Reflect: Stimulating Healthier Online Debate Using Algorithms

Alex Koen

University of British Colubmia - WRDS 150

Abstract

This paper introduces Reflect, a proof-of-concept software tool that provides users of online discussion forums with feedback on the potential impact of their comments *before* they are posted. Given a response to a prompt, the software identifies the markers of politeness and impoliteness in the text and assigns the interaction a probability score corresponding to the likelihood that future comments will contain personal attacks. This methodology is based on deindividuation theory, which postulates that increased self-awareness leads to more civil interaction. While other software (e.g. Jigsaw's Perspective API) can measure the toxicity of a comment in real-time, Reflect uniquely considers the entire conversation in context and explains to users the exact features of their language that contribute to its prediction. It was found that Reflect provides accurate analysis for the majority of prompt-response pairs, although further research will be necessary to determine whether it is currently feasible to implement the software on social media platforms.

Keywords: Conversational Forecasting, ConvoKit, Deindividuation, PerspectiveAPI

Reflect: Stimulating Healthier Online Debate Using Algorithms

Description

In “Public & Its Problems,” Dewey (1954) argued that “the improvement of the methods and conditions of debate, discussion, and persuasion [is] *the* problem of the public” (p. 208). Nowadays, with the popularization of social media platforms and their quasi-monopoly of public discourse, it is more important than ever to promote productive conversations on these platforms, which are well-known to incite toxicity (Gheitasy et al., 2015). Consequently, in the past few years there has been a vested interest in developing tools that can analyze online comments for markers of incivility and prejudice. For example, in 2019, researchers developed a novel recurrent neural-network-based algorithm that can detect biased statements in a given text with a precision of 91.7%, outperforming previous feature-based techniques by over 30% (Hube & Fetahu, 2019). However, until recently, research has focused almost exclusively on detecting negative markers in comments *after* they have been posted and for the purposes of moderation, with two notable exceptions. The first, an application programming interface named Perspective developed by Jigsaw, has gained popularity for its ability to provide users with real-time feedback on the perceived toxicity of their comments, yet it analyzes only single comments and does not consider the discourse that prompted them (Hosseini et al., 2017; Jigsaw, n.d.). The other, the Cornell Conversational Analysis Toolkit (ConvoKit), provides algorithms to analyze not only a conversation’s text, but also the social dynamics that influence it. (Chang et al., 2020). However, until now their algorithms have not been used to provide users with real-time feedback on the character of their comments.

Therefore, the purpose of this study was to build an open-source tool that synthesizes the paradigms of both Perspective and ConvoKit to provide users with insight into the perceived incivility of their comments in context of the surrounding conversation. Leveraging two preexisting algorithms provided by ConvoKit, Reflect notifies users of the linguistic markers of both politeness and impoliteness found in their comments along with the calculated probability that their comments will lead to personal attacks, all in real-time. It was hypothesized that by deploying this technology, users will be encouraged to stop, think, and even consider rephrasing their messages before sharing them with others, leading to healthier debate. It is well understood that when users are forced to confront their own behaviour, they are less likely to transgress; this is known as deindividuation theory (Beaman et al., 1979; Diener & Wallbom, 1976). This is novel in that users are not censored, but must merely confront the content of their messages before pressing send. To test this hypothesis, the tool was qualitatively analyzed against a publicly-available corpus of discussions between editors on Wikipedia—all of which started out civil but devolved into harassment and toxicity—to determine whether it shows promise given the maturity of the underlying technology (Chang et al., 2020). In addition, the ethical implications of this technology will be discussed given the views of several scholars that, in many cases, incivility plays a key role in public discourse (Papacharissi, 2004; Schudson, 1997).

Methods

The software was built using ConvoKit, a "unified framework for representing and manipulating conversational datasets" (p. 1), recently developed by researchers at Cornell University (Chang et al., 2020). ConvoKit is the first natural language processing framework targeted at the analysis of conversations and provides tools to analyze not only the statements that make up conversations (which the platform refers to as Utterances), but equally the relationship between these statements and the metadata of both the involved speakers and the conversation itself. (Chang et al., 2020). Specifically, it uses two of the included algorithms: an unsupervised neural network that extracts linguistic markers of politeness and impoliteness from text, and a forecaster that predicts the likelihood that a given conversation will result in personal attacks given an initial interaction (Chang & Danescu-Niculescu-Mizil, 2019; Zhang et al., 2018).

Politeness feature extraction

The first model, described by Zhang et al. (2018), was trained to identify a set of linguistic markers of politeness first proposed by Brown et al. (1987). These markers include gratitude, greetings, and hedging. Markers of impoliteness were also considered, including direct questions and the usage of second-person pronouns. Table 1 lists the full set of strategies considered.

Table 1
Politeness Strategies Analyzed

Strategy	**Example**
Direct question	Why did you do that?
2nd person start	Your opinion does not consider...
Please start	Please consider the implications...
2nd person	It's not clear if you...
1st person start	I believe that...
Hedge	It seems like...
Gratitude	Thanks for your help.
Greetings	How are you doing?

Conversational derailment forecasting

The second model, the Conversational Recurrent Architecture for ForecasTing (CRAFT), consists of an unsupervised neural network trained on a large dataset that analyzes conversational dynamics and a separate supervised model which then forecasts the outcome of a conversation (Chang & Danescu-Niculescu-Mizil, 2019). Specifically, it is trained to predict the likelihood that a conversation will result in personal attacks based only on the initial encounter, and was shown by Chang and Danescu-Niculescu-Mizil (2019) to do so with an accuracy of 66.5%, compared to the human accuracy of 72% (Zhang et al., 2018).

Dataset

These two algorithms were both trained on the *Conversations Gone Awry* (CGA) dataset, a curated database of 1,270 conversations between editors on Wikipedia with an average length of 4.6 comments each (Zhang et al., 2018). For each conversation, a crowdworker labelled whether the initial interactions were civil; if they were not the conversation was discarded. The remaining conversations were then categorized as either maintaining their civility, or resulting in personal attacks. This initial dataset was used by Zhang et al. (2018) to train the politeness strategies algorithm. In 2019, Chang and Danescu-Niculescu-Mizil (2019) tripled the size of the dataset to include an additional 2,918 conversations as part of the training for the CRAFT algorithm.

For this study, a prototype of Reflect was built using the two aforementioned algorithms and the expanded *Conversations Gone Awry* dataset. The program presents the user with a randomly selected prompt from the CGA corpus and stores their reply in a ConvoKit Utterance object. It then tokenizes each word and sentence in the reply, labels each word with its part-of-speech, and builds a dependency tree between each of its elements. The PolitenessStrategies transformer then analyzes the additional metadata added to the Utterance by this process and annotates the reply with the identified politeness strategies. Finally, the prompt-reply Conversation object is processed by the trained CRAFT model, which annotates the reply with a prediction score corresponding to the likelihood that the conversation will derail given the provided interaction. The software returns the text of the prompt and response, the list of politeness strategies used in the response, and the calculated probability score.

Results

In order to test the effectiveness of the prototype, a sample of seven prompt-response pairs from the *Conversations Gone Awry* corpus were processed by the algorithm. These samples were chosen for diversity of results and because each interaction is self-contained. Three contrasting examples are shown in Table 2 while the full set of examples are shown in Table 3.

In the first example, shown in Table 2, it is clear that the conversation is civil and that the reply is polite, addresses the prompt, and positively contributes to the discussion. Reflect correctly identifies that the user who replied uses hedging (“I agree”), employs the first-person, and is altogether positive. As expected, the software assigns the reply a low probability score of 14.4%.

Conversely, the second example is broadly negative and contains personal attacks. The software identifies that while the response uses hedging (“I guess” and “seems that”), it directly addresses the first speaker using the second person and has negative language. Naturally, it assigns

Table 2

Selected Wikipedia Conversations Analyzed by Reflect

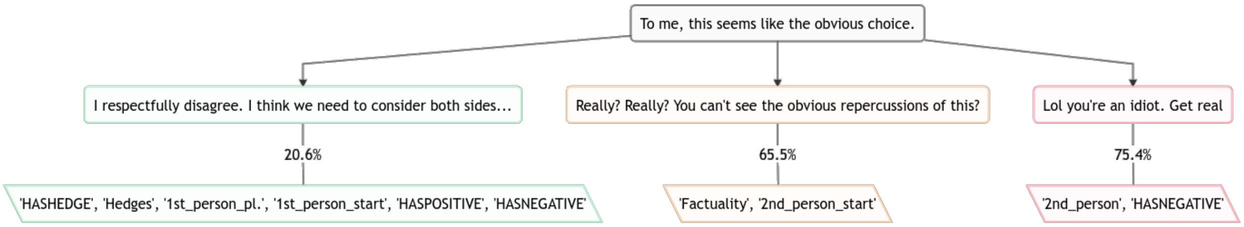
Prompt	**Reply**	**Reply politeness strategies**	**Probability of personal attacks (%)**
An individual keeps reverting my attempt to make a subsection on the recent Hersh article. I see mention of the Hersh article in the intro. In my opinion, it doesn't need to be mentioned in the intro since it is a minority report. A small section in the body is sufficient. Also, why is my addition getting reverted since the body needs to support any info placed in the lede?	I agree. A couple of paragraphs, with the main claims of the report, is more appropriate in it's own subsection.	HASHEDGE 1st_person_start HASPOSITIVE	0.144
The Brady Campaign was founded as The National Council to Control Handguns in 1974. No matter how much you wish it were founded with the word "ban" in it's title, it's simply not true. If you insist on making up a history for an organization you cleraly know nothing about, at least provide some historical proof of your spurious claim.	I guess that the Congress of the United States are a bunch of morons, and got the name wrong, huh? take a look a the [[*external_link*]]. Seems that YOU are trying to change history.	HASHEDGE Hedges 1st_person 2nd_person HASNEGATIVE	0.883
That Amazon has begun pre-ordering isn't a notable occurance, or useful to the average reader. Thanks!	They're the largest online retailer, which gives them a large amount of credibility. They do millions in pre-orders. For an article that people are going to be checking for news pertaining to a release, this needs to be included.		0.352

the comment a probability rating of 88.3%, which is likely correct given the initial interaction.

The third example is more nuanced. The first speaker respectfully states that a specific edit to a page is not notable, and should therefore be removed. The second user disagrees, but does so in a neutral manner. The politeness strategies network does not identify any markers of politeness or impoliteness. Consequently, the CRAFT model assigns the reply a probability rating of 35.2% which, again, seems reasonable given the neutral tone of the disagreement.

Figure 1 shows the algorithm's results for a series of fabricated replies to a common prompt. As the reply becomes more antagonistic from left to right, the probability score increases accordingly and Reflect identifies the positive and negative politeness strategies that contribute to the score.

Figure 1
Output of the Reflect Algorithm for Various Replies to a Common Prompt.


To me, this seems like the obvious choice.
I respectfully disagree. I think we need to consider both sides...
Really? Really? You can't see the obvious repercussions of this?
Lol you're an idiot. Get real
20.6%
65.5%
75.4%
'HASHEDGE', 'Hedges', '1st_person_pl.', '1st_person_start', 'HASPOSITIVE', 'HASNEGATIVE'
'Factuality', '2nd_person_start'
'2nd_person', 'HASNEGATIVE'

Note. The percentage corresponds to the likelihood that the comment will incite personal attacks while the lower box corresponds to the identified politeness strategies. All examples are fictitious.

Discussion

Given the above examples and those shown in Table 3, it is clear that Reflect provides accurate analysis for the majority of comments and consequently shows promise as a tool to provide users with feedback on the potential repercussions of their participation in online discussions. Figure 2 shows a mockup of a potential implementation of the software in the context of an online discussion forum. As the user types a response to a comment, the algorithm assigns it a letter grade from A to F. If the user hovers over the icon, a popup describes the likelihood that their comment will incite a personal attack. They may then choose to click the *See why* button to view a more comprehensive analysis of their comment including a detailed description of the identified politeness strategies and why they might be ill-received.

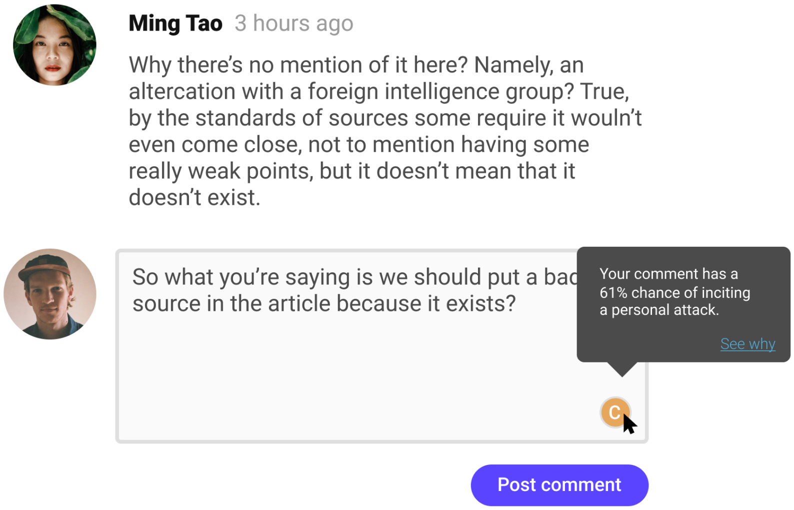
Notice that the user is never prevented from posting the comment, and that the linguistic analysis is only provided if they choose to hover over the icon. This way, there is no censorship—the tool simply increases self-awareness. It is well-known that such an increase in self-awareness decreases transgressive behaviour; the phenomenon is supported by numerous studies (Beaman et al., 1979; Diener & Wallbom, 1976). For example, Diener and Wallbom (1976) showed that students were 62% less likely to cheat on a test when they sat in front of a large mirror and listened to voice recordings of themselves (rather than someone else). It is thus theorized that Reflect will act as the proverbial mirror and tape-recorder, although further research will be required to determine its effectiveness.

Another tool similar in appearance to Reflect is Jigsaw's Perspective API, which creates a popup in a user's text field if their comment exceeds a certain threshold of perceived toxicity (e.g. "98.51% likely to be toxic. Please edit.") (Jigsaw, n.d.). However, Reflect improves upon Perspective in two notable ways. Firstly, we forecast the likelihood that a comment will provoke personal attacks given the context of the entire conversation while Perspective merely labels the toxicity of a given text (Jigsaw, n.d.). Moreover, the Reflect algorithm can explain to the user *why* their comment might derail the conversation given the extracted politeness strategies. While Perspective has made recent improvements to their algorithm by expanding their API to support other experimental models beyond toxicity including insult, profanity, identity attacks, sexual explicivity, and threats (Jigsaw, n.d.), Reflect differs in its methodology and might therefore provide more nuanced and persuasive analysis.

That being said, there are several clear limitations to the *Reflect* algorithm, the most salient of which is that the comments used to train the CRAFT algorithm were sourced from Wikipedia. Since *Reflect* is targeted at social media websites like Reddit and Twitter, whose discussion may be different in character to that of Wikipedia, the software may prove less effective for this type of

Figure 2

Mockup of a Potential Implementation of the Reflect Software



Ming Tao 3 hours ago

Why there’s no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources some require it wouln’t even come close, not to mention having some really weak points, but it doesn’t mean that it doesn’t exist.

So what you’re saying is we should put a bac source in the article because it exists?

Your comment has a 61% chance of inciting a personal attack.

Note. The use may click the *See why* button to view the politeness/impoliteness strategies that contribute to their score.

discussion. A clear avenue for further research is to build another corpus from these websites and use it to improve the software.

Reflect’s consideration of the preceding comments in its prediction may prove to be both a key feature and a limitation of the software. Like in spoken conversation, the meaning of a comment requires context, and Reflect is uniquely able to consider it. However, as a result, a comment may be scored as likely to provoke an altercation not because *it* is toxic, but because the *preceding* comments are. While this clearly requires consideration, it may allow the software to phrase its alerts such that the user is more likely to consider its advice. By framing the prompt as “your comment is likely to incite personal attacks” instead of “your comment seems toxic” it suggests that algorithm is there to protect not only others, but also the user themselves. Consequently, they may be more likely to stop and consider how their comment negatively affects everyone involved. This theory is merely speculative and further research will be required to determine whether this type of prompt incites a more positive response than that of Perspective and other competing software.

In the end, however, one further point needs consideration: should we even be building tools to limit antagonistic behaviour online? Dewey (1954) argued that the very principle of democracy relies on the improvement of the methods of societal discourse, yet some scholars (Papacharissi, 2004; Schudson, 1997) insist that civility is often counterproductive and may even detract from the quality of these discussions. This is highlighted by Schudson (1997), who believes that "democracy may sometimes require that your interlocutor does not wait politely for you to finish but shakes you by the collar and cries 'Listen! Listen for God's sake!'" (p. 308). This is evidenced during strikes, manifestations, and by prominent figures from history such as Martin Luther King and William Lloyd Garrison whose very incivility and self-righteousness led them to instigate lasting change (Schudson, 1997). Nowadays, in a time when "62% of US adults get news on social media" (Allcott & Gentzkow, 2017, p. 212) and these platforms have "pivotal roles in supporting news production and diffusion" (Lee & Ma, 2012, p. 246) it is clear that effort *must* be made to improve the effectiveness of the discourse they host. Tools like Perspective and Reflect show promise as a means to do so, but further discussion will be required to determine what unintended consequences they may have.

Conclusion

For many years, researchers have sought to develop tools to control the spread of toxicity on social media platforms. However, most of this research has focused on developing tools which detect toxicity *after the fact* or for the purpose of *moderation*. Until now, there has been little investigation into how productive conversation can be encouraged online without censoring the antagonistic language and behaviour that are often necessary to instigate substantial change. This paper introduces Reflect, a new tool that alerts users if their comments are likely to incite personal attacks and can provide them with detailed analysis into the precise linguistic markers of their language that affect this prediction. While the tool and underlying methodology are still novel and underdeveloped, they show promise as a means to improve the quality of the online discourse whose interactions, in our modern era, affect the lives of all.

References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Beaman, A. L., Klentz, B., Diener, E., & Svanum, S. (1979). Self-awareness and transgression in children: Two field studies. *Journal of Personality and Social Psychology*, *37*(10), 1835–1846. https://doi.org/10.1037/0022-3514.37.10.1835

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). ConvoKit: A Toolkit for the Analysis of Conversations. *arXiv:2005.04246 [cs]*.

Chang, J. P., & Danescu-Niculescu-Mizil, C. (2019). Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. *arXiv:1909.01362 [physics]*.

Dewey, J. (1954). Public & its problems.

Diener, E., & Wallbom, M. (1976). Effects of self-awareness on antinormative behavior. *Journal of Research in Personality*, *10*(1), 107–111. https://doi.org/10.1016/0092-6566(76)90088-X

Gheitasy, A., Abdelnour-Nocera, J., & Nardi, B. (2015). Socio-technical gaps in online collaborative consumption (OCC) an example of the etsy community. *Proceedings of the 33rd Annual International Conference on the Design of Communication*, 1–9.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs]*.

Hube, C., & Fetahu, B. (2019). Neural Based Statement Classification for Biased Language. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 195–203. https://doi.org/10.1145/3289600.3291018

Jigsaw. (n.d.). *Perspective*. Retrieved August 4, 2020, from https://www.perspectiveapi.com/

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, *28*(2), 331–339. https://doi.org/10.1016/j.chb.2011.10.002

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Schudson, M. (1997). Why conversation is not the soul of democracy. *Critical Studies in Mass Communication*, *14*(4), 297–309. https://doi.org/10.1080/15295039709367020

Zhang, J., Chang, J. P., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Thain, N., & Taraborelli, D. (2018). Conversations Gone Awry: Detecting Early Signs of Conversational Failure. *arXiv:1805.05345 [physics].*

Table 3

Full Set of Wikipedia Conversations Analyzed by the Reflect Algorithm.

Prompt	**Reply**	**Reply politeness strategies**	**Probability of personal attacks (%)**
An individual keeps reverting my attempt to make a subsection on the recent Hersh article. I see mention of the Hersh article in the intro. In my opinion, it doesn't need to be mentioned in the intro since it is a minority report. A small section in the body is sufficient. Also, why is my addition getting reverted since the body needs to support any info placed in the lede?	I agree. A couple of paragraphs, with the main claims of the report, is more appropriate in it's own subsection.	HASHEDGE 1st_person_start HASPOSITIVE	0.144

Prompt	**Reply**	**Reply politeness strategies**	**Probability of personal attacks (%)**
The Brady Campaign was founded as The National Council to Control Handguns in 1974. No matter how much you wish it were founded with the word "ban" in it's title, it's simply not true. If you insist on making up a history for an organization you cleraly know nothing about, at least provide some historical proof of your spurious claim.	I guess that the Congress of the United States are a bunch of morons, and got the name wrong, huh? take a look a the [[*external_link*]]. Seems that YOU are trying to change history..	HASHEDGE Hedges 1st_person 2nd_person HASNEGATIVE	0.883
I don't have a position on the content dispute, but it's clear that the addition of the citation is a disputed edit, which means we discuss it here, not in edit summaries. The facts that the edit warring is at a slow pace, involves three editors, and hasn't gone on very long don't change the fact that it's an edit war. So here you go, and best wishes on a consensus.	I'm not sure it's a matter of content- the thing it's cited to is fine, but we probably shouldn't be including these documents as a primary source for something if we can cite it to secondary sources. has made outside this topic.	HASHEDGE 1st_person_pl. 1st_person_start Direct_start HASPOSITIVE	0.232
one the most? ONE of the most? Are you kidding ME?	A defined statement like that would require at least 2 dozen valid wiki_link.	HASPOSITIVE	0.703
That Amazon has begun preordering isn't a notable occurance, or useful to the average reader. Thanks!	They're the largest online retailer, which gives them a large amount of credibility. They do millions in preorders. For an article that people are going to be checking for news pertaining to a release, this needs to be included.		0.352
Editing a person's image in order to add a bubble which reads "I am a gay loser" is never appropriate. The image has been tagged for speedy deletion.	You're a mean bean! Go to the *wiki_link* Baked Bean Factory and be a meanie!	2nd_person_start	0.740

Prompt	**Reply**	**Reply politeness strategies**	**Probability of personal attacks (%)**
That whole section needs as source if it's going to imply that Condit was involved. "most theories" are very weasel words. Condit should be removed completely.	Not completely. He is the reason her story was news. The article shouldn't imply that he killed her, but he certainly is relevant to the story.	HASHEDGE HASNEGATIVE	0.3262
I think a modernists template can work and be a very useful navigation tool, good idea and called for. Separate subjects - Poets, Writers, Musicians, Painters, Dada, Surrealists, Theology, Philosophy, Dancers, the modernist template that we have is rudimentary and nearly useless. I appreciate your input, and yes, rudeness and strife is not fun to deal with...	Yes, that's what I envision. Tie together the entire group. I don't have experience working with templates, but would be willing to help when I get the chance. My fingers are itching to remove the infobox from Epstein, but I'll be accused of edit warring if I do. Hopefully others will chime in. I wanted to say something last night but was trying to concentrate, so I let it go until today.	1st_person 1st_person_start Direct_start HASPOSITIVE HASNEGATIVE	0.360