COMPUTATIONAL COMMUNICATION RESEARCH 4.2 (2022) 486-527
HTTPS://DOI.ORG/10.5117/CCR2022.2.005.JAID

Talking politics: Building and validating data-driven lexica to measure political discussion quality

Kokil Jaidka
National University of Singapore
jaidka@nus.edu.sg

Abstract

Social media data offers computational social scientists the opportunity to understand how ordinary citizens engage in political activities, such as expressing their ideological stances and engaging in policy discussions. This study curates and develops discussion quality lexica from the Corpus for the Linguistic Analysis of Political Talk ONline (CLAPTON). Supervised machine learning classifiers to characterize political talk are evaluated for out-of-sample label prediction and generalizability to new contexts. The approach yields data-driven lexica, or dictionaries, that can be applied to measure the constructiveness, justification, relevance, reciprocity, empathy, and incivility of political discussions. In addition, the findings illustrate how the choices made in training such classifiers, such as the heterogeneity of the data, the feature sets used to train classifiers, and the classification approach, affect their generalizability. The article concludes by summarizing the strengths and weaknesses of applying machine learning methods to social media posts and theoretical insights into the quality and structure of online political discussions.

Keywords: deliberation, constructiveness, justification, political talk, comments, Twitter, Facebook

Studies of online political talk often consider social media platforms an ideal public sphere to study how citizens engage in politics, express their political opinions, formulate arguments, defend them and negotiate with each other.

Therefore, political social media posts are invaluable for computational social scientists to understand political attitudes and public opinion formation.

However, there is a methodological and a theoretical gap in the current body of work aimed at scalable methods to measure the quality of the political talk. First, beyond individual annotation-based studies (Muddiman, McGregor, & Stroud, 2018; Stroud, Scacco, Muddiman, & Curry, 2015), there are no expert-curated dictionaries similar to the Linguistic Inquiry and Word Count (Pennebaker, Boyd, Jordan, & Blackburn, 2015), nor Artificial Intelligence (AI)-based methods to measure the quality of online political talk beyond incivility. Scholars have released automatic language classifiers created using supervised machine learning that can predict the incivility, hate speech, and harassment in large unlabeled datasets of social media posts (e.g., Davidson, Warmsley, Macy, & Weber, 2017). However, although hate speech constitutes only a small proportion of political talk, there are no scalable, sophisticated ways to characterize it along other dimensions.

Second, AI-based methods may eschew a rich body of work in social science regarding how the quality of communication *should* be measured (Muddiman et al., 2018). As a result, they may not be particularly insightful for scholars looking to understand the facets of political talk. Rinke (2015) has identified different conditions that can enable 'reasoned dissent' (Wessler, 2008) in a media forum, such as inclusiveness, responsiveness, justification, and civility. Gastil (2008) considers political discussions to comprise both analytical and social aspects.

Addressing these drawbacks, the major methodological contribution of this study is an annotated Corpus (in English)[1] for the Linguistic Analysis of Political Talk ONline (CLAPTON) and an associated set of classifiers developed through machine learning methods trained on CLAPTON. The annotated discussion quality facets are based on the theoretical conceptualization of deliberative processes embodied in the discussion quality Index (Steenbergen, Bächtiger, Spörndli, & Steiner, 2003). The classifiers are built using an open vocabulary approach (Schwartz et al., 2013) on the content, syntactic, grammatical, discursive, and psycholinguistic features of social media posts. The methodological contribution of this study is the development of resources and methods that scale to measure discussion quality in larger samples of text. Consequently, scholars can analyze longitudinal trends in quality (Jaidka, Zhou, & Lelkes, 2019), identify heterogeneous effects between populations and platforms (Halpern & Gibbs, 2013), and better estimate the effects of online discussions on civic and political opinion formation (Jaidka, Zhou, Lelkes, Egelhofer, & Lecheler, 2022).

The theoretical contribution of this study is to connect the current understanding of political deliberation with the role of platform affordances and platform norms in facilitating these discussions. Furthermore, measuring and characterizing the different dimensions of political talk would enable scholars to understand the trade-offs involved in enforcing civil discussions.

Operationalizing the social and analytical quality characteristics of online political talk

Social media users who discuss politics online are expected to be unlikely to indulge in reflection or frame coherent arguments (Janssen & Kies, 2005; Stromer-Galley & Martinson, 2009). Nevertheless, social media platforms are relevant for understanding how these online discussions trigger 'internal reasoned dissent' (Rinke, 2015, p. 3) rather than building consensus. That is, social media users engage with a "number of publicly available ideas, opinions, and arguments (and) different points of view" (Rinke, 2015, p.4) in the form of mediated deliberation.

This study distinguishes the analytical aspects of political talk from its interactive qualities. Table 1 crystallizes the coding criteria for each discussion quality facet. The conceptualization of political talk in this manner is based on prior work by Rinke (2015); Rowe (2015); Steenbergen et al. (2003).

The *analytical* aspect of deliberation emphasizes the logic, evidence, and rational arguments to make claims and promote the exploration of solutions through dialogue to build consensus or move the conversation forward (Gastil, 2008). In prior studies, it is often operationalized as the qualities of 'constructiveness,' 'justification,' and 'relevance' offered in political comments. **Constructiveness** is considered to be evidence of the author's attempt to (a) build and bring about consensus and propose solutions, (b) resolve conflicts by pointing out facts, and (c) identify common ground (Esteve Del Valle, Sijtsma, & Stegeman, 2018; Friess & Eilders, 2015; Steenbergen et al., 2003; Stromer-Galley, 2007).

Justification includes forms of internal and external justification to support a claim in an argument. Internal justification can take the form of (a) personal anecdotes or (b) values and ideologies. On the other hand, external justification is based on data, links, and facts (Oz, Zheng, & Chen, 2018; Rinke, 2015; Rowe, 2015).

Prior work says little about whether the social media posts under study actually discuss politics and choose instead to presume relevance (Rowe, 2015; Steenbergen et al., 2003). However, discussion quality lexica derived

Table 1. Operationalization of the deliberative criteria used to measure the quality of political discussion, with examples from the Twitter Politics dataset. Detailed annotation instructions are provided in the supplementary materials.

Facet	Definition	Example
Constructiveness	Language that attempts to move the conversation forward, build and bring about consensus, and resolve conflicts by pointing out facts, identifying common ground, or proposing solutions.	• @USER They love anyone who hates America as much as them. It's crazy that they can hate they country that made them rich so much. Robbing us is what they do best sadly. • @USER your A two faced liargo kiss soros butt, You are A Traitor-You need to leave your position • @USER the GOP IS A COMPLICIT SHIT SHOW! History will remember you as greedy old men who sold this country to the Russians and rich corporations. Kiss your political careers goodbye!
Justification	Language that offers evidence for a claim in the form of personal experiences, values, and feelings or data, links, and facts.	• @USER and all the corporate DEMs, you're on notice. https://<LINK> • @USER Sen. <name> (R) said; We all know <name> is a pile of crap. • @USER @USER handed <name> a blank check. they're full of it at this point. • @USER #BeInformed #ShapeUpOrWeWillVoteYouOut https://<LINK> • @USER Well at least they aren't giving the money away to foreign countries like Obama and Clinton • @USER (…) Illegals can NOT get medical and food stamps from the gov't. Stop lying, please. • @USER You received $6,986,620 fm the NRA. You have a conflict of interest. You put donor interests above common sense gun laws.
Relevance	Language that is about politics.	• @USER Why are you sponsoring legislation to stop Russia investigation? • @USER You received $6,986,620 fm the NRA. You have a conflict of interest. You put donor interests above common sense gun laws. • @USER Calling for a military coup against the President is DANGEROUS to the REPUBLIC. STOP THIS FARCE.
Reciprocity	Language that asks for information or opinions, i.e., the author is eliciting an answer from someone.	• @USER Please share copies or links • @USER what affect did the naming of Chad in the travel ban have on Niger? • @USER Why are you sponsoring legislation to stop Russia investigation? • @USER "Would have preferred" means that you are okay with this but that would have been better. Is this really what you mean? <name> got to you?

Facet	Definition	Example
		• @USER – the tax bill does not need minor tweaks – it needs a complete rewrite. Just say no. • @USER Wish Trey Gowdy would get off his high horse & get tough w his actions not just w his words! When will that happen
Empathy & Respect	Language that acknowledges or is sensitive to another's viewpoint. The author asks a genuine question, or appears to elicit a response or further information.	• @USER I now know who I won't support • @USER Don't let this bill take any deductions away from us(…). Thank you! • @USER Calling for a military coup against the President is DANGEROUS to the REPUBLIC. STOP THIS FARCE. • @USER #HandsOff People with disabilities will be hurt more than those without by these bills. Vote them down.
Uncivil behavior	Language that is abusive, racist, threatening, or exaggerating.	• @USER #Paid #Ass #Kisser = #Prostitute ?! • @USER exactly Hiding behind the new Reich? • @USER "Best treatment" eh? You hypocrit. No Obamacare for you – you're too special for that. No VA care either. SOB • @USER #SchummerShutdown. Somebody in DoD got the Games on today (…) Illegals over American Citizens great election strategy(…) Looser! #MORONTraitor • @USER #Sheisacrook cannot be #trusted #California #VoteheroutNOW. (…) • @USER #2. giving people in crap red states several times the electoral votes per person than YOUR home state. You lying piece of • @USER Paranoid, racist, apocalyptic ramblings of Charles Manson are the DNA of the reactionary Alt Right.

from political talk data would be meaningless in a discussion that does not involve political topics. Therefore, this paper uses the concept of **Relevance** to reflect whether a piece of text constitutes political talk.

Next, this study considered the *social* dimensions of deliberation in reciprocity, empathy and respect, and incivility, as participants also need to "listen to each other, show respect for each other and reflect on their interests" (Steenbergen et al., 2003).

Reciprocity, the degree of interactivity in a discussion, is considered an important component of deliberation, finding mention by Stroud et al. (2015), Friess and Eilders (2015), and Himmelroos (2017) in their discussions of online political deliberation.

Reciprocal dialogue comprises counter-assertions rather than response affirmations (Esteve Del Valle et al., 2018; Friess & Eilders, 2015; Stromer-Galley, 2007). Reciprocity indicates the author's attempt to (a) ask a genuine question, or (b) post a comment intended to elicit a response or further information.

Scholars have also considered the necessity to show **Empathy and respect** towards opposing viewpoints (Friess & Eilders, 2015; Steenbergen et al., 2003), which reflects an author's attempt to be sensitive to others, manifested in (a) positive comments, or (b) an empathetic or a respectful response acknowledging other viewpoints (Esteve Del Valle et al., 2018; Steenbergen et al., 2003). A complement to the measurement of empathy and respect is the antithesis of it, measured as the **Incivility** shown in political talk. Incivility is evidenced through obscene language, insults, stereotypes, and exaggerations (Oz et al., 2018; Stroud et al., 2015; Theocharis, Barberá, Fazekas, Popa, & Parnet, 2016). Scholars have theorized that online anonymity gives social media users the confidence to wield uncivil comments as a weapon against others, as a status symbol, and a way to assert their ideology over others and maintain the dominant status (Chen, 2017, p. 7).

Training classifiers to measure discussion quality

The first methodological choice for a scholar is to select the AI approach for training classifiers. An advantage of machine learning methods is that the obtained model coefficients can act as a lexicon or a scale for subsequent measurements. On the other hand, neural networks represent an *assortment of machine learning algorithms* applied on interconnected layers of nodes. Moreover, neural networks do not allow for the back-propagation of model coefficients to infer the importance of individual words and phrases. Therefore they offer limited insights for scholars looking to understand the quality cues in political talk. They are, however, helpful in augmenting data and have been used for that purpose in the present work.

The second methodological choice is determining which linguistic features should be used to train their supervised machine learning models. “Open” vocabulary features comprise all the words in the messages. “Closed” vocabulary features constitute words belonging to categories that reflect syntactic, grammatical, or signal psychological aspects such as the author's emotional, cognitive, or social processes. Prior work regarding the statistical analysis of political text (Monroe & Schrodt, 2008) has discussed how bag-of-words approaches, such as the open-vocabulary approach, do not allow for the nuanced understanding of actors, for instance "who attacked whom" (p. 353). It can similarly be argued that in an intense political discussion, it is necessary also to encode the intent, the rhetorical moves, and the tonality of an argument. In this study, the choice of the final set of features used to train text classifiers (and subsequently, to build lexica for discussion quality prediction) is based on how well they generalize the predicting the discussion quality of data from a new context.

The third methodological choice is regarding the heterogeneity of the training data. There may be concerns that lexica developed in one context or platform would not accurately measure the quality of political talk on other platforms. For example, some prior work has suggested that Facebook affords higher quality deliberation than Twitter (Oz et al., 2018), news websites afford higher quality deliberation than Facebook (Rowe, 2015), and Facebook comments are more deliberative and civil than YouTube comments (Halpern & Gibbs, 2013). By corollary, models trained on short social media posts, such as tweets, may not generalize to posts in Reddit communities. A major implication would be that all Reddit posts are deemed constructive and high-quality by a model trained on Twitter posts or that even incivility on Reddit is judged civil.

The primary analyses reported in this paper evaluate the second and third methodological choices made in developing data-driven lexica for discussion quality. The findings compare the performance of homogeneous training data (from the Twitter Deliberative Politics dataset) against heterogeneous data (from the CLAPTON dataset).

Since CLAPTON subsumes and extends the former, it is naturally expected that CLAPTON would lead to richer lexica for subsequent measurements and allow us to address the following research questions:

> **RQ1:** *How do classifiers trained on closed vocabulary features compare against those trained on open vocabulary features in their internal validity on held-out data?*

> **RQ2:** *How do lexica developed from a homogeneous dataset compare against those developed on heterogeneous data in their external validity on other datasets of political talk?*

> **RQ3:** *How do lexica developed from closed-vocabulary features compare against those developed from open-vocabulary features in their external validity on other datasets of political talk?*

Method

The following paragraphs describe how data were sampled and annotated to obtain the training set for machine learning. Next, the best-performing models are evaluated on four other datasets measuring political discussion quality contributed by authors of previous studies.

Dataset curation

The Corpus for the Linguistic Analysis of Political Talk ONline (CLAPTON) developed in this study comprises a heterogeneous training set (N = 6,000) of English political messages posted to Twitter, Reddit (N = 2,974), and during an online chat experiment (N = 400). These different platforms were chosen because participants are anonymous but face different degrees of moderation and chat synchronicity. The following paragraphs discuss these datasets and the rationale for including them in the training dataset.

First, the Twitter Deliberative Politics dataset developed by Jaidka et al. (2019) was used. A 1% sample of tweets for 15 months was filtered to comprise only replies to 536 US Congressmen and Congresswomen in office during that period. Next, the 1% sample was filtered down to a random sample of 6000 English tweets, annotated, and used to train language classifiers. By choosing a training dataset of replies to politicians, the posts are more likely to involve mediated deliberation (Rinke, 2015) and comprise political discussion. This dataset constituted the training data in the homogeneous dataset condition in RQ2.

Second, a random sample of the Reddit CMV dataset was annotated, comprising 2974 comments posted to the Reddit Change My View (CMV) community from 636 political discussions. Reddit communities are highly moderated, and participants are provided with detailed rules for participation. Instructions are also pinned to the home page of the forum and reinforced after every few posts by moderating bots. Users who do not follow the norms have their posts deleted and can get banned from the community. Therefore, comments are more likely to stay on the topic than on Twitter, where there is no moderation. The dataset was collected using stratified sampling to select two comments each from 1000 discussion topics where posts were at least ten characters long.[2]

Third, a random sample of the Trivium dataset from Jaidka et al. (2022) was annotated, which comprised 400 comments posted in a live online discussion about gun control. This setting involved an intermediate level of moderation, where discussion prompts were pinned to the top of the chat platform. It also differs from the others in terms of the synchronicity of the discussion compared to social media platforms. For example, messages are more likely to depend on messages by other participants for context (Baxter, 2006), thereby adding challenges to the problem of measuring discussion quality.

Annotation. Amazon Mechanical Turk was used to annotate the data input into the machine learning classifiers, following the procedure reported in (Jaidka, 2022). The quality criteria comprised residents of the United States with a minimum approval rate of 80% and a minimum of 1000 accepted hits. Four annotations per message were collected from the workers who participated in this task.

In keeping with best practices for text classification setups that are reported elsewhere (Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec, & Potts, 2013), only the labels with at least 75% agreement (which constituted 64.8% of all labels) were subsequently used in training and testing the machine learning classifiers. The inter-annotator reliability results of the final training set are provided in Table 2. The columns provided the pairwise percentage agreement and the percentage proportion of the total observations for which at least three coders agreed, out of 8175 observations. An average Krippendorff alpha score of 0.2 was obtained; however, it may be inappropriate for inter-annotator agreement measurement in this case for two reasons: firstly, there are hundreds of non-overlapping coders for the dataset, which defies its core assumption.

Secondly, the skew in label distribution means that chance correction would hurt the final agreement calculation (Potter & Levine-Donnerstein, 1999). An average Fleiss' kappa value of 0.32 is reported, which is considered a 'fair' agreement (Landis & Koch, 1977). The inter-class correlations for the facets averaged 0.67, which is considered moderate agreement. Examples from the dataset exemplifying each of the deliberative facets are reported in Table 1. A low inter-annotator agreement was observed in cases with sarcasm, irony, and implied context/intertextual references lost in post-level annotations (Jaidka, 2022).

Augmenting the data. The original data was small, comprising about 10,000 data points with a small proportion of positive cases for many critical dependent variables. Therefore, popular methods for data augmentation available through the HuggingFace libraries (Wolf et al., 2019) were used to increase the overall dataset size and thereby the number of positive cases available for training. First, data were augmented using back-translation into and from Spanish, chosen for its linguistic proximity to English. Pre-trained Neural Machine Translation models created by Helsinki-NLP are trained on

Table 2. Inter-annotator statistics for the training data. Only the observations with at least 75% agreement, comprising an average of 64.8% for the full dataset, were used in the supervised machine learning step.

Facet	% Agree 3 or 4 coders	Fleiss' kappa	Inter-class Correlations
Constructiveness	52.39	0.36	0.69
Justification	80.37	0.29	0.62
Relevance	89.74	0.26	0.59
Reciprocity	51.56	0.41	0.74
Empathy & Respect	76.70	0.36	0.69
Incivility	50.86	0.34	0.67
Average reliability scores	**64.85**	**0.32**	**0.67**

news text and short news headlines. They were therefore anticipated to work well for the cross-translation of political tweets. The second data augmentation method was to further expand the vocabulary of the corpus by using the contextual word embeddings model (Kobayashi, 2018) available through the *nlp.aug* package[3]. In this manner, the original training datasets (both for the homogeneous and the heterogeneous setting) were augmented by 200%.

Extracting the features

Two main sets of features were evaluated. First, **closed-vocabulary classifiers** were trained on discursive features, comprising many stylistic, argumentative, rhetorical, and psycholinguistic features of the text:

- Stylistic features comprise scores for different politeness features and discourse connectives (Danescu-Niculescu-Mizil et al., 2013; Niculae, Kumar, Boyd-Graber, & Danescu-Niculescu-Mizil, 2015) in writing. In prior work, they have been applied to model politeness and trustworthiness in text.
- Grammatical and syntactic features, such as the presence of different parts of speech in writing.
- Psycholinguistic features denote the emotional, cognitive, social, and perceptual processes and the time orientation and personal concerns elicited in writing. These features are available from the Linguistic Inquiry and Word Count dictionaries (Pennebaker et al., 2015).

Because the closed-vocabulary approach did not include content or topic features, it was anticipated to be more transferable across domains than a context- or content-sensitive approach.

The second approach was an **open-vocabulary approach** that trained classifiers on the stylistic features, and the term frequency-inverse document frequency (TFIDF) features on tens of thousands of salient words and phrases representative of the corpus, together with the 125 discourse features mentioned above. First, one-, two-, and three-word phrases were extracted and converted into a frequency distribution. Next, the product of their term frequency and the inverse of their message frequency was calculated to reflect their importance and uniqueness to a message.

While the first approach is context agnostic, the second approach is anticipated to be more sensitive to the relevance and the quality of the justifications provided in the text.

Examples of these categories, their definitions, and some of the underlying linguistic cues are provided in Table 3.

Training classifiers

This study evaluated many of the classifiers available in the *sklearn* Python package to predict different facets of discussion quality as a function of the linguistic features of the input. The frequency distributions of the linguistic features of labeled tweets are the independent variables, and the labels about the presence or absence of each facet are the dependent variable.

In the internal validation step, ten classification approaches with differing underlying assumptions (K-Nearest Neighbors, Decision Trees, Linear Discriminant Analysis, Linear- and C-Support Vector classification, Gaussian Naive Bayes, Bernoulli Naive Bayes, Gradient and Ada Boosting, and Logistic Regression) were evaluated.[4]

In the training setup, following the best practices documented in similar machine learning studies (Davidson et al., 2017), feature selection was applied before training each classifier to discard those independent variables that were not univariately associated with the dependent variable. Feature selection was performed by fitting logistic regression models with an L2 penalty to each dependent variable – a recommended practice for reducing high-dimensional spaces and improving classifier accuracy (Pedregosa et al., 2011). This identified only the most relevant features. The logistic regression, linear-and c-support vector classification approaches were set up using 'balanced' class weights with an L2 penalty and a maximum of 100 iterations in the next step. First, the approach will adjust the weights of observations toward the final performance evaluation in data with class imbalance. The weights are inversely proportional to the label frequencies in the input data. Second, when regularizing increasingly complex models, L2 penalties were applied because they

Table 3. Exemplar linguistic features used to train the machine learning classifiers. Over 42000 features were input into the feature selection and classifier training pipelines for each label.

Feature	Definition
Discursive features in the Closed-vocabulary and Open-vocabulary classifiers	
Syntactic and grammatical features	
Syntactical features	Features that are used organize information in English message, such as punctuation marks, but also social media-specific syntactical features, such as hashtags,.
Grammatical features	Parts of speech in the English language, such as noun, pronoun, verb, adjective, adverb, preposition, conjunction, and interjection. They can also include linguistic units indicating the types of words, such as words indicating quantities, persons, and tenses.
Politeness features (Danescu-Niculescu-Mizil et al., 2013)	
Hedges	Words expressing ambiguity, probability, caution, or indecisiveness, e.g., “unlikely,” “think,” and “in general.”
Factuality	Linguistic cues that are used to report a fact, e.g., “point,” “reality,” “truth,” “actually,” and “honestly.”
Deference	Linguistic cues that are used to defer to another person, e.g., “great,” “good,” “interesting,” and “awesome.”
Apology	Linguistic cues that are used to issue an apology, e.g., “sorry,” “forgive,” and “excuse.”
Gratitude	Linguistic cues that are used to show gratititude, e.g., “thanks,” “thank,” and “appreciate.”
Greeting	Linguistic cues that are used to greet someone, e.g., “hi,” “hello,” and “hey.”
Discursive features (Niculae et al., 2015)	
Contingency	Discourse connectors used to indicate a contingency relationship between clauses, e.g., “thus,” and “indeed.”
Expansion	Discourse connectors used to indicate an expansion relationship between clauses, e.g., “rather” and “for instance.”
Temporal	Discourse connectors used to indicate temporal relationship between clauses, e.g., “still,” “while.”
Psycholinguistic features (Pennebaker et al., 2015)	
Emotional processes	Categories used to indicate emotional expression. Individual scores for the emotional categories (anger, sadness, anxiety) are reported, as well as a summative affect, positive emotion, and a negative emotion score.
Cognitive processes	Categories used to indicate styles of thinking and processing information. Scores are reported for cognitive processing and analytical thinking, as well as a number of sub-categories, e.g., differentiation, comparative language and so on.

Feature	Definition
Social processes	Categories used to indicate social processes, such as mentions of other individuals and groups. Individual scores for the types of reference (second-person, third-person, male, female) are reported, as well as a summative score for affiliation, group identity, and so on.
	Content features in the Open-vocabulary classifiers
TF-IDF features	Words and phrases weighted by their importance in a message. The weight is measured as a ratio of the frequency of the term in the text to the frequency of the term in the overall collection of messages.

shrink coefficients evenly, which is more appropriate when features are collinear or correlated (as is indeed the case with linguistic features). Third, an artificial limit on the maximum number of iterations allowed for the classifiers was set to 100 to force an evaluation even when classifiers failed to converge.

The model coefficients can be used to calculate the presence of a discussion quality facet as a function of the weighted average of the presence of different linguistic features in a text input, thereby 'generating a prediction' about the presence of a facet of discussion quality.

Performance evaluation

The validity of the classifiers was established through an internal validation on held-out data from the same dataset in a ten-fold cross-validation setup and external validation on hand-annotated datasets from other studies.

For internal validation, feature extraction steps were followed to obtain the discursive and content features. Next, classifiers were trained on closed- and open-vocabulary features (including words and phrases together with closed-vocabulary features) to obtain the closed- and open-vocabulary classifiers. Finally, all the classifiers were trained on 90% of the data and tested on 10% held-out data in a ten-fold cross-validation setup. This procedure was repeated ten times, and the average performance scores across the ten runs were reported.

External validation

This study applied the trained closed- and open-vocabulary classifiers to predict the discussion quality facets in other political comments' datasets from recent work.

Comparing the classifier-predicted and the hand-annotated labels can establish how well the classifiers can generalize to measurements in different contexts.

Two methodological considerations guided the choice of datasets. First, this study wanted to test whether the classifiers can generalize to measure English posts from a different English-speaking nation, such as Canada. Previous work has suggested that language models trained in one context may not apply to the same language spoken in another region because of cultural differences, let alone another part of the world. Second, different platforms would have different deliberative norms that warrant extra validation.

The authors of previous studies kindly provided four datasets that were used for external validation:

- The dataset from Theocharis et al. (2016) comprises tweets posted in reply to UK political candidates before the European Parliament elections.
- The dataset from Halpern and Gibbs (2013) comprises comments posted to the Facebook and YouTube accounts managed by the White House between June 15 and July 15, 2010.
- The dataset from Fournier-Tombs and Di Marzo Serugendo (2019) comprises political comments posted on Facebook Live, political blogs, and during live townhalls in the US and Canada between 2011 and 2017.
- The dataset from Stromer-Galley (2007) comprises utterances from the Virtual Agora project, in which 564 participants in face-to-face discussions deliberated on school policies in July 2004.

Annotations for a random sample of 500 messages were crowdsourced from Amazon Mechanical Turk. Inter-coder agreement statistics are reported in the supplementary materials[5]. The distribution of the labels is reported in Table 4. The number of 'positive instances' for each dataset reflects the number of observations (of the total) labeled to have the facet present.

Next, the predicted labels were obtained by applying the classifiers to the same text.

First, feature extraction was performed to obtain the set of features. Then, following standard practices, the data transformations for TFIDF from the training set were applied to weight the content features in this datasets (Davidson et al., 2017; Pedregosa et al., 2011). Finally, the lexica developed from pre-trained classifiers (from the previous steps) generated a 0 or a 1 label for each message signifying the presence or absence of a discussion quality facet.

Table 4. Descriptive statistics about the datasets used for external validation in this study. The number of 'positive instances' reflects the number of observations (of the total) which actually denoted the presence of that discourse quality facet.

	Positive Instances						
	Training sets (after augmentation)			Validation sets			
Facet	**Twitter Politics**	**Trivium**	**Reddit CMV Politics**	**Theocharis et al. (2016)**	**Halpern & Gibbs (2013)**	**Fournier-Tombs & Di Marzo Serugendo (2019)**	**Stromer-Galley (2007)**
Constructiveness	531 (1,593)	92 (276)	148 (444)	156	148	171	142
Justification	67 (201)	252 (756)	493 (1,479)	428	416	397	416
Relevance	953 (2,859)	90 (270)	617 (1,851)	381	431	427	300
Reciprocity	4,689 (14,067)	21 (63)	485 (1,455)	308	311	318	304
Empathy & Respect	2,842 (8,526)	168 (504)	535 (1,605)	381	343	424	388
Incivility	382 (1,146)	8 (24)	137 (411)	105	132	105	81
Total number of observations	6,000 (18,000)	400 (1,200)	2,974 (8,922)	500	500	500	500

Evaluation metric. To benchmark the predictive performance against prior work, this study reports the accuracy of predictions against hand-annotated labels as the primary evaluation metric. The F1 score reports the harmonic mean of the precision and the recall (how many positive cases were correctly predicted). Therefore, it denotes the trade-off involved when a machine learning algorithm aims to improve the precision for the minority class, which inadvertently also increases the number of false positives.

Results

The following paragraphs discuss the results. First, To answer **RQ1**, this study reports the individual and average prediction performance from a ten-fold cross-validation setup for the classifiers trained on the open- and closed-vocabulary features extracted from the training dataset. Then, to

Table 5. Internal validation – predictive performance of the best-performing closed- and open-vocabulary machine learning classifiers on held-out data in a ten-fold cross-validation setup. Scores closer to 1 implies that a greater number of cases were correctly predicted as positive or negative.

	Best-performing classifiers							
	1	2	3	4	5	6	7	8
Facet	**Feature set**	**Approach**	**Accuracy**	**F-1 score**	**Precision**	**Recall**	**Minority-F1**	**Area Under the Curve**
Constructiveness	**Closed**	C-Support Vector	0.73	0.59	0.36	0.69	0.25	0.64
	Open	Logistic regression	**0.93**	0.85	0.74	0.88	0.67	**0.81**
Justification	**Closed**	C-Support Vector	0.74	0.72	0.63	0.72	0.82	0.80
	Open	Logistic regression	**0.89**	0.88	0.84	0.88	0.93	**0.9**
Relevance	**Closed**	C-Support Vector	0.81	0.75	0.64	0.78	0.90	0.84
	Open	Logistic regression	**0.93**	0.9	0.86	0.91	0.96	**0.94**
Reciprocity	**Closed**	Decision tree	0.77	0.72	0.60	0.72	0.60	0.59
	Open	Logistic regression	**0.91**	0.89	0.84	0.89	0.82	**0.87**
Empathy & Respect	**Closed**	Decision tree	0.70	0.69	0.65	0.69	0.73	0.73
	Open	Logistic regression	**0.89**	0.89	0.88	0.89	0.91	**0.9**
Incivility	**Closed**	C-Support Vector	0.84	0.64	0.37	0.71	0.28	0.56
	Open	Logistic regression	**0.94**	0.83	0.69	0.86	0.64	**0.75**

answer RQ2 and RQ3, it reports the prediction performance from lexica trained on homogeneous, heterogeneous datasets, and closed- and open-vocabulary settings on four other datasets of political talk.

Internal validation

To answer RQ1, the first set of results in Table 5 reports the performance of the best classifiers (i.e., the classifiers with the highest average of accuracy and the AUC metrics) trained on the closed- and open-vocabulary features.

We observe that the open-vocabulary classifiers trained on content features had an advantage over the closed-vocabulary classifiers for all the predictive tasks. The c-support vector classifiers often had the best performance in the closed-vocabulary classifiers; however, they did not converge within 100 iterations in the open-vocabulary classifiers.

Instead, the best performance was observed with logistic regression classifiers. The findings concur with Jaidka (2022) for the Deliberative Politics dataset, but with substantially higher minority-F1 scores.[6] Columns 2 and 3 report the predictive performance metrics for the best-performing classifiers for each label. In drilling down into the performance of the open-vocabulary classifiers, a high average accuracy score (0.92) with a low standard deviation underplays the broader variability in the minority-F1 (Mean = 0.81, standard deviation = 0.07). The minority-F1 identifies the predictive performance on the positive cases alone and suggests that incivility has the poorest performance (minority F1 = 0.69).

Exemplar lexical features and weights (the model features and coefficients) from the best-performing logistic regression classifiers are provided in Table 6. The content features appear to pick up on the salient political topics in the USA (where these datasets were collected from). However, many of the classifiers include words related to political parties (e.g., *GOP, democrats*), ethnicities (e.g., *black, white*) and genders (e.g., *white man, black women, trans women*) as indicators of higher or lower discussion quality. These are reported in italics in Table 6.

External validation

To answer RQ2 and RQ3, Table 7 compares the accuracy of lexica developed on homogeneous and heterogeneous data on other datasets of political talk.[7]

The validation suggests that the classifiers had a moderate-to-good cross-platform accuracy, ranging from an average accuracy of 0.45 for reciprocity to 0.82 for justification. Classifiers for justification, incivility, and relevance had an average accuracy greater than 0.7. These findings can be benchmarked against the reported accuracy scores in similar text classification problems. For instance, experiments in training the state-of-the-art politeness classifier in Danescu-Niculescu-Mizil et al. (2013)) report accuracies between 0.68 to 0.78. To predict trustworthiness from text, the paper by Niculae et al. (2015) reports an accuracy of 0.57. In problems involving downstream predictions of persuasion from text, Peskov et al. (2020) report F1 scores in the range of 0.48 to 0.53.

Table 6. **Some of the features and classifier coefficients from the best-performing logistic regression classifiers on open vocabulary features.**

Constructiveness (12,409 features)		**Justification** (10,623 features)		**Relevance** (9,960 features)		**Reciprocity** (13,347 features)		**Empathy & Respect** (13,327 features)		**Incivility** (9,521 features)	
Feature	Coefficient	Feature	Coefficient	Feature	Coefficient	Feature	Coefficient	Feature	Coefficient	Feature	Coefficient
					TF-IDF features						
@user @us	0.25	@user how	0.13	money	0.17	do with	0.13	@user tell	0.13	you work	0.12
flank	0.12	@user thats	0.1	november.	0.12	@user @us	0.13	@user how	0.11	eyes?	0.11
congress-man is	0.12	works?	0.09	senate.	0.08	you follow	0.12	is well	0.11	please explain	0.11
%.	0.12	technically speaking,	0.09	potentially lose	0.08	useless.	0.11	prayers with	0.09	slut	0.1
@user that	0.11	inalienable	0.09	republic.	0.08	ago?	0.11	agree, the	0.08	brexit.	0.1
democrat.	0.08	*democrats and*	0.06	*white man*	0.02	*black voters*	0.02	*white as*	0.05	*hillary.*	0.12
black women	0.03	*is white*	0.04	*black lives*	0.04	*whites.*	0.04	*democrat.*	0.08	*republicans and*	0.05
when white	0.03	*republican to*	0.04	*democrat.*	0.05	*democrat.*	0.05	*@user republicans*	0.05	*democrats and*	0.06
					Grammatical and syntactic features						
Quantifier	0.01	Verb	0.1	Preposition	0.26	Question mark	0.29	Auxiliary verb	0.06	Second person pronoun	0.07
Impersonal pronoun	0.02	Preposition	0.15	Question mark	0.15	Second person pronoun	0.02	Comma	0.05	Question mark	0.1

Discursive features											
Contingency	0.02	Temporal	0.05	Temporality	0.15	Contingency	0.04	Contingency	0.05	Contingency	0.01
Expansion	0.01	Comparison	0.04	Contingency	0.07			Comparison	0.02		
Politeness features											
Deference	0.09	Deference	0.05	1st person pl.	0.05	Factuality	0.05	Hedges	0.1	First person pronoun and positive emotion	0.01
Factuality	0.05	Greeting	0.05	Deference	0.04	Hedges	0.05	Greeting	0.07	Start with first person pronoun	0.01
Psycholinguistic features											
Comparison	0.02	Comparison	0.04	Past focus	0.09	Social processes	0.03	Insight	0.1	Female pronouns	0.19
Certainty	0.02	Past focus	0.15	Clout	0.23	Discrepancy	0.01	Positive emotion	0.07	Anger	0.15
TF-IDF features											
your opinion	-0.06	everyone is	-0.1	yeah i	-0.11	and/or have	-0.09	haha	-0.1	@user tell	-0.12
pleasant,	-0.07	thank you!!!	-0.1	can provide	-0.11	martyr, blocking	-0.08	but doesn't	-0.1	yes you	-0.08
@ one	-0.08	god.	-0.11	definite	-0.1	yes you	-0.08	least.	-0.11	wall.	-0.09
the narrative	-0.07	you lose	-0.11	the ridiculous	-0.1	so what's	-0.08	politician,	-0.11	witch hunt	-0.06

female	-0.06	bot, and	-0.13	risk,	-0.1	hypothetically,	-0.07	america would	-0.11	wont be	-0.06
republican, i	-0.02	*republicans and*	-0.1	*the republican*	-0.03	*women who*	-0.05	*black populations*	-0.02	*the democrats.*	-0.02
black woman,	-0.01	*black women.*	-0.04	*black populations*	-0.04	*trans women*	-0.02	*whites.*	-0.02	*black woman,*	-0.01
white supremacy.	-0.01	*for women.*	-0.05	*trans women*	-0.03	*republicans have*	-0.02	*when women*	-0.04	*the whites*	-0.01
Grammatical and syntactic features											
Personal pronoun	-0.04	Parentheses	-0.19	Second person pronoun	-0.11	Quote	-0.09	Parentheses	-0.11	Question mark	-0.1
Conjunction	-0.07	Second person pronoun	-0.08	Adjective	-0.06	Adjective	-0.06	Conjunction	-0.1	Adverb	-0.1
Discursive features											
Temporality	-0.05					Temporality	-0.05	Temporality	-0.02	Temporality	-0.15
Comparison	-0.04					Comparison	-0.01	Expansion	-0.03	Expansion	-0.02
Politeness features											
Hedges	-0.04	Hedges	-0.04	Gratitude	-0.04	Hedges	-0.04	Factuality	-0.03	Greeting	-0.09
Start with second person pronoun	-0.03					Gratitude	-0.06	Start with first person pronoun	-0.03	Gratitude	-0.05
Psycholinguistic features											
Negative emotion	-0.04	Tentativeness	-0.07	Anger	-0.08	Anger	-0.08	Anger	-0.1	Positive emotion	-0.11
Cognitive processing	-0.05	Positive emotion	-0.04	Certainty	-0.06	Affiliation	-0.14	Tentativeness	-0.11	Cognitive processing	-0.07

Table 7. External validation: the predictive performance (accuracy) of the best performing classifiers in the closed- vs. open, and the homogeneous vs heterogeneous training setups, on other datasets. A value closer to 1 implies that a greater proportion of cases were correctly classified. The macro-average scores for F1 are reported in the supplementary materials.

	1	2	3	4	5	6	7
			Classifiers with the best performance (Accuracy)				
Facet	**Best training data**	**Best feature set**	**Halpern & Gibbs (2013)**	**Fournier-Tombs & Di Marzo Serugendo (2019)**	**Theocharis et al. (2016)**	**Stromer-Galley (2007)**	**Average (SD)**
Constructiveness	Heterogeneous	Closed	0.68	0.64	0.63	0.59	**0.64 (0.04)**
Justification	Heterogeneous	Closed	0.79	0.84	0.75	0.89	**0.82 (0.06)**
Reciprocity	Heterogeneous	Open	0.62	0.36	0.38	0.34	0.43 (0.13)
Relevance	Heterogeneous	Open	0.72	0.79	0.66	0.62	**0.7 (0.07)**
Empathy & Respect	Heterogeneous	Open	0.68	0.85	0.76	0.16	**0.61 (0.31)**
Incivility	Homogeneous	Closed	0.74	0.79	0.79	0.95	**0.82 (0.09)**
Average			**0.68**	**0.72**	**0.67**	**0.58**	

The Table highlights three main takeaways. First, there is a remarkable drop in predictive performance compared to the accuracies for in-domain data in Table 5. The best performing classifier on average (highest accuracy) was for justification (Mean accuracy = 0.82). The poorest performance was for reciprocity (Mean accuracy = 0.45). The highest performance variability was empathy and respect (Mean accuracy = 0.61, standard deviation = 0.31).

Second, as was expected, in all cases except incivility, lexica trained on heterogeneous datasets outperformed those trained on homogeneous datasets in external validation. The detailed results reported in the supplementary materials demonstrate an average improvement of 20.5% in accuracy across the different facets. Finally, for constructiveness, justification, and incivility, lexica trained with stylistic features appears more insightful than content features. Classifiers trained with content features may overfit the dataset vocabulary, thus limiting their generalizability to new contexts.

Third, performance was variable across datasets. The best performance – as well as the one with the highest variance – was on Fournier-Tombs and Di Marzo Serugendo (2019). The lowest average performance was on the dataset from Stromer-Galley (2007) which was the only dataset collected in a face-to-face context. The possible causes of error are discussed in the supplementary materials, with examples of false negatives (positive cases marked negative). Classification errors occurred on short messages, messages with French words, and messages that mentioned politicians by name, as those were not detected as politically relevant.

Discussion and Conclusion

This study developed lexica to measure the discussion quality of online political talk (in English). Annotating text regarding the facets of discussion quality is a complex task that suffers from moderate agreement even when well-specified. However, we noted that a high agreement did not necessarily equate to a high-performance classifier (e.g., for reciprocity), suggesting (not surprisingly) that while humans might intuitively know and recognize high-quality text when they see it, it is harder to teach an algorithm to pick up on similar cues.

Lexica built on heterogeneous datasets have better performance than homogeneous datasets. Closed and open vocabulary features offer different advantages for precise measurements of quality facets, such as constructiveness and relevance. Although trade-offs are involved when participating in political discussions, being analytical would not necessarily require a participant always to be respectful of or civil to others. The finding supports recent work that argues that incivility is not anti-correlated to the quality of political talk (Maia & Rezende, 2016).

Inspecting the derived lexica allows us to raise further concerns about the generalizability, face validity, and *biases* they encode. For instance, some of the content features (e.g., *trump, donor, christian*) would not be meaningful in other contexts and countries. Furthermore, inspecting the lexica reveals that social identity words (e.g., *gop, democrat, black, white*) are indicators of discussion quality. Scholars may want to prune these lexica to remove social and ethnic identity markers and first names. Keeping them in the lexica could inadvertently mischaracterize all the content about the populations and ideological groups as uncivil or low quality. Alternatively, Dobbrick, Jakob, Chan, and Wessler (2021) suggest that a sensitivity analysis could be performed to remove all the words in the training step itself. These methods could be evaluated and compared in future work.

Previous findings have suggested that the different affordances of social media platforms, which can interact in ways that affect the participants' behavior, offer grounds for new measurements and experiments at the intersection of theory and practice. The datasets we chose for validation have varied affordances for discussion participation. They also vary stylistically when discussions occur in a *synchronous* versus an asynchronous setting. Synchronous conversations can build on previous speakers' comments without establishing relevance (Baxter, 2006), as was evident in the low proportion of reciprocity instances in the Trivium dataset. These differences affect reciprocity, empathy, and respect, i.e., how users react to content and perceive and respond to each other. For example, users who self-select into political talk on YouTube are likely to eschew (or unlikely to expect) turn-taking behavior in civil discussions favoring brief, humorous exchanges.

Machine learning classifiers can offer empirical insights into theoretical expectations. The inter-relationship between the different discussion quality facets is reported in the supplementary materials, highlighting the trade-offs in choosing a more analytical vs. a more social stance in a political discussion. Syntactical and grammatical features predict discussion quality, suggesting that social media users (and the annotators) consider linguistic sophistication the normative criterion for deliberation. It is also interesting that the choice of which machine learning approach contributes significantly to model performance beyond feature selection. As discussed in Jaidka (2022), we may see logistic regression win out in the internal validation because it makes no assumptions about class distribution, which is helpful in cases with imbalanced data. Moreover, it is effective when classes can be linearly separated. However, support vector classification outperformed logistic regression in external validation, especially for complex categories such as constructiveness, justification, and incivility. These methods work well when there is a clear separation between the classes in high-dimensional spaces. The code and dashboard released with this study will let readers compare the efficacy of different approaches for any input text.[8]

Recent work has explored other facets of political discussion quality, such as outrage (Berry & Sobieraj, 2013; Jakob, Dobbrick, Freudenthaler, Haffner, & Wessler, 2022) and integrative complexity (Jakob, Dobbrick, & Wessler, 2021). A recommended practice is to curate labeled training datasets from multiple social media platforms and use discursive and content features to train generalizable classifiers. Building on the suggestions by Grimmer and Stewart (2013), while no method is perfect, validation against hand annotations and external validation on new data is necessary to ensure that the approaches do not overfit the training sample and that the predictions constitute meaningful signal.

While this study has incorporated some challenging ideas in measuring discussion quality, such as relevance, other modeling approaches may invoke inter-label associations to predict the facets, as suggested by Erlich, Dantas, Bagozzi, Berliner, and Palmer-Rubin (2021). The task of pruning lexical features could test more systematic approaches, such as Bayesian shrinkage and regularization suggested by Monroe, Colaresi, and Quinn (2008).

There is also much to be done to model political deliberation through its linguistic features and as a back-and-forth exchange between two agents in a given social environment. A network-based approach, as suggested by Beauchamp (2020), would potentially be needed for such an operationalization to work in tandem with a purely natural language processing (NLP)-based technique to understand both cross-sectional and longitudinal trends in data. Alternatively, data can be modeled using actor-partner interdependence models (Liao, Zhang, Oh, & Palomares, 2021) to understand the evolving impact of encountered discussion quality on participant responses. This study must be replicated in contexts with different languages, and multi-lingual contexts, with suitable adjustments to the methodology for data annotation, augmentation, and training. Future work could also examine the persuasion effect of different deliberative facets on citizens' opinion formation.

Data Availability Statement

The supplementary materials are available at:
https://doi.org/10.17605/OSF.IO/28ESD
The code used to extract features and train the classifiers is available at:
https://github.com/kj2013/deliberative-politics
A dashboard to test and compare the outputs generated by the different classifiers is available at: https://share.streamlit.io/sriramelango/nus-political-discourse-quality/app.py

Acknowledgment

This project was funded by a Nanyang Presidential Postdoctoral fellowship and a start-up grant from the National University of Singapore. Sriram Elango created the dashboard to surface the classifier outputs. I would like to thank Yphtach Lelkes for his time and extensive feedback on the many early versions of this paper. This paper was possible because of inputs from the following awesome colleagues: Alvin Zhou, Daniel Halpern, Eleonore

Fournier-Tombs, Gina M. Masullo, Ian Rowe, Jennifer Stromer-Galley, Mark Boukes, Matti Nelimarkka, Mustafa Oz, Natalie Stroud, Sophie Lecheler, Staffan Himmelroos, Subhayan Mukerjee, Yannis Theocharis, and Yujin Kim. Thank you!

Supplementary materials

Annotation instructions

An Amazon Mechanical Turk (AMT) task was launched to obtain four annotations for each of the deliberative facets in each message across all the datasets. The instructions are provided below.

Inter-annotator agreement

The inter-annotator agreement for the annotations on the external datasets are reported in Table 2.

Choice of machine learning approaches

The different machine learning approaches under the *scikitlearn* package (Pedregosa et al., 2011) formulate the classification problem under different assumptions. For instance, some may classify data points by minimizing distances between points (K-Nearest neighbors, Linear Discriminant Analysis) or a point and a line (Support Vector Machines). They may also classify data points based on their attributes (Decision tree) or look for a linear relationship between their attributes and the label. Alternatively, they may assume that all the data points' attributes are independent of each other and assign probabilities based on these attributes using a Gaussian (Gaussian Naive Bayes) or a Bernoulli (Bernoulli Naive Bayes) distribution of probabilities. Finally, classifiers may also iteratively improve weak learners using negative gradients (Gradient boosting) or exponential gradients (Ada boosting) of the loss function to reduce the losses made during prediction.

Additional results

Internal validation: all results with the mixed training dataset

Detailed results about the evaluation of different linguistic features, as well as the ones finally chosen to train the model, are provided in Table 3 and Table 4.

External validation

Detailed results of the external validation approaches applying both the closed- and open-vocabulary classifiers are reported in Table 5.

Insights

Prediction errors

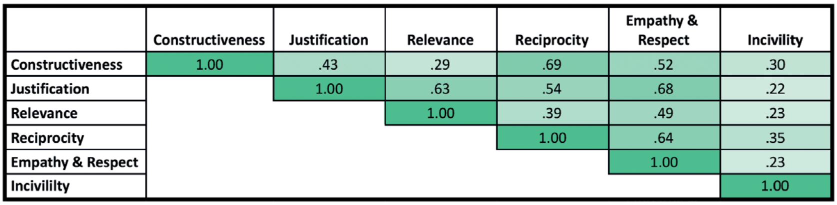
Table 6 illustrates some of the prediction errors (positive cases marked as negative) made by the classifiers. Some of the errors may have been caused because the text was in French (see the examples from Fournier-Tombs and Di Marzo Serugendo (2019)), other messages may have been too short (see examples from Halpern and Gibbs (2013)). Some were irrelevant to politics (e.g., from Stromer-Galley and Martinson (2009)) which made prediction a difficult task. There were fewer cases of misclassification for the incivility category.

Theoretical insights

A theoretical puzzle remains regarding the inter-relationship of discussion quality facets. While Wessler (2008) and Rinke (2015) theorize the meta-structure of deliberation, there is also a need to examine the structure of the arguments themselves. Studies examining online political posts have often reported mixed or null findings regarding the association of rationality and (in)civility (Jaidka, Zhou, & Lelkes, 2019; Maia & Rezende, 2016; Rossini, 2020). Papacharissi (2004) and Groshek and Cutino (2016) suggested that politeness (or the absence of incivility) could restrict conversation by making it less spontaneous. Rinke (2015) concurs that "some tolerance of incivility is necessary for a deliberative public sphere" (p. 7). This would imply that there may be trade-offs in having more analytical vs. more social and civil discussions. Such insights are only possible when scholars can understand how the different deliberative facets relate to each other.

To test these ideas, we conducted a pairwise correlation of the discussion quality facets provide insights into the structure of political discourse. Figure 1 provides the Pearson's *r* among the facets in the training data used in this study (N = 9,274). The color and the shade of the cell denote the direction and strength of the correlation (all p < 0.01, Bonferroni-corrected). Constructiveness and reciprocity are strongly associated with each other (*r* = .69). The relationship of empathy and respect is the strongest with justification (*r* = .68) and reciprocity (*r* = .64). Among the analytical facets, relevance has the least inter-correlation with the other facets. This is suggests the importance of having a separate measure of relevance when characterizing political talk, and the trade-off involved when invoking more discursive than subjective norms in their responses. For instance, when participants choose to invoke more constructiveness in political talk they may sacrifice their relevance to political discourse (*r* = 0.29).

An unexpected finding is the moderately positive association of incivility with other discussion quality facets. It has a positive correlation with the analytical aspects, such as constructiveness (r = .30) and justification (r = .22). Qualitative examples to contextualize this finding are provided in Table 7. They illustrate how longer tweets can offer insults yet offer justification for their attacks. Obviously, we should not interpret this to mean that political talk *should* include extremely intolerant expressions such as hateful speech. But evaluating the analytical facets of an uncivil social media post (and vice-versa) could offer a more nuanced understanding of its quality.



	Constructiveness	Justification	Relevance	Reciprocity	Empathy & Respect	Incivility
Constructiveness	1.00	.43	.29	.69	.52	.30
Justification		1.00	.63	.54	.68	.22
Relevance			1.00	.39	.49	.23
Reciprocity				1.00	.64	.35
Empathy & Respect					1.00	.23
Incivililty						1.00

Figure 1. Pairwise Pearson's r between the discussion quality facets in the mixed training dataset (N = 9274).

Table 1. The instructions used as a part of the Amazon Mechanical Turk to annotate the Twitter Deliberative Politics dataset.

Short Instructions
This tweet is a reply on Twitter (i.e., a Tweet) to a United States member of the Congress. Please classify this tweet according to whether it
(a) is about politics (b) is positive/respectful (c) uncivil (d) has a genuine question (e) has a justification (f) is constructive. Each HIT takes about 30 seconds.

Steps

- Read the tweet.
- Determine which categories best describe the tweet

Relevance

- YES: Whether this tweet is probably about politics, or
- NO: this tweet is irrelevant to politics.

Positive/Respectful

- YES: Whether this tweet shows respect or empathy towards others, or
- NO: This tweet is not particularly positive or respectful.

Uncivil

- YES: Abuses and sledging: Whether this tweet uses ideological extremes like "liberal potheads", abuses like "ass" or "moron", stereotypes like "faggot" or "backward" or "terrorist"

- YES: Threatening: Whether this tweet threatens individual freedoms ("You people better shut up"), threatens someone or threatens democracy ("American people must take him down")
- YES: Exaggeration: Whether this tweet uses exaggerated arguments (e.g. "It's very easy to solve all of this just keep your legs closed if you don't want a baby."), or
- NO: This tweet is not particularly uncivil.

Reciprocity

- YES: Whether this tweet asks questions that were designed to elicit opinions or information (Where is the money coming from? Increased taxes?"), or
- NO: This tweet does not ask a genuine question or asks rhetorical questions ("You have no idea how limiting Medicaid coverage can be, do you?").

Constructiveness

- YES: Fact-checking: Whether this tweet contains fact-checking "(1) that's not a real quote 2) more importantly, since then the DNC has embraced racially progressive stances… Mostly.") ("Not exactly true…she's tried to invent a Native American heritage that failed epically")
- YES: Common ground: Whether this tweet contains a search for common ground ("You are undoubtedly right (correct, too). No matter how Conservative I am I am still a Mom and my heart strings get tugged easily.") ("I'm all for progressive change but too much will lead to repeat 2016") ("We can keep getting lost in the weeds") ("We are not all like that :)")
- YES: Solution: Whether this tweet contains a solution ("It would be WONDERFUL if the House & Senate committees looked into..")("Also, no one is blaming Pence, Sec. Price for not getting it passed. Why not?")
- NO: This tweet is not constructive.

Table 2. Inter-annotator agreement statistics for the second round of annotations, performed on the datasets used in the external validation.

	Average percentage agreement (%)			
	1	2	3	4
Label	**Theocharis et al. (2016)**	**Halpern & Gibbs (2013)**	**Fournier-Tombs & Di Marzo Serugendo (2019)**	**Stromer-Galley (2007)**
Constructiveness	77.71	76.66	77.49	74.92
Justification	69.08	73.79	73.84	70.17
Relevance	76.49	81.68	84.08	84.15
Reciprocity	77.17	76.2	76.02	76.26
Empathy & Respect	80.34	78.33	77.49	81.95
Incivility	80.14	79.61	78.55	79.07

Table 3. Internal validation with closed-vocabulary features on heterogeneous data in a ten-fold cross-validation setup. Scores closer to 1 implies that a greater number of cases were correctly predicted as positive or negative.

Approach	1	2	3	4	5	6		1	2	3	4	5	6
	Accuracy	**F-1 score**	**Preci-sion**	**Re-call**	**Minority-F1**	**AUC**	**Approach**	**Ac-curacy**	**F-1 score**	**Preci-sion**	**Re-call**	Minority-F1	AUC
				Constructiveness						**Reciprocity**			
Logistic regression	0.59	0.49	0.27	0.61	0.17	0.63	Logistic regression	0.63	0.59	0.47	0.61	0.4	0.56
K-Nearest neighbors	0.88	0.59	0.24	0.57	0.46	0.16	K-Nearest neighbors	0.75	0.67	0.5	0.66	0.61	0.43
Gaussian naive Bayes	0.63	0.51	0.26	0.6	0.17	0.57	Gaussian naive Bayes	0.53	0.52	0.45	0.57	0.34	0.65
Bernoulli naive Bayes	0.87	0.5	0.06	0.51	0.26	0.04	Bernoulli naive Bayes	0.73	0.61	0.38	0.6	0.6	0.28
Adaboost	0.88	0.47	0	0.5	0	0	Adaboost	0.73	0.59	0.34	0.59	0.61	0.24
Gradient boosting	0.88	0.47	0.01	0.5	0.67	0.01	Gradient boosting	0.75	0.6	0.37	0.6	0.71	0.25
Decision tree	0.85	0.65	0.39	0.66	0.37	0.41	Decision tree	0.77	0.72	0.6	0.72	0.6	0.59
Linear support vector	0.59	0.49	0.26	0.6	0.17	0.6	Linear support vector	0.63	0.59	0.47	0.61	0.4	0.55
C- support vector	**0.73**	0.59	0.36	0.69	0.25	**0.64**	C- support vector	0.7	0.66	0.54	0.67	0.49	0.6
Linear discriminant analysis	0.88	0.47	0	0.5	0	0	Linear discriminant analysis	0.72	0.5	0.17	0.54	0.71	0.1
				Justification						**Empathy & Respect**			
Logistic regression	0.69	0.65	0.54	0.65	0.77	0.76	Logistic regression	0.6	0.6	0.56	0.6	0.66	0.63
K-Nearest neighbors	0.76	0.71	0.59	0.7	0.78	0.88	K-Nearest neighbors	0.68	0.68	0.63	0.68	0.71	0.74
Gaussian naive Bayes	0.7	0.65	0.51	0.64	0.75	0.8	Gaussian naive Bayes	0.63	0.6	0.51	0.6	0.64	0.77
Bernoulli naive Bayes	0.7	0.65	0.52	0.65	0.76	0.82	Bernoulli naive Bayes	0.61	0.6	0.53	0.6	0.65	0.7
Adaboost	0.72	0.63	0.44	0.63	0.73	0.92	Adaboost	0.62	0.6	0.51	0.6	0.64	0.76
Gradient boosting	0.74	0.66	0.49	0.65	0.75	0.92	Gradient boosting	0.65	0.62	0.52	0.63	0.65	0.81
Decision tree	0.74	0.7	0.6	0.7	0.8	0.8	Decision tree	0.7	0.69	0.65	0.69	0.73	0.73

Approach	1	2	3	4	5	6	Approach	1	2	3	4	5	6
	Accuracy	**F-1 score**	**Precision**	**Recall**	**Minority-F1**	**AUC**		**Accuracy**	**F-1 score**	**Precision**	**Recall**	Minority-F1	AUC
Linear support vector	0.69	0.65	0.54	0.65	0.77	0.77	Linear support vector	0.6	0.6	0.55	0.6	0.65	0.63
C- support vector	**0.74**	0.72	0.63	0.72	0.82	**0.8**	C- support vector	0.66	0.65	0.61	0.65	0.7	0.69
Linear discriminant analysis	0.71	0.6	0.39	0.6	0.72	0.93	Linear discriminant analysis	0.62	0.6	0.5	0.6	0.64	0.77
				Relevance						**Incivility**			
Logistic regression	0.77	0.71	0.59	0.74	0.89	0.8	Logistic regression	0.73	0.55	0.27	0.67	0.18	0.59
K-Nearest neighbors	0.82	0.74	0.59	0.72	0.86	0.92	K-Nearest neighbors	0.91	0.61	0.26	0.58	0.49	0.18
Gaussian naive Bayes	0.77	0.7	0.54	0.7	0.86	0.84	Gaussian naive Bayes	0.86	0.59	0.27	0.61	0.23	0.31
Bernoulli naive Bayes	0.78	0.71	0.56	0.71	0.86	0.85	Bernoulli naive Bayes	0.9	0.62	0.29	0.6	0.37	0.24
Adaboost	0.81	0.7	0.51	0.67	0.84	0.93	Adaboost	0.92	0.58	0.19	0.55	0.54	0.12
Gradient boosting	0.82	0.72	0.55	0.7	0.85	0.93	Gradient boosting	0.92	0.57	0.19	0.55	0.64	0.11
Decision tree	0.82	0.75	0.62	0.75	0.88	0.88	Decision tree	0.9	0.67	0.39	0.66	0.4	0.38
Linear support vector	0.77	0.71	0.58	0.74	0.89	0.8	Linear support vector	0.75	0.56	0.28	0.67	0.18	0.57
C- support vector	0.81	0.75	0.64	0.78	0.9	0.84	C- support vector	0.84	0.64	0.37	0.71	0.28	0.56
Linear discriminant analysis	0.79	0.66	0.45	0.64	0.82	0.93	Linear discriminant analysis	0.92	0.57	0.18	0.55	0.53	0.11

Table 4. Internal validation of classifiers trained on the open-vocabulary features of heterogeneous data in a ten-fold cross-validation setup. Scores closer to 1 implies that a greater number of cases were correctly predicted as positive or negative.

Approach	1	2	3	4	5	6	Approach	1	2	3	4	5	6
	Accuracy	**F-1 score**	**Preci-sion**	**Recall**	Minority-F1	AUC		**Ac-curacy**	**F-1 score**	**Preci-sion**	**Recall**	Minority-F1	AUC
	Constructiveness							**Reciprocity**					
Logistic regression	0.95	0.88	0.79	0.88	0.8	0.78	Logistic regression	0.93	0.91	0.88	0.91	0.88	0.87
K-Nearest neighbors	0.88	0.49	0.04	0.51	0.3	0.02	K-Nearest neighbors	0.8	0.71	0.54	0.69	0.86	0.4
Gaussian naive Bayes	0.8	0.68	0.48	0.79	0.35	0.77	Gaussian naive Bayes	0.82	0.81	0.75	0.85	0.63	0.93
Bernoulli naive Bayes	0.89	0.78	0.63	0.84	0.53	0.78	Bernoulli naive Bayes	0.88	0.85	0.77	0.83	0.86	0.7
Adaboost	0.88	0.51	0.09	0.52	0.45	0.05	Adaboost	0.73	0.6	0.37	0.6	0.58	0.28
Gradient boosting	0.88	0.48	0.02	0.51	0.83	0.01	Gradient boosting	0.75	0.6	0.36	0.6	0.7	0.25
Decision tree	0.9	0.74	0.54	0.73	0.57	0.51	Decision tree	0.79	0.74	0.63	0.73	0.66	0.59
Linear support vector	0.85	0.73	0.55	0.81	0.43	0.74	Linear support vector	0.83	0.8	0.73	0.82	0.67	0.79
C- support vector	0.12	0.11	0.21	0.5	0.12	1	C- support vector	0.21	0.2	0.27	0.29	0.18	0.49
Linear discriminant analysis	0.84	0.7	0.5	0.77	0.4	0.67	Linear discriminant analysis	0.69	0.66	0.54	0.67	0.48	0.62
	Justification							**Empathy & Respect**					
Logistic regression	0.89	0.88	0.84	0.88	0.93	0.9	Logistic regression	0.91	0.91	0.9	0.91	0.92	0.93
K-Nearest neighbors	0.79	0.74	0.62	0.72	0.79	0.93	K-Nearest neighbors	0.84	0.83	0.81	0.83	0.84	0.89
Gaussian naive Bayes	0.83	0.82	0.78	0.86	0.96	0.78	Gaussian naive Bayes	0.9	0.9	0.89	0.9	0.91	0.92
Bernoulli naive Bayes	0.76	0.71	0.58	0.69	0.78	0.89	Bernoulli naive Bayes	0.68	0.65	0.55	0.65	0.67	0.85
Adaboost	0.75	0.7	0.56	0.68	0.77	0.89	Adaboost	0.65	0.62	0.52	0.62	0.65	0.8
Gradient boosting	0.76	0.71	0.58	0.7	0.78	0.9	Gradient boosting	0.66	0.63	0.51	0.63	0.65	0.86
Decision tree	0.82	0.8	0.72	0.79	0.86	0.88	Decision tree	0.72	0.72	0.68	0.72	0.75	0.77

Approach	1	2	3	4	5	6	Approach	1	2	3	4	5	6
	Accuracy	**F-1 score**	**Preci-sion**	**Recall**	Minority-F1	AUC		**Ac-curacy**	**F-1 score**	**Preci-sion**	**Recall**	Minority-F1	AUC
Linear support vector	0.8	0.79	0.73	0.8	0.89	0.8	Linear support vector	0.81	0.81	0.79	0.81	0.84	0.82
C- support vector	0.33	0.25	0.5	0.5	0	0	C- support vector	0.44	0.3	0.61	0.5	0	0
Linear discriminant analysis	0.84	0.82	0.76	0.82	0.88	0.89	Linear discriminant analysis	0.73	0.73	0.7	0.73	0.77	0.75
	Relevance							**Incivility**					
Logistic regression	0.94	0.92	0.88	0.92	0.96	0.97	Logistic regression	0.96	0.86	0.75	0.86	0.77	0.73
K-Nearest neighbors	0.8	0.59	0.31	0.59	0.8	0.99	K-Nearest neighbors	0.92	0.54	0.13	0.53	0.74	0.07
Gaussian naive Bayes	0.87	0.84	0.77	0.88	0.97	0.85	Gaussian naive Bayes	0.83	0.62	0.35	0.7	0.26	0.55
Bernoulli naive Bayes	0.83	0.77	0.64	0.76	0.89	0.89	Bernoulli naive Bayes	0.79	0.64	0.41	0.83	0.27	0.89
Adaboost	0.84	0.77	0.64	0.76	0.88	0.91	Adaboost	0.92	0.61	0.26	0.58	0.52	0.17
Gradient boosting	0.85	0.8	0.69	0.79	0.9	0.91	Gradient boosting	0.92	0.56	0.17	0.55	0.68	0.1
Decision tree	0.89	0.84	0.76	0.84	0.92	0.93	Decision tree	0.94	0.77	0.58	0.74	0.66	0.51
Linear support vector	0.86	0.82	0.74	0.85	0.94	0.87	Linear support vector	0.91	0.77	0.59	0.84	0.48	0.76
C- support vector	0.24	0.19	0.38	0.5	0	0	C- support vector	0.91	0.48	0.01	0.5	0.06	0.01
Linear discriminant analysis	0.91	0.87	0.8	0.86	0.93	0.96	Linear discriminant analysis	0.94	0.8	0.64	0.81	0.64	0.65

Table 5. External validation: A comparison of the predictive performance (accuracy and macro-F1) of the lexica developed from homogeneous, heterogeneous, open- and closed-vocabulary setups on other datasets. A value closer to 1 implies that a greater proportion of cases were correctly classified.

Predictive performance (Accuracy)								
Facet	Training set	Features	**Halpern (2013)**	**Gibbs**	**Fournier-Tombs & Di Marzo Serugendo (2019)**	**Theocharis (2016) et al.**	**Stromer-Galley (2007)**	**Average**
Constructiveness	Homogeneous	closed-vocabulary	0.31		0.35	0.35	0.87	0.47 (0.27)
	Heterogeneous	closed-vocabulary	0.68		0.64	0.63	0.59	**0.64 (0.04)**
		open-vocabulary	0.34		0.39	0.39	0.3	0.36 (0.04)
Justification	Homogeneous	closed-vocabulary	0.53		0.52	0.61	0.21	0.47 (0.18)
	Heterogeneous	closed-vocabulary	0.79		0.84	0.75	0.89	**0.82 (0.06)**
		open-vocabulary	0.74		0.74	0.66	0.82	0.74 (0.07)
Relevance	Homogeneous	closed-vocabulary	0.24		0.39	0.24	0.54	0.35 (0.14)
	Heterogeneous	closed-vocabulary	0.85		0.85	0.74	0.23	0.67 (0.3)
		open-vocabulary	0.72		0.79	0.66	0.62	**0.70 (0.07)**
Reciprocity	Homogeneous	closed-vocabulary	0.38		0.36	0.38	0.58	0.43 (0.10)
	Heterogeneous	closed-vocabulary	0.62		0.36	0.38	0.34	0.43 (0.13)
		open-vocabulary	0.49		0.42	0.51	0.46	**0.47 (0.04)**
Empathy & Respect	Homogeneous	closed-vocabulary	0.31		0.15	0.23	0.36	0.26 (0.09)
	Heterogeneous	closed-vocabulary	0.45		0.37	0.55	0.59	0.49 (0.10)
		open-vocabulary	0.68		0.85	0.76	0.16	**0.61 (0.31)**
Incivility	Homogeneous	closed-vocabulary	0.74		0.79	0.79	0.95	**0.82 (0.09)**
	Heterogeneous	closed-vocabulary	0.74		0.79	0.79	0.7	0.76 (0.04)
		open-vocabulary	0.73		0.76	0.79	0.82	0.79 (0.03)

Predictive performance (Macro-F1)

Facet	Training set	Features	**Halp-ern (2013)**	**Gibbs**	**Fournier-Tombs & Di Marzo Serugendo (2019)**	**Theocharis (2016) et al.**	**Stromer-Galley (2007)**	**Average**
Constructiveness	Homogeneous	closed-vocabulary	0.25		0.27	0.3	1	0.46 (0.36)
	Heterogeneous	closed-vocabulary	0.41		0.41	0.43	0.42	**0.42 (0.01)**
		open-vocabulary	0.32		0.35	0.39	0.24	0.33 (0.06)
Justification	Homogeneous	closd-vocabulary	0.5		0.52	0.49	0.19	0.43 (0.16)
	Heterogeneous	closed-vocabulary	0.46		0.84	0.48	0.47	**0.56 (0.19)**
		open-vocabulary	0.53		0.52	0.49	0.49	0.51 (0.02)
Relevance	Homogeneous	closed-vocabulary	0.24		0.55	0.19	0.49	0.37 (0.18)
	Heterogeneous	closed-vocabulary	0.49		0.85	0.49	0.32	**0.54 (0.22)**
		open-vocabulary	0.52		0.57	0.53	0.49	0.53 (0.03)
Reciprocity	Homogeneous	closed-vocabulary	0.51		0.49	0.51	0.78	**0.57 (0.14)**
	Heterogeneous	closed-vocabulary	0.51		0.47	0.37	0.41	0.44 (0.06)
		open-vocabulary	0.42		0.49	0.51	0.46	0.47 (0.04)
Empathy & Respect	Homogeneous	closed-vocabulary	0.42		0.21	0.32	0.48	0.36 (0.12)
	Heterogeneous	closed-vocabulary	0.45		0.37	0.49	0.42	**0.43 (0.05)**
		open-vocabulary	0.41		0.47	0.44	0.22	0.39 (0.11)
Incivility	Homogeneous	closed-vocabulary	0.85		0.88	0.88	0.97	0.93 (0.07)
	Heterogeneous	closed-vocabulary	1		0.79	1	0.95	**0.94 (0.1)**
		open-vocabulary	0.43		0.45	0.95	0.45	0.57 (0.25)

Table 6. Illustrative examples of errors (positive cases marked negative) across the training set and the external validation datasets. No examples are provided where no false negatives were reported.

Constructiveness	
Twitter Politics Deliberation	@USER Let me rephrase your tweet: "government can't take over if citizens have guns. They're a threat. You're welcome… Idiot?"
Reddit CMV Politics	-
Trivium	-
Halpern & Gibbs (2013)	There really is no winning with some of you. He acted to fast, he acted to slow, he didn't do anything, he's going too far, he's not going far enough.
Fournier-Tombs & Di Marzo Serugendo (2019)	toi en estrie nous en ostie ….hien! pis ca fait quoi etre le larbin de Soros …ton boss y?? dans la mire de trump…et ca fait quoi etre sous enquete de la v??rificatrice pour conflit d'interet …moi ca serait pour trahison EN PRISON LE TRAI
Theocharis et al. (2016)	@Nigel*F aragef ingerscrossednigel*.
Stromer-Galley (2007)	Also, it's obvious that everyone disagrees probably with the same agreement that the middle schools having such chaos, that we need to follow some other models that have worked and if we do studies and see that the private sector
Justification	
Twitter Politics Deliberation	@USER Why no seams of sexism? Hypocrisy! https://t.co/IvIKYbWYfy
Reddit CMV Politics	I can respond to this with an anecdote from my family. My great Uncle was one of the main physicists for the Manhattan Project.
Trivium	It may cost a lot to bulk up on security but since almost 1/5 of our budget is going to national defense, close to $700 billion, it seems logical to use this money here.
Halpern & Gibbs (2013)	this guy has now appointed an "oil commision" filled with eco nazis and even someone from "National Geograpic"..????
Fournier-Tombs & Di Marzo Serugendo (2019)	Federal regulations on MEP dodgers….
Theocharis et al. (2016)	If @Nigel*F arage@U KIPhadtheirwaywefdbeoutof it*!*N ow, THAT couldswayme*!
Stromer-Galley (2007)	(…) my fear that is that so many people will want their children to go to what are the so called best schools that it would be discriminatory.

Twitter Politics Deliberation	@USER Why no seams of sexism? Hypocrisy! https://t.co/lvIKYbWYfy
Reddit CMV Politics	I can respond to this with an anecdote from my family. My great Uncle was one of the main physicists for the Manhattan Project.
Trivium	It may cost a lot to bulk up on security but since almost 1/5 of our budget is going to national defense, close to $700 billion, it seems logical to use this money here.
Halpern & Gibbs (2013)	this guy has now appointed an “oil commision” filled with eco nazis and even someone from “National Geograpic”..????
Fournier-Tombs & Di Marzo Serugendo (2019)	Federal regulations on MEP dodgers….
Theocharis et al. (2016)	If @Nigel*F arage@U KIPhadtheirwaywefdbeoutof it*!*N ow, THAT couldswayme*!
Stromer-Galley (2007)	(…) my fear that is that so many people will want their children to go to what are the so called best schools that it would be discriminatory.
Relevance	
Twitter Politics Deliberation	@USER I cant believe the stock market hasnt crashed yet. This is huge. Worse than nixon.
Reddit CMV Politics	This comes off as incredibly naive. In a lot of situations, an officer cannot always take cover and wait for the suspect to run out of ammunition.
Trivium	Students should learn in school not dodge bullets. Teacher should not be armed. And we should have strong gun laws.
Halpern & Gibbs (2013)	didnt think you could comment on a government channel
Fournier-Tombs & Di Marzo Serugendo (2019)	Loser PM, please go away
Theocharis et al. (2016)	@USER @USER @USER are welfare, overseas aid, EU contribution NHS, pensions, MPs
Stromer-Galley (2007)	I believe that it will consolidate schools and that m.s. will not remain open (…)
Reciprocity	
Twitter Politics Deliberation	@USER Doesn’t anyone read the actual bill on which they are voting? BO could explain every detail of the ACA. He actually gave a shit (or two.)
Reddit CMV Politics	-
Trivium	Heard they have few issues with that. Not sure about the concealed situation there.
Theocharis et al. (2016)	@USER gets my vote He goes above and beyond the call of duty for his country @USER
Halpern & Gibbs (2013)	For every chemtrail I see, I’m going to try and make a baby. I encourage everyone I talk to do the same. Every time you see chemtrails go have sex!
Fournier-Tombs & Di Marzo Serugendo (2019)	-
Stromer-Galley (2007)	(…) and I’d like to know what other people think about this.

Empathy & Respect	
Twitter Politics Deliberation	@USER This is a rigged battle. It is up to you to unrig it. You are our representatives in government. This is no laughing matter anymore.
Trivium	USER4987, I agree.
Reddit CMV Politics	Violent, sexist and homophobic lyrics preaching materialism isn't exactly the best advertisement for what these people perceive as "black culture"
Theocharis et al. (2016)	@USER @USER @USER I've voted Tory and put a tick against the UKIP box
Halpern & Gibbs (2013)	I just had a thought and it worries me. What if space men are watching and see Obama and know he is the leader of the greatest nation on earth.
Fournier-Tombs & Di Marzo Serugendo (2019)	I like PM Trudeau, he is so simple like Obama
Stromer-Galley (2007)	I also wanted to say to Michael that I'm also intrigued by the idea of slc in h.s., and you mentioned the idea that students with similar interests might for example self select a learning community within a h.s., which sounds very interesting, but (…)
Incivility	
Twitter Politics Deliberation	@USER Cardin is now & always has been a liar. Makes me wonder why he does not want tax cut.ms. Political party more important than integrity?
Reddit CMV Politics	-
Trivium	-
Theocharis et al. (2016)	-
Halpern & Gibbs (2013)	-
Fournier-Tombs & Di Marzo Serugendo (2019)	-
Stromer-Galley (2007)	-

Table 7. Explicating positive correlations between incivility and other discourse quality facets in the Twitter Political Deliberation dataset.

Cases marked as Constructive and Uncivil
• @USER They love anyone who hates America as much as them. It's crazy that they can hate they country that made them rich so much. Robbing us is what they do best sadly. • @USER your A two faced liar- go kiss soros butt, You are A Traitor-You need to leave your position • @USER the GOP IS A COMPLICIT SHIT SHOW! History will remember you as greedy old men who sold this country to the Russians and rich corporations. Kiss your political careers goodbye!

Cases marked as Reciprocal and Uncivil
• @USER- #GrahamCassidy will devastate #MilitaryFamilies w/ kids like Justin who need #Medicaid. Pls vote no! <LINK> • @USER- The US people & Minnesotans must see the Senate Ethics investigation committee hearing: the womens' allegations and Senator Al Franken's responses. Dems have Ethics but lose 1 FINE Senator! Creepy Reps support 1 more sex assaulter to Senate. Explain the Math?????? • @USER Trump ran on doing what he did he/Repub in congress won. GOP tone deaf to what the people said in 2016 800K vs 62M

Notes

1. This paper discusses findings based on a dataset in English. The methods, however, are generalizable to other languages and contexts.
2. Political discussions were identified based on whether or not the original post comprised any one of a list of political keywords curated by researchers in previous work (Preoţiuc-Pietro, Liu, Hopkins, & Ungar, 2017)
3. https://github.com/makcedward/nlpaug
4. An explanation of these approaches is provided in the Supplementary Materials at https://doi.org/10.17605/OSF.IO/28ESD.
5. https://doi.org/10.17605/OSF.IO/28ESD
6. The complete results with all the classifiers and feature sets are reported in the supplementary materials at https://doi.org/10.17605/OSF.IO/28ESD.
7. The detailed macro-F1 metrics for all the classifiers are reported in the supplementary materials at https://doi.org/10.17605/OSF.IO/28ESD.
8. They are available at https://github.com/kj2013/deliberative-politics

References

Baxter, L. A. (2006). Communication as dialogue. *GJ Shepherd, J. St. John, & TG Striphas (Eds.), Communication as—: Perspectives on theory*, 101–109.

Beauchamp, N. (2020). Modeling and measuring deliberation online. In *The oxford handbook of networked communication.*

Berry, J. M., & Sobieraj, S. (2013). *The outrage industry: Political opinion media and the new incivility*. Oxford, UK: Oxford University Press.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Cham, Switzerland: Palgrave Macmillan.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (*Volume 1: Long Papers*) (pp. 250–259). Sofia, Bulgaria: Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (pp. 512–515).

Dobbrick, T., Jakob, J., Chan, C.-H., & Wessler, H. (2021). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures*, 1–18.

Erlich, A., Dantas, S. G., Bagozzi, B. E., Berliner, D., & Palmer-Rubin, B. (2021). Multi-label prediction for political text-as-data. *Political Analysis*, 1–18.

Esteve Del Valle, M., Sijtsma, R., & Stegeman, H. (2018). Social media and the public sphere in the Dutch parliamentary Twitter network: A space for political deliberation? Hamburg, Germany: ECPR General Conference.

Fournier-Tombs, E., & Di Marzo Serugendo, G. (2019). Delibanalysis: Understanding the quality of online political discourse with machine learning. *Journal of Information Science*, 0165551519871828.

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7* (3), 319–339. doi: 10.1002/poi3.95

Gastil, J. (2008). *Political communication and deliberation*. Los Angeles, CA: SAGE Publications.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21* (3), 267–297.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* , *29* (3), 1159–1168. doi: 10.1016/j.chb.2012.10.008

Himmelroos, S. (2017). Discourse quality in deliberative citizen forums – A comparison of four deliberative mini-publics. *Journal of Public Deliberation*, *13* (1), Article 3.

Jaidka, K. (2022, June). Developing a multilabel corpus for the quality assessment of online political talk. In *Proceedings of the language resources and evaluation conference* (pp. 5503–5510). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.589

Jaidka, K., Zhou, A., & Lelkes, Y. (2019). Brevity is the soul of twitter: The constraint affordance and political discussion. *Journal of Communication*, *69* (4), 345–372.

Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J., & Lecheler, S. (2022). Beyond anonymity: Network affordances, under deindividuation, improve social media discussion quality. *Journal of Computer-Mediated Communication*, *27* (1), zmab019.

Jakob, J., Dobbrick, T., Freudenthaler, R., Haffner, P., & Wessler, H. (2022). Is constructive engagement online a lost cause? toxic outrage in online user comments across democratic political systems and discussion arenas. *Communication Research*, 00936502211062773.

Jakob, J., Dobbrick, T., & Wessler, H. (2021). The integrative complexity of online user comments across different types of democracy and discussion arenas. *The International Journal of Press/Politics*, 19401612211044018.

Janssen, D., & Kies, R. (2005). Online forums and deliberative democracy. *Acta Politica*, *40* (3), 317–335. doi: 10.1057/palgrave.ap.5500115

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201* .

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Liao, W., Zhang, J., Oh, Y. J., & Palomares, N. A. (2021). Linguistic accommodation enhances compliance to charity donation: The role of interpersonal communication processes in mediated compliance-gaining conversations. *Journal of Computer-Mediated Communication*.

Maia, R. C. M., & Rezende, T. A. S. (2016). Respect and disrespect in deliberation across the networked media environment: Examining multiple paths of political talk. *Journal of Computer-Mediated Communication*, *21* (2), 121–139. doi: 10.1111/jcc4.12155

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116* (20), 9785–9789.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16* (4), 372–403.

Monroe, B. L., & Schrodt, P. A. (2008). Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, *16* (4), 351–355.

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2018). (Re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*. doi: 10.1080/10584609.2018.1517843

Niculae, V., Kumar, S., Boyd-Graber, J., & Danescu-Niculescu-Mizil, C. (2015). Linguistic harbingers of betrayal: A case study on an online strategy game. *arXiv preprint arXiv:1506.04744* .

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, *20* (9), 3400–3419. doi: 10.1177/1461444817749516

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., &Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12* , 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015* (Tech. Rep.). Austin, TX: University of Texas at Austin.

Peskov, D., Cheng, B., Elgohary, A., Barrow, J., Danescu-Niculescu-Mizil, C., & Boyd-Graber, J. (2020). It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3811–3854).

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 729–740).

Rinke, E. M. (2015). Mediated deliberation. *The International Encyclopedia of Political Communication*.

Rowe, I. (2015). Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of broadcasting & electronic media*, *59* (4), 539–555.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., & others (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8* (9), e73791.

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, *1* (1), 21–48. doi: 10.1057/palgrave.cep.6110002

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation*, *3* (1), Article 12.

Stromer-Galley, J., & Martinson, A. M. (2009). Coherence in political computer-mediated communication: Analyzing topic relevance and drift in chat. *Discourse & Communication*, *3* (2), 195–216. doi: 10.1177/1750481309102452

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication*, *20* (2), 188–203. doi: 10.1111/jcc4.12104

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66* (6), 1007–1031. doi: 10.1111/jcom.12259

Wessler, H. (2008). Investigating deliberativeness comparatively. *Political communication*, *25* (1), 1–22.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* .

References supplementary materials

Fournier-Tombs, E., & Di Marzo Serugendo, G. (2019). Delibanalysis: Understanding the quality of online political discourse with machine learning. *Journal of Information Science*, 0165551519871828.

Groshek, J., & Cutino, C. (2016). Meaner on mobile: Incivility and impoliteness in communicating contentious politics on sociotechnical networks. *Social Media + Society*, *2* (4), 1–10. doi: 10.1177/2056305116677137

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* , *29* (3), 1159–1168. doi: 10.1016/j.chb.2012.10.008

Jaidka, K., Zhou, A., & Lelkes, Y. (2019). Brevity is the soul of twitter: The constraint affordance and political discussion. *Journal of Communication*, *69* (4), 345–372.

Maia, R. C. M., & Rezende, T. A. S. (2016). Respect and disrespect in deliberation across the networked media environment: Examining multiple paths of political talk. *Journal of Computer-Mediated Communication*, *21* (2), 121–139. doi: 10.1111/jcc4.12155

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12* , 2825–2830.

Rinke, E. M. (2015). Mediated deliberation. *The International Encyclopedia of Political Communication*.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*. doi: 10.1177/0093650220921314

Stromer-Galley, J., & Martinson, A. M. (2009). Coherence in political computer-mediated communication: Analyzing topic relevance and drift in chat. *Discourse & Communication*, *3* (2), 195–216. doi: 10.1177/1750481309102452

Wessler, H. (2008). Investigating deliberativeness comparatively. *Political communication*, *25* (1), 1–22.