Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5503–5510
Marseille, 20-25 June 2022

Developing A Multilabel Corpus for the Quality Assessment of Online Political Talk

Kokil Jaidka
Department of Communications and New Media
11 Computing Way
National University of Singapore
jaidka@nus.edu.sg

Abstract

This paper motivates and presents the Twitter Deliberative Politics dataset, a corpus of political tweets labeled for its deliberative characteristics. The corpus was randomly sampled from replies to US congressmen and women. It is expected to be useful to a general community of computational linguists, political scientists, and social scientists interested in the study of online political expression, computer-mediated communication, and political deliberation. The data sampling and annotation methods are discussed and classic machine learning approaches are evaluated for their predictive performance on the different deliberative facets. The paper concludes with a discussion of future work aimed at developing dictionaries for the quality assessment of online political talk in English. The dataset and a demo dashboard are available at https://github.com/kj2013/twitter-deliberative-politics.

Keywords: deliberation, justification, political discussions, comments, Twitter, Facebook

1. Introduction

Social media offers an ideal public sphere for citizens to engage in politics because of its inclusivity, accessibility, equality, and extensive outreach. Political discussions on Twitter and other online platforms allow people with diverse perspectives and opinions to participate in casual conversation. However, the volume, variety, and velocity of social media posts often make hand-annotation an intractable solution to understand online political discussion dynamics. The Discourse Quality Index (Steenbergen et al., 2003) is an instrument to measure political discussion quality. However, so far it has been limited to the quantitative analysis of small samples of hand-annotated posts. On the other hand, a study of the political engagement, civic communities, echo chambers, and their dynamics requires a language quality instrument which can scale to annotate thousands of political social media posts. To address this research gap, this study reports on a new corpus of political tweets, labeled with the different facets of the discussion quality. The dataset offers a first step in building dictionaries to aid in the automatic measurement of the Discourse Quality Index.

An emerging body of work in political communication has proposed methods that allow for automated content analysis through dictionaries (Gründl, 2020), classifiers (Barberá et al., 2015) and hybrid methods (Baden et al., 2020), while also discussing the possibilities and limitations of using supervised machine learning approaches as a research method (Pilny et al., 2019; Burscher et al., 2014). These methods and approaches mainly focus on frames, emotions, and incivility (Muddiman et al., 2018) or politeness (Niculae et al., 2015). A measure of political discussion quality is scarce, and prior attempts suggest that they would not generalize well across platforms (Nelimarkka and Ahonen, 2019; Stroud et al., 2015; Halpern and Gibbs, 2013). In addressing this gap, this article contributes with the first in what is to be a diverse set of online political comments labeled for their discussion quality characteristics. It is hoped that a diverse dataset will overcome of the limitations identified in prior work regarding the generalizability of linguistic measurements across platforms. Therefore, the main contributions of this article are:

- A corpus of tweets about American politics (5585 observations) that can assist the study of online political discussions from a social science perspective.
- Labels of discussion quality on the corpus, which makes it applicable to apply multilabel and multiclass approaches in machine learning approaches by computational linguists.
- A report on the evaluation of classic machine learning approaches to predict the labels on held-out data in ten-fold cross-validation.

2. Related work

Supervised methods to study the quality of online political posts have largely focused on their civility, politeness, and use of partisan language. A review of the papers on developing and applying abusive language classifiers is reported in (Fišer et al., 2018). A few studies have reported on discussion quality facets, such as politeness (Danescu-Niculescu-Mizil et al., 2013) and empathy (Buechel et al., 2018). A challenge to applying these pre-trained classifiers to the study of political discussions is that they are usually trained on a dataset of general social media posts. Therefore, they may miss out on many implicit signals of discussion quality in political talk. They may also wrongly attribute quality facets where they are none. For example, while ‘snowflakes’ and ‘red hats’ may be innocuous in general language, they take on a special nuance in post-2016 America as a liberal and conservative voter stereotype, respectively. Furthermore, using the word ‘liberal’ may be associated with politeness in general usage, while it is used to indicate an ideological stance in political discussions unambiguously.

On the other hand, studies of the quality of online political talk report on smaller hand-annotated datasets, which offer empirical evidence in support of the differences in how communities and platforms engage in political discussions. The prior work which has thus studied online political talk is grounded in the Habermasian ideas of a deliberative public sphere (Habermas, 1984; Gastil, 2008) in terms of the necessary and sufficient criteria of public deliberation: the presence of language that builds consensus, argues stances and provides evidences, invites responses, and expresses empathy towards others (Stromer-Galley, 2007; Steenbergen et al., 2003; Esteve Del Valle et al., 2018; Friess and Eilders, 2015). This article draws from the annotation instructions and findings of prior work and builds on them to construct an annotation task aimed at annotating the analytical and social aspects of political discussions, i.e., the use of justification, constructiveness, and relevance(Friess and Eilders, 2015), separate from its social aspects, i.e., reciprocity, empathy and respect, and incivility (Steenbergen et al., 2003). The annotation instructions focused on defining and exemplifying the following deliberative characteristics of online political posts:

- **Constructiveness:** Reflects an intention to move the conversation forward, build and bring about consensus, and resolve conflicts by pointing out facts, identifying common ground, or proposing solutions.
- **Justification:** Reflects an intention to offer a justification, either based on personal experiences, values, and feelings or data, links, and facts.
- **Relevance:** Reflects whether the post is relevant to politics.
- **Reciprocity:** A post that asks a genuine question, or a comment intended to elicit a response or further information.
- **Empathy & Respect:** Reflects the author’s acknowledgment of or sensitivity to others, manifested in positive comments, an empathetic or a respectful response acknowledging other viewpoints.
- **Incivility:** Abusive, racist, threatening, or exaggerative behavior.

Observations from the dataset exemplifying each of the deliberative facets are reported in Table 1.

Table 1: Examples of cases marked positive in the Twitter Deliberative Politics dataset.

Constructiveness
• @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.
Justification
• @USER #morningjoe @USER @USER Aft Sen <name> mtg confirmed what we all KNEW: “I didn’t expect an epiphany”! Yeah, he be • @USER #DREAMers are doctors, lawyers, teachers, social workers, friends, and colleagues. Time for Congress to protect them. • @USER #Democrats @USER should not be running the budget for a bat mitzvah and @USER is unfit to run a book club. #CommieZombie traitors https://<LINK>
Reciprocity
• @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?
Empathy & Respect
• @USER I now know who I won’t support • @USER Don’t let this bill take any deductions away from us(...). Thank you! • @USER #HandsOff People with disabilities will be hurt more than those without by these bills. Vote them down.
Incivility
• @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

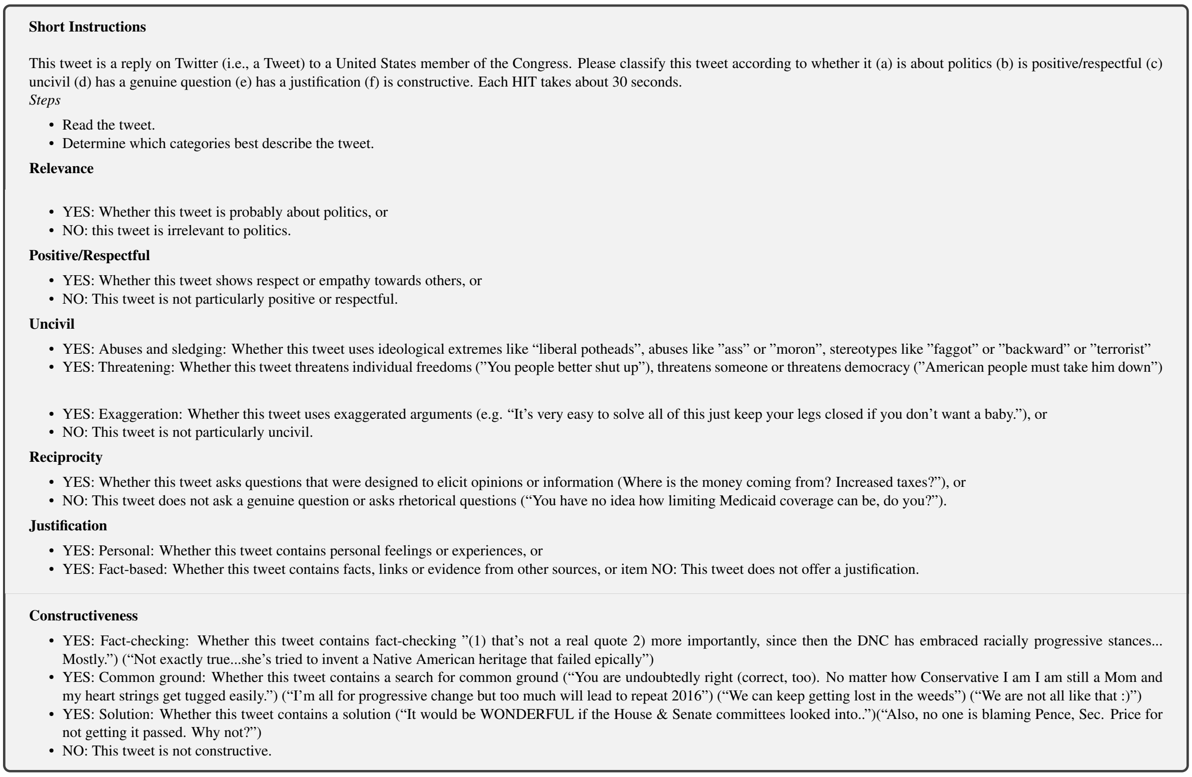
3. Data

3.1. Sampling and annotation

Tweets from Twitter’s 1% sample between January 2017 - March 2018 were filtered to retain the replies to 536 US Congressmen and Congresswomen holding office. The Python Natural Language Tool Kit package was first applied to retain English tweets. Then, a random sample of 6000 English tweets of at least 10 characters in length was drawn for the purpose of annotation.

Amazon Mechanical Turk was used to recruit residents of the United States and with a minimum approval rate of 80%. Annotators were trained with detailed instructions and numerous examples before they were able to work on the labeling task. The instructions are provided in Table 2. Four annotations per tweet were collected from 564 workers who participated in this task for one week in March 2019. The inter-annotator reliability results of the final training set are provided in Table 3. The columns provided the pairwise percent-

Table 2: 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?”).

Justification

- YES: Personal: Whether this tweet contains personal feelings or experiences, or
- YES: Fact-based: Whether this tweet contains facts, links or evidence from other sources, or item NO: This tweet does not offer a justification.

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.

age agreement, and an average percentage agreement of 64.2% was obtained.

Table 3: Descriptive information and inter-coder reliability statistics for the Twitter Deliberative Politics dataset.

	Positive Instances	Pairwise % Agree
Constructiveness	531	59.3
Justification	67	64.7
Relevance	953	79.3
Reciprocity	4689	62.2
Empathy & Respect	2842	61.3
Incivility	382	58.5
Average agreement		**64.2**

The low percentage agreement is a cause for concern and highlights how subjective the annotation task truly is. The sources of error are discussed in the Results section, and potential directions for improvement are discussed towards the end of this article.

3.2. Data selection and augmentation

In keeping with best practices for text classification experiments with hand-annotated data (Danescu-Niculescu-Mizil et al., 2013; Davidson et al., 2017), only the labels with at least 75% agreement (which constituted 64.8% of all labels on 5585 observations) were subsequently used in training and testing the machine learning classifiers.

Table 3 suggests a class imbalance, which can adversely affect the performance of trained classifiers. Therefore, the data was augmented using the HuggingFace libraries (Wolf et al., 2019). First, the data was augmented using back-translation into and from Spanish, by applying the pre-trained Neural Machine Translation models created by Helsinki-NLP. These models are trained on news text, which were anticipated to approximate the vocabulary of political tweets. Next, the vocabulary of the corpus was further expanded by using the contextual word embeddings model (Kobayashi, 2018) available through the *nlp.aug* package[1]. In this manner, the data was augmented to triple its size.

3.3. Feature extraction

Before feature extraction, the dataset was anonymized to substitute user handles with a generic “<USER>” string. Links were also substituted with “<LINK>.” Then, the term frequency-inverse document frequency (TFIDF) features were extracted from the augmented data. One-word and two- and three-word phrases were extracted, converted into a frequency distribution, and weighted according to their importance in the message and its uniqueness in the overall dataset by calculating the product of their term frequency and the inverse of

[1] https://github.com/makcedward/nlpaug

their frequency over all the messages (their document frequency). Only the top 10,000 features were retained to avoid overfitting the model to sparse features.

3.4. Model training

Logistic regression classifiers were trained to predict the constructiveness, justification, relevance, empathy/respectfulness, reciprocity, and incivility based on the presence or absence of linguistic features (words and phrases) in each tweet. The frequency distribution of the TFIDF features of the labeled tweets were the independent variables, and the labels about the presence or absence of each facet were the dependent variable.

A deliberate choice was made to only evaluate classic machine learning approaches in these experiments. This was done in keeping with the ultimate aim of corpus development – to develop interpretable and explainable dictionary measures for the quality assessment of online political talk. Furthermore, recent work by prominent scholars comparing classic and neural network architectures in English (Tuggener et al., 2020) and other languages (Septiandri et al., 2020) has suggested that logistic regression may fare as well as or better than neural network approaches for predictive modeling with modest-sized samples. This is likely because classic machine learning approaches appear to benefit from picking out associations among linguistic features with fewer training examples, while transformer based approaches appear to struggle to learn these associations when there is a class-imbalance (Tuggener et al., 2020).

Ten classification approaches available with scikitlearn were evaluated, using the L2 penalty where available (Pedregosa et al., 2011). These were 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.

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 L2 penalty and a maximum of 100 iterations in the next step. 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. Finally, the face validity of the logistic regression classifiers was evaluated by visualizing their features in terms of their importance to the final label prediction.

4. Results

4.1. Sources of inter-annotator disagreement

An inspection of the sources of inter-annotator disagreement is helpful to identify areas of improvement in the annotation instructions and the annotator training. Table 5 exemplifies instances where the percentage agreement between annotators was 50%.

Some patterns do emerge, and suggest that annotators are more likely to disagree across all the labels when the authors use sarcasm or irony. Sarcasm was a particular challenge in labeling constructiveness (*@USER more evidence to support your letter@USER come and deny it as usual. @USER @USER @USER https://<LINK>*).

A second challenge was the use of irony (*Lives at risk b/c he doesn't like being so low in the polls and you're quoting the Bible?, @USER - more evidence to support your letter @USER come and deny it as usual. @USER @USER @USER https://<LINK>*).

Beyond these, there are other challenges specific to each labeling task, Firstly, annotations for constructiveness appear to suffer when users share information or demands but do not provide evidence for it (*@USER #25thAmendment Now before he gets us all killed, @USER $9,900 from the NRA during the 2016 election cycle.*). Annotating justification is challenging when evidence appears to have been supplied as a link at the end of the tweet, but is not explicitly mentioned in the text (*@USER quite sad actually seeing a lot of great work being undermined. @USER @USER https://<LINK>, @USER Subpoena away! Personally I hope that POS perjures him-self before the committee, being he thinks he's smarter than you.https://<LINK>*). Similarly, relevance labeling is challenging when the context was implicit, so the comment could have been irrelevant or relevant to politics (*hypocrite. You are A porn surfer and claim to be holier than thou*). On the other hand, reciprocity labeling is challenging because the author often directs their comment as a personal attack or request to a person, but does not appear to expect a reply (*@USER You (supposedly) work for US, not The other way around.You are supposed to act in our interests not The RINOs like you.*), or may ask a rhetorical question (*@USER doesn't@USER brother work for the DOJ????? Corruption!!!!!! https://<LINK>*). As previously mentioned, annotators were unsure when labeling sarcastic tweets as empathetic, because authors couched their contention in particularly polite language (*@USER - why don't you retweet this? Oh... wait... you're a partisan hack. That's right. https://<LINK>*). Finally, labeling incivility gets challenging when the authors appear to "shout" through the use of capital letters () (*@USER Before this goes any further I want NAMES of pols who PAID OFF accusers w taxpayer money! #IMWITHAL*), or include evidence in their name-calling (*@USER "Free and Open Internet" is nothing more than a code word for corporate land-*

Table 4: Predictive performance of classifiers trained on the TFIDF features of the Twitter Deliberative Politics dataset 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.

	1	2	3	4	5	6		1	2	3	4	5	6
Approach	**Accuracy**	**Macro F-1**	**Minority-F1**	**AUC**	**Recall**	**Precision**	**Approach**	**Accuracy**	**Macro F-1**	**Minority-F1**	**AUC**	**Recall**	**Precision**
			Constructiveness							**Reciprocity**			
Logistic regression	0.82	0.72	0.55	0.74	0.6	0.51	**Logistic regression**	0.78	0.76	0.69	0.77	0.72	0.66
K-Nearest neighbors	0.83	0.58	0.26	0.57	0.16	0.64	K-Nearest neighbors	0.68	0.6	0.41	0.6	0.34	0.54
Gaussian naive Bayes	0.73	0.65	0.49	0.72	0.69	0.38	Gaussian naive Bayes	0.67	0.66	0.61	0.7	0.79	0.5
Bernoulli naive Bayes	0.81	0.62	0.35	0.61	0.28	0.48	Bernoulli naive Bayes	0.77	0.72	0.6	0.71	0.51	0.72
Adaboost	0.81	0.51	0.12	0.53	0.07	0.47	Adaboost	0.7	0.58	0.35	0.59	0.24	0.65
Gradient boosting	0.82	0.46	0.01	0.5	0.01	0.67	Gradient boosting	0.71	0.54	0.26	0.57	0.16	0.8
Decision tree	0.84	0.74	0.58	0.74	0.58	0.58	Decision tree	0.74	0.7	0.58	0.69	0.54	0.63
Linear support vector	0.75	0.66	0.48	0.7	0.62	0.39	Linear support vector	0.72	0.7	0.62	0.71	0.68	0.57
C- support vector	0.88	0.75	0.57	0.71	0.44	0.81	C- support vector	0.8	0.75	0.65	0.74	0.55	0.79
Linear discriminant analysis	0.69	0.59	0.39	0.63	0.54	0.31	Linear discriminant analysis	0.64	0.62	0.53	0.63	0.61	0.46
			Justification							**Empathy & Respect**			
Logistic regression	0.79	0.7	0.53	0.73	0.83	0.9	Logistic regression	0.73	0.7	0.62	0.71	0.76	0.82
K-Nearest neighbors	0.78	0.61	0.35	0.6	0.89	0.84	K-Nearest neighbors	0.69	0.6	0.41	0.6	0.87	0.72
Gaussian naive Bayes	0.76	0.69	0.54	0.75	0.77	0.92	Gaussian naive Bayes	0.63	0.63	0.6	0.68	0.54	0.86
Bernoulli naive Bayes	0.77	0.65	0.44	0.66	0.84	0.87	Bernoulli naive Bayes	0.72	0.66	0.52	0.65	0.85	0.76
Adaboost	0.8	0.48	0.08	0.51	0.99	0.81	Adaboost	0.68	0.5	0.21	0.54	0.95	0.68
Gradient boosting	0.81	0.45	0.01	0.5	1	0.81	Gradient boosting	0.67	0.44	0.08	0.52	0.99	0.67
Decision tree	0.79	0.65	0.42	0.64	0.89	0.86	Decision tree	0.7	0.65	0.52	0.65	0.81	0.76
Linear support vector	0.72	0.64	0.47	0.69	0.74	0.89	Linear support vector	0.68	0.66	0.57	0.67	0.69	0.8
C- support vector	0.81	0.71	0.54	0.73	0.86	0.9	**C- support vector**	0.78	0.75	0.65	0.74	0.88	0.81
Linear discriminant analysis	0.76	0.66	0.47	0.68	0.81	0.88	Linear discriminant analysis	0.7	0.68	0.58	0.68	0.74	0.8
			Relevance							**Incivility**			
Logistic regression	0.9	0.73	0.51	0.81	0.92	0.98	**Logistic regression**	0.86	0.73	0.54	0.75	0.59	0.49
K-Nearest neighbors	0.92	0.61	0.25	0.58	0.98	0.94	K-Nearest neighbors	0.86	0.6	0.27	0.58	0.19	0.51
Gaussian naive Bayes	0.89	0.69	0.45	0.77	0.91	0.97	Gaussian naive Bayes	0.78	0.65	0.43	0.71	0.63	0.33
Bernoulli naive Bayes	0.91	0.53	0.11	0.53	0.98	0.93	Bernoulli naive Bayes	0.86	0.61	0.3	0.59	0.22	0.45
Adaboost	0.93	0.53	0.1	0.53	0.99	0.93	Adaboost	0.86	0.52	0.12	0.53	0.07	0.43
Gradient boosting	0.93	0.5	0.04	0.51	1	0.93	Gradient boosting	0.87	0.5	0.07	0.52	0.04	0.9
Decision tree	0.91	0.65	0.35	0.64	0.96	0.95	Decision tree	0.89	0.74	0.55	0.73	0.52	0.59
Linear support vector	0.81	0.63	0.36	0.77	0.82	0.97	Linear support vector	0.77	0.65	0.43	0.71	0.63	0.33
C- support vector	0.9	0.7	0.45	0.74	0.93	0.96	C- support vector	0.74	0.6	0.36	0.65	0.54	0.27
Linear discriminant analysis	0.89	0.67	0.4	0.71	0.92	0.96	Linear discriminant analysis	0.81	0.67	0.45	0.71	0.56	0.37

grabbing in the form of internet control and removal of open access and toll-less consumer routing. You are a shill, and you have no understanding of that which you speak.Either that, or you are a liar.)

4.2. Predictive performance

The results in Table 4 report the predictive performance of classifiers trained on the TFIDF features of the dataset in a ten-fold cross-validation setup. Column 1, 2 and 3 reports the Accuracy, macro-F1-score and minority F1-score (F1 for the positive class only). The classifiers in the bold font identify those which had the best performance on average across accuracy, F-1 scores, precision, recall, and AUC (Area Under the Curve).

The results suggest that logistic regression and c-support vector classification often had the best average performance among all the approaches. In terms of general implementation, logistic regression is often the easiest to implement and makes no assumptions about class distribution, which is helpful in cases with imbalanced data. It is effective when classes can be linearly separated. The logistic regression classification approach outperformed others in the case of relatively simpler discussion quality categories such as relevance, reciprocity, and incivility. On the other hand, support vector classification is known to work well in high dimensional spaces when there is a clear separation between the classes. The C-support vector classification outperformed others for the relatively complex categories of discussion quality, such as constructiveness, justification, and empathy and respect.

The largest standard deviation in performance is seen in the minority-F1 score. That is, while the accuracy and macro F-1 performance across classifiers are quite close together, where logistic regression and c-support vector classifiers outperform others is in terms of the minority F-1 metric, which identifies the predictive performance on the positive cases alone. It suggests that even the best-performing classifiers for most of the facets except reciprocity and empathy and respect fared only slightly better than random at identifying the positive instances, with minority-F1 scores all under 0.6.

4.3. Linguistic Insights

The word clouds in Figure 1 visualize the TFIDF features most predictive of each of the deliberative facets in the logistic regression classifiers. The words and phrases are sized according to their importance in the classifier. The color denotes whether the feature is a positive (green) or a negative (red) predictor of the facet. The following paragraphs identify the features most predictive of different facets of discussion quality.

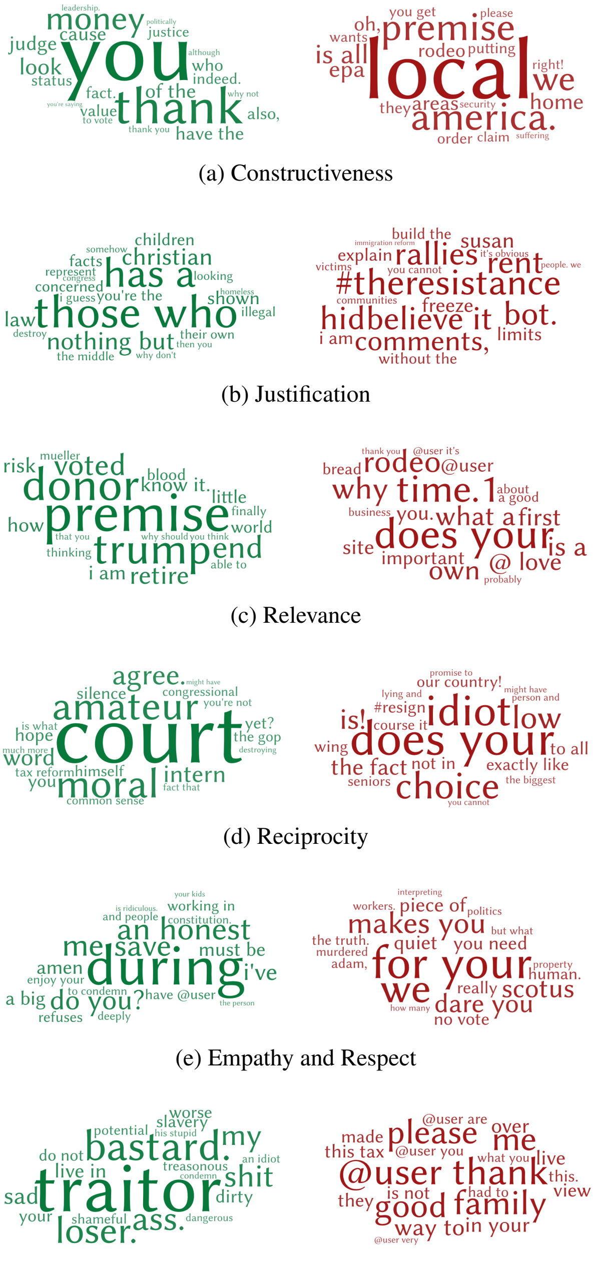
Consider the features predictive of the analytical aspects of discussion quality in Figure 1a-c. Some of the features most predictive of constructiveness seem to resonate with the goal of constructiveness to introduce facts (*fact, cause*), build consensus (*judge*) and invite co-participants to consider a solution (*look*). In Figure 1b, words denoting evidence in the form of opinions (*i guess*) appear together with words denoting evidence in the form of facts and statistics (*fact, facts, illegal, law*). The features predictive of relevance mention political issues in the US (*mueller, trump, voted, donor*).

Among the features predictive of the social aspects of discussion quality in Figure 1d-f, the words predictive

Table 5: Examples of cases with inter-annotator disagreement.

Constructiveness
• @USER #25thAmendmentNow before he gets us all killed! Lives at risk b/c he doesn't like being so low in the polls and you're quoting the Bible? • @USER $9,900 from the NRA during the 2016 election cycle. • @USER #AlFranken should step down when #ThePresident and #RoyMoore do. The Democratic party needs to get their act together otherwise we will be looking at many more years of RW politics.
Justification
• @USER quite sad actually seeing a lot of great work being undermined. @USER @USER https://<LINK> • @USER Subpoena away! Personally I hope that POS perjures himself before the committee, being he thinks he's smarter than you. https://<LINK> • @USER #PuertoRico is only important to Democrats come election time or deflecting from a corruption trial. #MenendezForPrison2017
Relevance
• @USER hypocrite. You are A porn surfer and claim to be holier than thou. • @USER more evidence to support your letter @USER come and deny it as usual. @USER @USER @USER https://<LINK> • @USER visiting your city...stepping around homeless people & homeless vets asking for money...on every corner. what.the.hell? https://<LINK>
Reciprocity
• @USER You (supposedly) work for US, not The other way around. You are supposed to act in our interests not The RINOs like you. • @USER Just saw you on Mornings with Maria. Fight for those state and local tax deductions. Don't let this bill take any deductions away from us. That is stealing from the citizens who pay taxes. Thank you! • @USER doesn't @USER brother work for the DOJ????? Corruption!!!!!! https://<LINK>
Empathy & Respect
• @USER - A poison in our island - Rising seas caused by climate change are seeping inside a United States nuclear waste dump, contamination, cover-up. https://<LINK> #enewetak #potus #gop #democrats #realDonaldTrump #RepLeeZeldin • @USER - more evidence to support your letter @USER come and deny it as usual. @USER @USER @USER https://<LINK> • @USER - why don't you retweet this? Oh... wait... you're a partisan hack. That's right. https://<LINK>
Incivility
• @USER Before this goes any further I want NAMES of pols who PAID OFF accusers w taxpayer money! #IMWITHAL • @USER You are a shameful representative of WV https://<LINK> • @USER "Free and Open Internet" is nothing more than a code word for corporate land-grabbing in the form of internet control and removal of open access and toll-less consumer routing. You are a shill, and you have no understanding of that which you speak. Either that, or you are a liar.

of reciprocity include assent words (*agree.*), questions (*yet?*) and words the mention the person the comment is directed at (*you, you're not*). The words predictive of empathy and respect also reflect the latter two patterns (*do you?, your kids, enjoy your*) and, uniquely for any classifier, mentions of the self (*i've, me*). Finally, the features predictive of incivility include mostly swear words (*traitor, bastard, ass, loser*).



(a) Constructiveness

(b) Justification

(c) Relevance

(d) Reciprocity

(e) Empathy and Respect

(f) Incivility

Figure 1: Words and phrases predictive of high (green) and low (red) scores of the different deliberative facets in the logistic regression classifiers. The size reflects the absolute magnitude of the feature's coefficient in the classifier.

5. Conclusion

This study presents a freely-available corpus for text classification in the domains of communication research and political science. The paper discusses insights from the inter-annotator agreement statistics, demonstrates the performance of classic machine learning approaches on the task of text classification, and visualizes the predictive features.

Inter-annotator statistics suggest that there is room for improvement in specifying the task instructions and training the annotators. Although many examples were provided to the annotators when they were being trained, perhaps the errors persist because there are infinite ways to convey each of the facets. Furthermore, the final choice is ultimately a subjective inference by the annotator, which may or may not reflect what a different annotator infers. In this sense, annotating discussion quality appears to be rather different and far more challenging as compared to annotating for emotion or hate speech.

The results from the predictive evaluation of classic machine learning approaches there is a need to improve the in-domain classification performance with further experimentation. Measuring and predicting discussion quality appears to be a challenging problem for both humans and supervised machine learning approaches, since the classification task was not easily solved with classic methods. A performance improvement may only be possible with further data collection and augmentation, as well as further experimentation in the linguistic feature space. Note that the social aspects of discussion quality are closely related to affect, yet the classifiers have not yet been evaluated with emotion or sentiment input.

While the development of the corpus for assessing discussion quality is still a work-in-progress, the features and the coefficients of the final, best-performing model would ultimately be publicly released as a dictionary, which can be used to calculate the presence of a discussion quality facet in unseen test as a function of the weighted average of the presence of different linguistic features, thereby ‘generating a prediction’ about the presence of a facet of discussion quality. However, applying dictionaries across different platforms requires bearing in mind the subtle differences in their deliberative and behavioral norms. Ultimately, the classifiers should be validated on out-of-domain data to truly establish its generalizability.

Scholars who explore this dataset and wish to apply it to their research should carefully consider the context in which the data was collected, the goals it was intended for, and the appropriateness and generalizability of the dictionaries to their dataset. They may want to examine the relevance of their context to the original context, and validate the labels against a small sample of hand-labeled data to ensure the validity of the labels, and correctly anticipate and interpret its predictions.

6. Bibliographical References

Baden, C., Kligler-Vilenchik, N., and Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, 14(3):165–183.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542.

Buechel, S., Buffone, A., Slaff, B., Ungar, L., and Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., and De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3):190–206.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and 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)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., and 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*, pages 512–515.

Esteve Del Valle, M., Sijtsma, R., and 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.

Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z., and Wernimont, J. (2018). Proceedings of the 2nd workshop on abusive language online (alw2). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*.

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

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

Gründl, J. (2020). Populist ideas on social media: A dictionary-based measurement of populist communication. *New Media & Society*, page 1461444820976970.

Habermas, J. (1984). *The theory of communicative action*, volume 2. Beacon Press, Boston, MA.

Halpern, D. and 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.

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

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

Nelimarkka, M. and Ahonen, P. (2019). Measuring deliberation using machine learning. In *Proceedings of the 13th General Conference of the European Consortium for Political Research (ECPR)*.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pilny, A., McAninch, K., Slone, A., and Moore, K. (2019). Using supervised machine learning in automated content analysis: An example using relational uncertainty. *Communication Methods and Measures*, 13(4):287–304.

Septiandri, A. A., Winatmoko, Y. A., and Putra, I. F. (2020). Knowing right from wrong: Should we use more complex models for automatic short-answer scoring in bahasa indonesia? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 1–7.

Steenbergen, M. R., Bächtiger, A., Spörndli, M., and Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48.

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

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

Tuggener, D., von Däniken, P., Peetz, T., and Cieliebak, M. (2020). Ledgar: a large-scale multi-label corpus for text classification of legal provisions in contracts. In *12th Language Resources and Evaluation Conference (LREC) 2020*, pages 1228–1234. European Language Resources Association.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.