without challenges. Ungrammaticality and colloquialism in on-line comments often introduce frustration, especially concerning the span of scope and focus. It can be quite arduous, when faced with incoherent thoughts or multiple run-on sentences, to determine where a span should reasonably end. Throughout the annotation process, there were no obvious trends with the usage of negation. Given that negation is fundamentally a logical operator, its usage seems to be more basic. Either the writer employed negation, or did not. There were a great many comments within the corpus which included no instances of negation at all, whereas others included many.

4.4 Agreement

Two annotators performed the annotation. One was in charge of overseeing the process and training the research assistant. The research assistant annotated the entire corpus. The senior annotator then curated and solved any disagreements. To calculate agreement, 50 comments from the beginning of the annotation process and 50 comments from the conclusion of the annotation process were compared. Agreement between the annotators was calculated individually based on the label and the span for the keyword, scope, and focus. The label represents the tag used in the annotation, either keyword, scope, or focus. The annotations for scope include both scope and xscope, as xscope is considered a subtype of scope. For span, agreement was considered only for the instances in which the label is agreed upon for an annotation, and there is some overlap in the span of that annotation. Agreement was calculated using percentage agreement for nominal data, with annotations regarded as either agreeing or disagreeing. A percentage indicating agreement was measured for both label and span, then combined to yield an average agreement for the tag. We did not employ more complex measures of agreement (Cohen's kappa or Krippendorff's alpha) because most of our labels are binary (the item is a keyword or not), and percentage agreement provides enough information for such cases.

The first 50 comments included a total of 51 instances of negation. Within the 43 sentences, there were 51 cases of keyword, 50 cases of scope and xscope, and 59 cases of focus. Out of the annotations for keyword, a total of 50 were considered as agreeing. All of the annotations for keyword which agreed on the label also agreed on the span. In the annotations for scope, all 50 spans agreed on the label, and 49 of the 50 agreed on the span. The selection of focus showed more variability, a total of 59 elements were annotated for focus, but the label was agreed upon for only 46. Of the 46 foci, 42 of them agreed on the span. The percentage based results for the first 50 comments, as well as the average percentage of agreement, are shown in Table 4.

The final 50 comments examined for negation included 68 sentences with instances of negation. Within these 68 sentences, there were 83 instances of keyword. Of the 83 instances, in 77 of the cases the two annotators agreed on both label and span. For the scope, 70 instances were found, with 69 in agreement. Out of the 69 scope annotations, 62 agreed on the span. The focus for the final 50 annotations, similar to the first 50, yielded a much lower percentage of agreement. A total of 93 instances of focus were marked, but only 59 were in agreement. The foci found agreeing on label agreed on span in only 52 of the 59 cases.

From the results of the agreement calculations, the intuition that the keyword is the most easily identifiable feature is accurate. Whenever the keyword is in agreement, there is 100% agreement on its span. Upon reviewing the individual files included in the analysis, it also seems highly likely that supposed disagreement displayed with the keyword is better attributed to an element being overlooked by either of the annotators. The results for scope also show a very high percentage of agreement, suggesting that the scope is also fairly simple to determine. The results for focus, while noticeably lower than scope and keyword, are not surprising. Given the context dependent nature of focus, disagreement is expected. It is often to the discretion of the annotator to designate the most appropriate candidate, particularly in

Table 4 Agreement for the first and last set of 50 comments

		Keyword	Scope	Focus
First 50	Label Agreement	98.0%	98.0%	79.3%
	Span Agreement	100%	98.0%	91.3%
	Average Agreement	99.0%	98.0%	85.3%
Last 50	Label Agreement	92.8%	98.6%	63.4%
	Span Agreement	100%	89.9%	88.1%
	Average Agreement	96.4%	94.2%	75.8%

instances where there are multiple candidates for focus. Recognizing the reliance on context, the agreement from these annotations is considerably positive. This suggests that the guidelines for annotation may have been beneficial for determining focus. One surprising effect was that agreement went down from the first 50 comments to the last 50, which were carried out over a period of three months. At the beginning of the annotation process, the lead annotator was working more closely with the other annotator to provide instruction as the novice became more comfortable with the procedure. This likely resulted in higher agreement in between the two annotators as they would be in direct contact and influence each other's opinions. Another reason for the loss in accuracy was perhaps haste towards the end of the annotation process. During curation, we observed that one of the annotators had overlooked keywords and their associated scope and focus. This was corrected in the curation process.

5 Appraisal annotations on SOCC

Following the framework of Martin and White (2005), we annotated Attitude[28], as well as Graduation of Attitude, using WebAnno (de Castilho et al., 2016). The Appraisal framework aims to capture the linguistic resources used to convey evaluation. In Appraisal, linguistic choices are characterized as systems of choices. The first of those choices is in the Attitude system, which in turn classifies evaluation as Affect (expressions of emotion by the speaker), Judgment (evaluation of other people's abilities and ethics), or Appreciation (aesthetic evaluations of an object). Attitude can be intensified or downtoned through resources classified under Graduation. A third system, Engagement, characterizes resources to engage or disengage with the evaluation being expressed. In our annotations, we labelled linguistic expressions that convey Attitude and Graduation. Figure 4 contains a brief representation of the choices in Appraisal.

The annotations consist of two layers: Attitude and Graduation. The Attitude layer labels spans by category (Affect, Appreciation or Judgment), as well as polarity (Positive, Negative or Neutral). The Graduation layer categorizes a span as

[28] We follow conventions in the Appraisal Framework, and in Systemic Functional Linguistics, to capitalize the first letter of the systems being described (Attitude, Appreciation, Graduation, etc.).

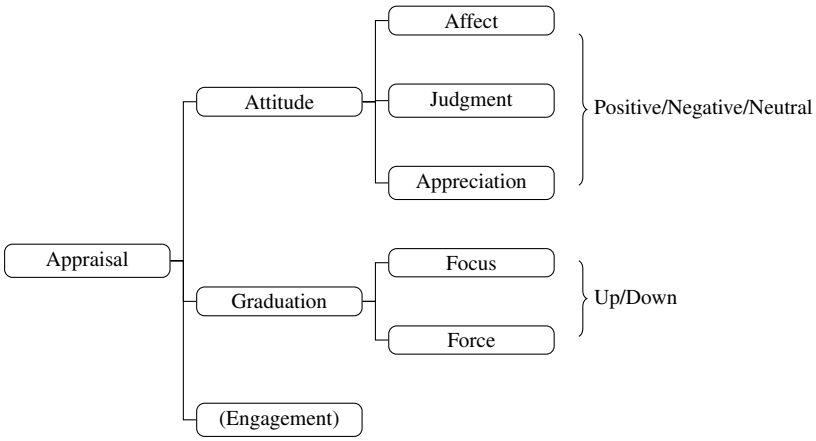

Fig. 4 The Appraisal system

either Force or Focus, with a polarity of up or down, i.e., an intensifying or downtoning effect. Further details on how labels were used can be found in the Appraisal Guidelines document.[29]

The Appraisal annotations provide a sophisticated level of analysis of evaluative language, based on a solid theory of how evaluation is conveyed through language. The annotations can prove useful in studies of the interplay of evaluative language, negation, and constructiveness and toxicity.

5.1 The annotation process

Using WebAnno (de Castilho et al., 2016), a total of 1,121 comments were annotated with the aforementioned labels according to the Annotation Guidelines (after duplicate removal, 1,043 comments remain; see Section 3.5). As in the negation annotations, the comments were annotated by up to two individuals, evaluated for agreement, and curated to increase the accuracy of annotation. We began with guidelines previously used on a corpus of film reviews (Taboada et al., 2014) and further developed them by iteratively testing and discussing those guidelines on our corpus. Once guidelines had been established, a research assistant annotated the original 1,121 comments under the supervision of one of the researchers, who was responsible for curating the annotations.

Over the course of the annotation process, a few significant departures and clarifications needed to be made to the original guidelines (Taboada et al., 2014). Originally, adjectives coordinated by a conjunction (12) were distinguished from those coordinated by a comma (13). However, commenters did not seem to make a meaningful distinction between these two structures, which may be influenced by the difference in register between movie reviews and online comments on opinion articles.

(12) Hillary Clinton's deeply flawed and frankly troubling history [...]

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

(13) Clinton is a power hungry, egotistical person [...]

Additionally, due to the content of the articles being commented on, many commenters expressed various attitudes towards organizations such as governments and corporations. At times it can be ambiguous whether these attitudes represent Judgment, as they do in (14), or Appreciation, as they do in (15) and (16). In the case of Judgment, the target of the commenter's attitude is the members that make up an organization, while in the case of Appreciation, the target is either a country as a location, or an (e.g., sacred or dangerous) institution regardless of its members.

(14) The brutal Chinese Communist Party has murdered over fifty million of its own people since 1949, since 1999 it has been attempting the blood-thirsty genocide of the tens of millions of innocent Falun Gong who live in Mainland China.

(15) Our parliament is our secular church [...]

(16) [...] Mexico were over 60,000 people have been tortured, killed , decapitated over the last 10-20 years.

Another issue in writing the guidelines was determining how long a span should be. In brief, we attempted to annotate spans that were as short as possible while containing all relevant words with attitudinal content. Many of the guidelines were designed to address cases where such a judgment is difficult to make, but as it is unfeasible to provide a guideline to cover every circumstance, there was inevitably some disagreement on the exact limits of a span.

5.2 Preliminary analysis

Over all 1,043 comments, 6,623 instances of Attitude and 771 instances of Graduation were found. Table 5 shows the number of spans that were annotated with each category and polarity of Attitude and Graduation.

Several very strong trends emerged from the data. Expression of Affect was quite rare in the corpus, comprising a mere 3.4% of the spans containing some kind of Attitude. As well, the Attitude expressed in these comments was overwhelmingly negative, as in nearly 75% of Attitude spans. Explicitly neutral positioning is vanishingly rare, accounting for barely over one percent of the Attitude found in this corpus. Out of the 771 spans in which Graduation was found, Force used for up-scaling was prevalent. 85% of Graduation was an instance of Force, and 91% sharpened or scaled up the relevant Attitude.

In future work, we intend to further analyze these patterns in terms of the genre of online comments and how negation affects Attitude.

5.3 Inter-annotator agreement

To determine agreement between the two annotators, two agreement studies were done, one once the research assistant indicated reasonable familiarity with the guidelines, and one approximately one and a half months later, at the end of the process.

Table 5 Instances of Appraisal

		Frequency	Percentage		Frequency	Percentage
Attitude	Appreciation	3,577	54.0%	Negative	4,870	73.6%
	Judgment	2,820	42.6%	Positive	1,679	25.4%
	Affect	226	3.4%	Neutral	71	1.1%
Graduation	Force	600	71.5%	Up	637	75.9%
	Focus	239	28.5%	Down	202	24.1%

For each study, the research assistant and her supervisor annotated 50 comments in parallel.

Agreement was calculated separately for each layer (Attitude and Graduation) and each field within the annotation (category and polarity). All spans annotated with the same beginning, end, and label were counted as agreeing. Spans which were labeled by one annotator but not the other were counted as disagreeing. When each annotator agreed on the start and end of a span but assigned it a different label, this was also counted as a disagreement. For spans which at least partially overlapped, agreement was decided on a case-by-case basis; if annotators were labeling similar phrases with similar labels, this was counted as an agreement, while using different labels or labeling different phrases were counted as disagreements. As with negation, we calculated percentage agreement.

Agreement study 1. The 50 comments in the first agreement study included 2,688 words in 139 sentences. The results are summarized in Table 6.

Table 6 Results of agreement study

		Attitude	Graduation
First 50	Category Agreement	81%	35%
	Polarity Agreement	89%	46%
	Average	85%	41%
Last 50	Category Agreement	81%	42%
	Polarity Agreement	87%	48%
	Average	84%	45%

There were many causes of disagreement in the annotations for these comments. Sometimes, the span of attitudinal content was unclear, as in (17), where the annotators disagreed as to whether *publicly call me a liar* or simply *call me a liar* should be annotated as negative Judgment. In this case, it was decided that *publicly* should be included in the span as publicly calling someone a liar could is perhaps more face-threatening and thus more reprehensible than doing so in private.

(17) I CANNOT allow you to publicly call me a liar!

Another issue causing disagreement was different interpretations of the appropriate label for an attitude-bearing span, as in (18) and (19). In both cases, the annotators

disagreed as to whether the span in question was Appreciation or Judgment. Example (18) could be seen as Appreciation on the grounds that a continent with a dictator is institutionally bad, or as Judgment on the grounds that Merkel, as a dictator, is reprehensible for her alleged dictatorship. Ultimately the interpretation as Judgment was deemed superior. Example (19) could be coded as Appreciation if it is interpreted more as a criticism of the interlocutor's argument (along the lines of saying: *If the problem isn't Islam, then why aren't there radical Christian terrorists?*), or as Judgment due to its sarcastic tone and the implicit accusation of intellectual dishonesty or incompetence. This span was also deemed negative Judgment, especially due to its format as a rhetorical question, which can be seen elsewhere in the corpus as frequently accompanying and even conveying insults.

(18) The capital of Europe is Berlin and Merkel is the dictator.

(19) There are radical Christians causing world terror?

In addition to these errors, some spans were overlooked by one annotator or the other, likely due to the large volume of Attitude contained in a comment. Some disagreements are also attributable to errors based on unfamiliarity with the guidelines, especially with regard to Graduation, which the research assistant did not have as much time to become familiar with. There are also several errors based on combining spans conjoined by commas or the word *and,* as this guideline was not established until after agreement was calculated.

Agreement study 2. The 50 comments in the second agreement study included 3,852 words in 207 sentences. The results of this study are also summarized in Table 6.

In the second set of 50 comments, many of the same issues as those in the first set reappeared. Graduation category agreement improved moderately as the research assistant became more familiar with finding and labeling graduation in comments. However, as the barely-improved agreement on Graduation polarity indicates, a new issue arose in sentences such as (20) and (21). In both of these cases, there is downscaling of a positive item (i.e. *substance, flexible,* and *able to handle change*). But in fact the negative Appreciation conveyed in those sentences is increased by this, as something which is *less positive* is effectively *more negative*. This was not made clear in the guidelines, hence the error.

(20) Apple stuff consist of 95% marketing nonsense and 5% substance

(21) [...] I suspect a narrow view is less flexible, less able to handle change.

Another cause of disagreement is ambiguity in comments. For example, in (22), it is not obvious where the span(s) should begin or end, or whether the first part of the sentence is criticizing forced obsolescence (which would make it Appreciation) or Apple, the corrupt purveyor of products which it forcibly obsolesces. Similarly, in (23) it is not clear whether the commenter is sarcastically targeting the hypothetical iMplant and the institutional flaws of Apple, or the corrupt decision makers of Apple who supposedly want to invade people's privacy by creating such a device.

(22) Will the $13,000 gold model be obsolete when the next iteration is released in 3 months time [...]?

(23) Bahhh! I' m waiting for the iMplant.™

Both the negation and Appraisal annotations were done with the WebAnno interface. On our GitHub page, We have download links for WebAnno projects that can be explored with WebAnno, or as text files with annotations. The GitHub project also contains instructions for WebAnno import.[30]

6 Conclusion

We have described The SFU Opinion and Comments Corpus (SOCC), an excellent resource for exploring opinion news articles, online news comments, and their relationship. A number of research questions related to journalism, online discourse, the dialogic structure of online comments, and evaluative language can be explored from such a resource.

Our corpus is composed of two kinds of sub-corpora: raw and annotated corpora. Our raw corpus comprises 10,339 opinion articles published in the Canadian newspaper *The Globe and Mail* in the five-year period between 2012 and 2016, along with 663,173 comments and 303,665 comment threads in response to these articles. The annotated corpora comprises a subset containing 1,043 comments from our raw corpus, enriched with constructiveness, negation, and Appraisal annotations.

While carrying out annotations for constructiveness and toxicity, we learned that constructiveness is an interplay between a variety of other phenomena of interest in computational linguistics, such as argumentation, relevance of the comment to the article, and the tone of the comment. We believe that we may obtain better quality annotations if we ask specific questions leading to constructiveness (e.g., whether the comment is relevant to the article or whether the claims made in the article are supported by evidence), instead of asking a single binary question, and in our current work, we are pursuing this research direction.

With respect to the negation annotation, we developed extensive and detailed guidelines for the annotation of negative keywords, scope and focus. We used the guidelines to annotate the chosen subset of the comments corpus, producing a completely annotated corpus for negation, including its scope and focus. This corpus has been curated to provide the most accurate annotations according to the guidelines. We have also achieved reasonable results for agreement between annotators on these annotations.

With the Appraisal annotations we have shown that it is possible to achieve favourable rates of agreement using the system, though agreement requires a high degree of familiarity with the guidelines and can still be hindered by ambiguity in comments. Aside from the guidelines, we have provided a novel, extensively annotated corpus of online comments that will be used to investigate the relationship between negation and Appraisal, yet which has the potential for other avenues of research as well.

[30] `https://github.com/sfu-discourse-lab/SOCC#appraisal`

The annotations were carefully curated, and interannotator agreement suggests that they are reliable and replicable. In the article, we have provided extensive detail about the corpus, because we think data collection and curation is an important process, which should be well documented and accountable. Our corpus is freely available for non-commercial use.[31]

Acknowledgements This work was supported by the Social Sciences and Humanities Research Council of Canada (Insight Grant 435-2014-0171). We thank all the members of the Discourse Processing Lab at Simon Fraser University for their help in testing annotation questions, and especially Erin Jastrzebski and Sarah Mulhall for annotating the data.

[31] Full description of the corpus and structure: `https://github.com/sfu-discourse-lab/SOCC` and direct link to the data: `https://researchdata.sfu.ca/islandora/object/islandora:9109`

References

Pranav Anand and Craig Martell. Annotating the focus of negation in terms of questions under discussion. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 65–69, Jeju, Korea, 2012.

David B Aronow, Feng Fangfang, and W Bruce Croft. Ad hoc classification of radiology reports. *Journal of the American Medical Informatics Association*, 6(5): 393–411, 1999.

Emma Barker and Robert Gaizauskas. Summarizing multi-party argumentative conversations in reader comment on news. In *Proceedings of ACL 2016*, pages 12–20, Berlin, 2016.

Eduardo Blanco and Dan Moldovan. Retrieving implicit positive meaning from negated statements. *Natural Language Engineering*, 20(4):501–535, 2014.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. Detecting online harassment in social networks. In *Proceedings of the 35th International Conference on Information Systems*, pages 1–14, Auckland, 2014.

Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. Extending the negex lexicon for multiple languages. *Studies in Health Technology and Informatics*, 192:677, 2013.

Isaac G Councill, Ryan McDonald, and Leonid Velikovich. What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala, Sweden, 2010.

Noa P Cruz Díaz, Manuel J Maña López, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the Association for Information Science and Technology*, 63(7):1398–1410, 2012.

Maral Dadvar, Claudia Hauff, and Franciska MG De Jong. Scope of negation detection in sentiment analysis. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop, DIR 2011*, pages 16–19, Twente, 2011.

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

Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, 2016.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30, Florence, Italy, 2015. ACM.

Laurence Horn. *A Natural History of Negation*. University of Chicago Press, Chicago, 1989.

Rodney Huddleston and Geoffrey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, 2002.

Salud María Jiménez-Zafra, Mariona Taulé, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and M Antónia Martí. Sfu reviewsp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation*, pages 1–37, 2017.

Varada Kolhatkar and Maite Taboada. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, 2017a.

Varada Kolhatkar and Maite Taboada. Using New York Times Picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105, Copenhagen, 2017b.

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

James R. Martin and Peter R. R. White. *The Language of Evaluation*. Palgrave, New York, 2005.

Maite Martín Valdivia, Salud María Jiménez Zafra, and Noa Cruz Díaz. Proceedings of the Workshop, Taller de NEGación en ESpañol NEGES-2017. 2017. URL `http://sepln2017.um.es/neges.html`.

Namita Mittal, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania. Discourse based sentiment analysis for Hindi reviews. In Pradipta Maji, Ashish Ghosh, M. Narasimha Murty, Kuntal Ghosh, and Sankar K. Pal, editors, *Pattern Recognition and Machine Intelligence*, pages 720–725, Berlin, 2013. Springer.

Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *Journal of the American Medical Informatics Association*, 8(6):598–609, 2001.

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

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Conversational markers of constructive discussions. In *Proceedings of NAACL-HLT 2016*, pages 568–578, San Diego, 2016.

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

Stephanie Pieschl, Christina Kuhlmann, and Torsten Porsch. Beware of publicity! Perceived distress of negative cyber incidents and implications for defining cyberbullying. *Journal of School Violence*, 14(1):111–132, 2015.

Christopher Potts. On the negativity of negation. In *Semantics and Linguistic Theory*, volume 20, pages 636–659, 2010.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using Machine Learning to detect cyberbullying. In *Proceedings of the 10th International Conference on Machine Learning and Applications*, pages 241–244, Honolulu, HI, 2011.

Mats Rooth. *Association with Focus*. PhD thesis, Department of Linguistics, University of Massachusetts, Amherst, 1985.

Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. Using crowdsourcing to improve profanity detection. In *Proceedings of the AAAI Spring Symposium: Wisdom of the Crowd*, pages 69–74, Stanford, 2012.

Brian H. Spitzberg and Jean Mark Gawron. Toward online linguistic surveillance of threatening messages. *The Journal of Digital Forensics, Security and Law*, 11(3): 43, 2016.

Maite Taboada, Marta Carretero, and Jennifer Hinnell. Loving and hating the movies in English, German and Spanish. *Languages in Contrast*, 14(1):127–161, 2014.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):S9, 2008.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Cursing in English on Twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 415–425, Baltimore, MD, 2014.

William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, 2012.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of NAACL-HLT*, pages 88–93, San Diego, CA, 2016.

Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A corpus study of evaluative and speculative language. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Aalborg, Denmark, 2001.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on the World Wide Web*, pages 1391–1399, Perth, 2017.

A Factiva article with metadata

Table 7: An example Factiva article with metadata.

SE	Editorial
HD	Not just words
WC	352 words
PD	30 December 2016
SN	The Globe and Mail
SC	GLOB
ED	Ontario
PG	A12
LA	English
CY	
LP	There are 60 indigenous languages in Canada, more or less; the most spoken are Inuktitut and the related Cree and Ojibway. But while they are many, they risk disappearing. What is needed is a national effort to preserve them.
TD	Senator Serge Joyal has been heroically trying to get a private member's bill through the Senate to help revitalize indigenous languages. But that can't possibly work, for the simple reason that a Senate bill can't force the government to spend any money. More promising is Prime Minister Justin Trudeau's vow to introduce a bill to re-energize indigenous languages. He offered no details when he made the announcement, but it's still progress. For many, the urgent interest in this issue is about preserving dying languages, which are understood mostly by elderly people. In an era of reconciliation with indigenous peoples, putting money into the preservation of native languages would be a concrete gesture that could produce equally concrete benefits. Section 13 of the United Nations Declaration on the Rights of Indigenous Peoples, if adopted in a measured way, can be a helpful guide. It guarantees the right to the preservation of native languages and literature, and the right to retain native place names. It also guarantees the right to a trial in one's native language, which would be too difficult to accommodate in all cases. The teaching of indigenous languages should be a priority on reserves, especially the most remote ones. Young people there may be most in need of tools to help overcome their alienation. But there can also be excitement and value in learning a language and culture that you don't have any hereditary link to, something non-native Canadians might be interested in. The University of Winnipeg's compulsory policy that all its students take at least one course in an aboriginal subject goes too far. But the intention behind that requirement is good. These are some of the principles on which the government can form new policies on native languages – and not just for people with indigenous ancestry.
NS	nedi : Editorials | ncat : Content Types
RE	cana : Canada | namz : North America
PUB	The Globe and Mail Inc.
AN	Document GLOB000020161230eccu0000c