Comparing a BERT Classifier and a GPT classifier for Detecting Connective Language Across Multiple Social Media

Josephine Lukitol[1], Bin Chen[2], Gina M. Masullo[1], and Natalie Jomini Stroud[1]
[1]Center for Media Engagement, University of Texas at Austin
[2]The University of Hong Kong

February 21, 2026

Abstract

This study presents an approach for detecting connective language—defined as language that facilitates engagement, understanding, and conversation—from social media discussions. We developed and evaluated two types of classifiers: BERT and GPT-3.5 turbo. Our results demonstrate that the BERT classifier significantly outperforms GPT-3.5 turbo in detecting connective language. Furthermore, our analysis confirms that connective language is distinct from related concepts measuring discourse qualities, such as politeness and toxicity. We also explore the potential of BERT-based classifiers for platform-agnostic tools. This research advances our understanding of the linguistic dimensions of online communication and proposes practical tools for detecting connective language across diverse digital environments.

1 Introduction

The growth and popularity of social media over the past two decades has created many opportunities for natural language processing and computational social science researchers to study short-form text. During this time, researchers have built a wide variety of text classifiers to understand social media posts, including for sentiment analysis [65], discrete emotion detection [4], life events identification [8], and even depression detection [28]. Overwhelmingly, these efforts have focused on negative or unwanted online content. For example, research efforts have focused on the identification of misinformation, disinformation, or bot activity [33, 56, 57]. Similarly, there are hundreds of studies discussing NLP classifiers for malicious [23] or toxic language [22]. At face value, the emphasis on building classifiers for unwanted content makes sense: one very common use case for NLP classifiers is to identify content for removal, whether it be spam messages [21] or content seen as toxic [3].

And yet, there is little discussion regarding what desired language on social media would look like. Although NLP research has focused a great deal on building classifiers to remove unwanted content on social media, it has paid less attention to classifiers that detect wanted or desired content. To fill this gap, we advocate for and build a classifier for one such language feature: connectivity. We define connective language as language features that express a willingness to talk with people who are not ideologically aligned, such as expressions of intellectual humility or openness to other perspectives. As we explain, connectivity is an essential aspect of human communication, and recent social science research highlights the importance of connective language to facilitate prodemocratic conversations [43]. This research suggests that connective language can help facilitate discussion [42], empower citizens [29], and contribute to a healthier public square. A connective language classifier could be used in multiple ways, such as allowing users to filter or sort content, awarding a badge to users employing the language, or recommending content on a platform. These use cases could help people identify others who are interested in having thoughtful exchanges.

Drawing from the literature in communication research and in natural language processing, this paper introduces and illustrates the use of a multi-platform connective language classifier. First, we build a human-labeled training set using a mix of social media messages from Reddit, Twitter, and Facebook. We use this novel training dataset to build a BERT classifier and a Generative AI (GPT-3.5 Turbo) classifier for connective language. Finally, we compare the connective language classifier to concepts for which there are existing classifiers, such as politeness, to show how they are semantically distinct.

2 Related Work

2.1 Pro-Democratic NLP Efforts

Given how much language and conversation, both political or otherwise, that occurs online and through digital platforms, natural language processing is increasingly important for pro-democratic efforts, from studying free speech efforts [14] and improving public service accessibility [35] to encouraging citizen participation [1].

One pivotal area of NLP research is political opinion and information detection [52, 17]. These efforts can be used to decrease political animosity [31] and increase contact with different perspectives on a political issue [47]. While acknowledging that language models may themselves have political biases [25], they nevertheless can help citizens sort through the overwhelming amount of content now produced online.

2.2 Polite, Civil, and Deliberative Language

Identifying quality discourse has been a key feature of past research. Much of the work draws from deliberative theory [26], which has been defined in numerous ways, but often includes the idea that interlocutors, treated equally, respectfully engage in fact-based discussions to reach consensus [12]. As summarized in Table 1, many past studies draw from this approach when analyzing discourse, whether in face-to-face conversations, within comment sections, or, most popular recently, on social media. Studies examine whether there is evidence of rational information exchange, including the citation of evidence, the presence of reasoned arguments, and whether people are asking genuine questions. Also consistent with some definitions of deliberation, past work has examined utterances that provide solutions or build toward consensus. Quality exchanges, according to several studies, also include interactivity and reciprocity among participants.

Beyond the informational content and the presence of interactivity, some studies also have looked at the tone of the conversation. Incivility, for instance, is seen as antithetic to deliberation [20]. Civility and respect characterize some operationalizations of quality discussion, yet most of the research looks for the presence of incivility and disrespect, as opposed to language indicating civility and respect. This is critical because a comment that does not use uncivil or disrespectful language is not necessarily civil and respectful. The final discourse quality category we identified across studies, labeled Acknowledgment in Table 1, looks at how people treat others and others' arguments in a discussion. The concepts used vary broadly. Some involve acknowledging others' views, regardless of whether one is sympathetic. Others involve meta-reflection on the conversation overall. Yet others involve empathy for different viewpoints.

In a highly polarized context such as the United States, the opportunity for deliberation as conceived of by deliberative theorists is optimistic, but slim (e.g., [40]). Political partisans routinely do not engage in deliberation, let alone agree upon facts, engage with each other, or respectfully work toward consensus. Rather than focusing on deliberation as solely important, scholars have noted that it may be better to consider related concepts—other forms of desired language that may lead to (but are not necessarily) deliberation [54, 43].

For example, identifying language that recognizes the humanity of the interlocutors or indicates an acknowledgement of differing opinions may help connect ideologically divergent groups, such as Democrats and Republicans in the United States. Although a few concepts from Table 1 may hold promise, such as empathy and respect for counterarguments, it is equally important to consider (1) how these individual concepts may operate together to facilitate pro-democratic connectivity and (2) how one might computationally-detect such concepts.

A handful of NLP studies have sought to identify desired language styles, including polite language [46] and empathy [69]. These studies rely on background literature from social science disciplines, but leverage computational and NLP expertise to build pro-social classifiers that have the potential to improve online conversation [32].

2.3 Connective Language

Connective language is distinct from these past work in that it emphasizes linguistically building connections. It includes encouraging engagement, understanding, and conversation, using techniques such as expressing openness to alternative viewpoints. Although it has some aspects in common with the use of polite language, there are many forms of polite language that would not be connective (e.g. saying please). The idea also is related to (but distinct from) empathy, as connective posts are not about how one internalizes others' views. Rather, connective posts are about presenting one's own point in a manner that invites others to engage productively.

Research suggests that this type of language can reduce affective polarization. First, there's good evidence that exposure to sympathetic outparty members can curb affective polarization [64]. Outpartisans writing connective posts should be seen as more sympathetic. Second, the use of humility—one form of connective language—can improve people's attitudes toward commenters from an opposing political party [38] and research on inter-group contact theory finds that positive interactions with individual outparty members can generalize to evaluations of the opposing party as a whole [45].

Table 1: Related Work on Attributes of Quality Discourse

Category	Description	References
Rationality	Evidence, Justification, Relevance, Opinion expression, Reflexivity, Argument repertoire	[60, 27, 16]
Questions	General questions, Genuine questions, Inflammatory questions	[59, 15, 24, **?**]
Consensus/Solutions	Working toward consensus, Proposing solutions, Resolving conflicts	[27, 68, 16, 39]
Interactivity/Reciprocity	Replying, Referencing	[68]
Respect/Civility	Incivility, Interruption, Impoliteness, Negative empathy, Civility, Respect for others	[11, 68, 7, 36, 16]
Acknowledgement	Value another's statement, Respect for arguments	[27, 9, **?**, 59, 20, 36, 16]

3 Proposed Method

To build a connective language classifier, we apply the following approach: first, we build a multi-platform dataset consisting of content from users who are likely to be engaging in discussion on a topic about which they disagree. This includes a mix of political topics (e.g., for whom should a citizen vote?) and apolitical discussion (e.g., should pineapple be a pizza topping?).

We then construct a gold-standard training set of connective language using human labelers. After achieving inter-coder agreement, four undergraduate students labeled 14,107 social media posts. We then use these messages to build a connective language BERT classifier. We compare this classifier to one built using GPT 3.5 turbo, a large-language model. We also analyze how connective language is distinct from other similar concepts, including politeness and constructiveness.

3.1 Dataset

To identify social media posts with connective language, we took an inductive approach. We first constructed a list of five Reddit and Twitter accounts that engaged in cross-cutting discussion that (1) did not alienate and (2) sometimes encouraged deliberation with ideologically-opposed social media users. These were: r/ChangeMyView, Olympia Snowe, Kathryn Murdoch, Nolabels, Braver Angels. From this list, the authors then derived eight attributes that could relate to connective language: humility, humanizing, common humanity, acknowledgement of emotions/thoughts, consensus building, reflective listening, reactivity, and truthfulness in conversation. These aligned with recommendations from journalists (`https://journalistsresource.org/politics-and-government/receptive-opposing-views-research/`) and organizational communication researchers (e.g., [18]) for building trust.

Using these five examples and eight attributes, four undergraduate students were tasked with identifying similar accounts across Twitter. A total of 31 Twitter accounts were identified by the undergraduate coders and confirmed to contain connective language by the authors. These were: "The65Project", "PreetBharara", "BarbMcQuade", "mashagessen", "ianbremmer", "NateSilver538", "Yascha_Mounk", "KHayhoe", "uniteamerica", "NickTroiano", "KarenKornbluh", "BrennanCenter", "NowThisPolitics", "kylegriffin1", "politico", "hrw", "cliffordlevy", "ZekeJMiller", "CREWcrew", "PhilipRucker", "tribelaw", "glennkirschner2", "HeartlandSignal", "nprpolitics", "ezraklein", "JohnkingCNN", "txpolproject", "ap_politics", "mattyglesias", "HeerJeet", "UNHumanRights", "bbcpolitics". Additionally, we constructed a keyword-based query to supplement our user collection. The case-insensitive keyword query included the following 12 terms: imo, imho, inmyopinion, "in my opinion", "I hear you", "never thought about it", "my perspective", "see where you're coming from", "see where ur coming from", "thanks for sharing", "complicated issue", "correct me if". Posts from the original 31 accounts were subsampled for posts using the aforementioned 12 terms.

Public Twitter data from these accounts were gathered using the Twitter 2.0 Academic Track API from January 1, 2012 to December 31, 2022. To collect this data, we used two queries (one keyword-based and one user-based).

For Reddit, we considered posts published from January 1, 2012 to December 31, 2022, which were gathered from July 1 to 17, 2023 using Pushshift [6] from the following subreddits: r/ChangeMyView and r/politics (two English-based subreddits, with the former including apolitical posts and the latter focused on political posts), using the above list of 12 query terms. Both subreddits

are highly active with many users; at the time of the collection, r/ChangeMyView had 3.6 million followers and r/politics had 5 million followers in 2024.

For Facebook, we did not conduct a user-based query and simply queried for the use of the 12 terms across all public Facebook groups and pages available through Crowdtangle from January 1, 2012 to December 31, 2022. This collection was conducted from July 1 to 30, 2022.

To construct the dataset used to train this classifier, we took a subsample from each corpus and combined them into an English-language dataset that consisted of public Reddit submissions ($n = 6,107$), Twitter posts ($n = 5,000$), and Facebook posts ($n = 3,000$).

Using different query parameters for each data collection has become an increasingly common practice to account for temporal, discursive, and platform diversity (for similar collections, see [2, 48]). Identifying information from this dataset, including the pseudonym or name of the account producing the content, has been removed from the dataset.

3.2 Labeled Data

To build a connective language classifier, we developed a codebook and hired four undergraduate students to code posts. The faculty co-authors initially conducted a comprehensive literature review on how various fields had conceptualized and operationalized concepts like connective language. A synthesis of this literature was developed into a preliminary codebook and shared with the students, who then brainstormed with the faculty authors to determine broad categories for operationalizing the concept of "connective posts" versus "not connective posts." Then the students coded repeated random samples of 100 posts each drawn from our universe to practice coding and iterate on the coding guide, based on post content. Next the students conducted eight rounds of coding, meeting weekly until they achieved a reliable Krippendorff's α (0.73) using a sample of 1,000 posts. Once the students achieved an inter-coder reliability above a 0.7 threshold, we then had students code 6,107 Reddit posts, 5,000 Twitter posts, and 3,000 Facebook posts, over three rounds, using the following coding guide:

A connective post was coded "1" and defined as a post that:

- Encourages engagement, understanding, and conversation, sometimes by asking questions, or expressing openness to alternative views.
- Contains language that conveys openness by including phrases, such as "in my opinion," "imo," "imho," "in my viewpoint," "here's how I see it," "in my mind," "my 2 cents is."
- Other indicators of a connective posts include phrases such as "I respectfully disagree," "I disagree to an extent," "You're right about xxx," "I see where you're coming from," "You've changed my view," "I never thought about it like that," "Can you clarify," "I'm not trying to debate, but want to offer an opinion," "That's an interesting perspective," "I appreciate your feedback."
- Clarification: Hate speech (e.g., racist, sexist, homophobic, or xenophobic language) would invalidate a post as "connective," but profanity alone would not.

A non-connective post was coded 0 and defined as a post that:

- Lacks any of the elements of connective posts described above or included hate speech.
- Demonizes another person or is disrespectful to other points of view.
- Contains no discussion.

To validate this operationalization of connective posts, accounting for variations in gender, race/ethnicity, and political beliefs, we conducted an online survey ($n = 621$) and find little to no demographic differences across evaluations regarding connective language. These details can be found in the Appendix A.1.

3.3 BERT Classifier

Using human-labeled data, we trained a BERT (Bidirectional Encoder Representations from Transformers [13]) classifier to predict the presence of connective language in text content. Compared to traditional text classification methods, such as logistic regression and Naive Bayes models, a BERT classifier excels due to its deep understanding of context and language nuances [53, 55, 37], which is particularly useful in complex tasks, such as detecting connective language in texts.

As seen in Figure 1, we use the following approach: from the entire human-coded dataset, we first created a balanced sample ($N = 10,894$) by undersampling the "1" group, due to fewer instances of "0"s in the labeled data. A balanced dataset is crucial as it ensures that the model learns to recognize patterns associated with both classes equally, which leads to more accurate and generalizable results [5].

Figure 1: Pipeline of fine-tuning a BERT classifier for detecting connective language

We then utilized the bert-base-uncased model [13] for fine-tuning with our balanced labeled sample. The data was divided into training, validation, and test sets to effectively train the model while preventing overfitting. During training of the BERT classifier for binary classification, we employed TFBertForSequenceClassification with an Adam optimizer set at a learning rate of 2×10^{-5}. Essential callbacks like EarlyStopping, ModelCheckpoint, and ReduceLROnplateau were incorporated to enhance training efficiency and optimization on a MacBook Pro with an Apple M1 Pro chip. Default parameters from the scikit-learn package [44] were used. The training process involved multiple iterations where the model predicted labels on the training data and these predictions were compared against the actual labels, continuing until the fine-tuned model demonstrated satisfactory precision and recall.

3.4 Few-shot Classifier

We employed a generative AI tool, specifically OpenAI's "GPT 3.5 Turbo," accessed via the OpenAI API, to classify social media texts for connectivity[1]. The GPT 3.5 Turbo model is the most recently available version of OpenAI's language models, known for its enhanced speed and accuracy, which makes it ideal for real-time text classification tasks.

While social science research may benefit from the efficiency of large language models [49], LLMs may exhibit biases [62] and reliability issues [34].

The classification process involved a prompt that defined "connectivity" and requested that the model classify an unlabeled post as either "1" (connective) or "0" (non-connective). After several attempts (see Appendix A.2), the final prompt provided to the model was as follows:

> Please perform a text annotation task: Below is the definition of 'connectivity' and an unlabeled post. Your task is to classify the post based on whether it demonstrates connectivity. Respond only with '1' for connective or '0' for non-connective. Definition of Connectivity: Connectivity indicates the tone of a message. A post is considered connective if it shows a willingness to engage in conversation with others, especially those with differing opinions, uses hedging, or maintains a polite tone when sharing opinions or facts. Phrases like 'in my honest opinion' are also markers of connective language. This definition is derived from the codebook used by the human coders. Here is the post: "TEXT"

We sampled a balanced set of 1000 texts (500 connective, 500 non-connective), stratified by platform, from our human-labeled dataset. We then compared the classifications made by the GPT model to the human labels, treating the human labels as actual values and the GPT's outputs as predictions.

3.5 Comparing BERT and LLMs

We choose to compare a BERT classifier and a GPT-based classifier as both are popular language models for building classifiers in the social sciences. While the BERT model has been used to build other political communication classifiers for topics such as deliberation [19], GPT-based classifiers are comparatively newer. Furthermore, scholars have raised concerns about GPT 3.5's unreliability and tendency to produce biased outputs [66], especially when dealing with topics related to stereotyping and protected demographic groups. However, at the time of our study, it was unclear whether these biased outputs would also impact the ability to produce classifiers for normatively desired content (such as connective language).

3.6 Comparison to Other Concepts

To demonstrate the conceptual uniqueness of "connectivity," we compared the result of connective language detection (human-labeled results) with several other related concepts, including politeness, civility, and a set of attributes related to political discussion quality such as constructiveness, justification, relevance, and reciprocity [30]. Through correlation analysis between the score of connective language and other concepts for the same texts, we show the connectivity is a distinct attribute of political and social discussions.

For detecting toxicity, we employed the Perspective API[2], a tool developed by Jigsaw and Google that uses machine learning models to identify and score the degree of perceived harmfulness or unpleasantness in written content. The output from Perspective API provides a set of scores for various sub-attributes, such as personal attacks, among others, in addition to an overall toxicity score. For our analysis, we specifically utilize the overall toxicity score, ranging from 0 (not toxic at all) to 1 (extremely toxic), to assess the general level of toxicity in the texts. This score synthesizes insights from all the sub-attributes into a single comprehensive measure, enabling a clear and focused evaluation of toxicity. We also compare the classifier to the new Perspective API attributes, which are experimental: affinity, compassion, curiosity, nuance, personal story, reasoning, and respect.

To detect politeness, we utilized the R package "politeness" [67], a statistical tool designed to analyze linguis-

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

[2] https://support.perspectiveapi.com/

tic cues and determine the levels of courtesy and respect present in text. We utilized the politenessModel function, which is a wrapper that can be used around a pre-trained model for detecting politeness from texts [10]. This function outputs a score ranging from -1 to 1, where higher values represent higher politeness, and lower values indicate less politeness or rudeness.

In addition to toxicity and politeness, we also compared the connective language with a set of attributes related to the quality of political discussions proposed by [30]. We are specifically concerned with six attributes that are related to connective language, constructiveness, justification, relevance, reciprocity, empathy/respect, and incivility. We used the classifiers featured in this paper to do the classifications.

4 Results

4.1 Descriptives

Table 2 provides a descriptive summary of human-coded posts used for training machine learning classifiers, showing the distribution of posts labeled as connective (1) and non-connective (0) across three major platforms: Facebook, Reddit, and Twitter. Notably, the data highlights variability in connective language usage, with Twitter exhibiting a higher percentage of connective posts (61.5%), compared to Reddit and Facebook.

Table 2: Descriptive of Human-coded Posts by Platform

Platform	Label	Count	Percentage
Facebook	0	1196	43.9%
	1	1527	56.1%
Reddit	0	2133	50.1%
	1	2661	49.9%
Twitter	0	1903	38.5%
	1	3041	61.5%

4.2 Model Evaluation: BERT vs GPT

To evaluate and compare the performance of two classifiers, BERT and GPT-3.5 Turbo, we assessed their ability to predict whether social media posts convey "connective language" by comparing the predicted values from each classifier against the human-labeled results on the same data. The evaluation metrics used included precision, recall, and F1-score, as detailed in Table 3.

The BERT model, using the pre-trained model "bert-base-uncased," analyzed 1,000 posts and demonstrated a precision of 0.85, recall of 0.84, and an F1-score of 0.85.

In contrast, the GPT-3.5 Turbo model, when evaluating the same 1,000 posts, recorded lower scores across all metrics with a precision of 0.55, recall of 0.42, and F1-score of 0.48. These results indicate that the BERT model outperforms the GPT-3.5 Turbo in accurately identifying the conveyance of connective language in social media posts.

Table 3: Evaluation metrics of BERT and GPT classifier by platform

Metric	BERT			GPT	
	Precision	Recall	F1	Precision	Recall
Overall ($N = 1000$)	0.85	0.84	0.85	0.55	0.42
Facebook ($N = 203$)	0.92	0.86	0.89	0.64	0.22
Twitter ($N = 229$)	0.81	0.72	0.76	0.55	0.51
Reddit ($N = 568$)	0.97	0.99	0.98	0.51	0.32

4.3 Comparing Connectivity to Other Concepts

We conducted a correlation analysis (see Table 4) to explore the relationship between the new metric of connectivity and established measures within the context of political discussions. This analysis highlighted the unique aspects of the connectivity metric and its interactions with other key qualities of online discussions.

The findings reveal that connectivity negatively correlates with toxicity and incivility. Additionally, connective language identified with the BERT classifier shows a positive correlation with politeness, at 0.28, as well as empathy-respect, at 0.29. This implies that conversations with greater connectivity are also labeled as more polite and respectful, and less toxic or incivil.

Furthermore, weak to no negative correlations were found between connectivity and other concepts such as constructiveness, justification, relevance, and reciprocity. These findings provide robust evidence that connectivity captures elements of communication that are not fully addressed by traditional metrics. This distinctiveness is vital for a deeper understanding of the structural and relational dynamics that are often neglected in conventional content-focused analyses of online discussions.

Table 5 shows the results of a correlation test between three connective measurements: BERT, Human, and GPT, and seven measurements related to the "bridging system" [41] computed by Perspective API[3]: Affinity, Compassion, Curiosity, Nuance, Personal Story, Reasoning, and Respect. The results show that the measurements of connective language are, in some instances, weakly correlated

[3] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

with the "bridging" measurements such as affinity and respect, yet the magnitude is modest, indicating the conceptual uniqueness of connective language.

5 Discussion

Connectivity emerged as an important attribute of online discussions. In this study, we proposed two types of classifiers to detect connective language from social media posts. First, we found that the BERT classifier outperforms GPT-3.5 turbo in classifying texts into connective and non-connective categories. This indicates the superior effectiveness of BERT in identifying connective language within political discussions. Additionally, we found that connective language is conceptually distinct from other related concepts such as politeness, toxicity, constructiveness, reciprocity, among others, suggesting that connectivity represents a unique dimension of discourse quality. Furthermore, our results demonstrate the ability to use BERT to construct multi-platform classifiers, enhancing the versatility and applicability of our approach and potentially laying the foundation for platform-generalizable classifiers.

While our analysis did not necessarily find biases among the outputs of the GPT classifier, the decreased accuracy of this classifier may be a result of the more complex, nuanced, or new conceptualization of connectivity, as opposed to more overt or well-studied labels like sentiment or toxicity.

This classifier could be used to test whether those using connective language have more deliberative conversations, as theory would suggest [12]; to evaluate the effects of exposure to social media posts that contain connectivity; and to examine practical ways of increasing connectivity to the extent that it has pro-democratic effects such as increasing understanding of alternative views.

5.1 Limitations

As with any study, we recognize that there are several limitations to this study that we were unable to address or were beyond the scope of our study. First, we constructed our sample in an effort to oversample for connective language. To do so, we sought out digital spaces where discussion and disagreement occurs, and we used keywords that literature suggests may be used when disagreement occurs. Therefore, the proportion of connective posts in our sample is not necessarily representative of a typical virtual conversation or topic. Future studies can build on this work by applying the classifier to more generalizable contexts.

Additionally, while we were able to build a classifier using multi-platform annotations from Facebook, Reddit, and Twitter, we do not consider a wide variety of other platforms, including audio-based and video-based platforms such as YouTube and TikTok. The consideration of spoken language-based classifiers, while important, was beyond the scope of our analysis and should be considered in future work.

This work is foremost motivated by a desire to advance NLP classifiers that identify desirable language and contribute to quality discussion. Drawing from literature on the importance of interactivity, respectfulness, and expressions of openness [60, 59, 39, 20], our work is among the first to propose an NLP classifier to detect connective language.

In addition to building a classifier for a relatively understudied concept, our connective language classifier also contributes to ongoing scholarly efforts to build multi-platform classifiers (e.g., [63, 51]). While single-platform analyses have served as a useful starting point, this work can fail to consider the ever-expanding nature of our multi-platform digital ecosystem.

We consider this work to be "in conversation" with the plethora of NLP scholarship building classifiers for harmful or toxic language (e.g., [3, 31]). While the study of harmful or toxic language is certainly important, especially for removal efforts, it is equally important (and comparatively uncommon) to study and build classifiers for desired language styles. We hope this work inspires others to build and develop classifiers for both undesired and desired online content.

6 Conclusion

This work was supported by the John S. and James L. Knight Foundation and partially supported by the UT-Austin Bridging Barriers Research Development Initiative. We are grateful to Megan A. Brown, Jessy Li, Lynnette Ng, and Ashwin Rajadesingan for their helpful comments and suggestions.

References

[1] Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22.

[2] Michele Avalle, Niccolò Di Marco, Gabriele Etta, Emanuele Sangiorgio, Shayan Alipour, Anita Bonetti, Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli, Matteo Cinelli, et al. 2024. Persistent

Table 4: Correlations Between Connectivity and Other Concepts

Variable	M	SD	1	2	3	4	5	6	7	8	9
1. Conn. (BERT)	0.48	0.50	1								
2. Conn. (Human)	0.41	0.49	.73**	1							
3. Conn. (GPT)	0.50	0.50	.48	.50	1						
4. Toxicity	0.38	0.42	-.10**	-.06*	.06*	1					
5. Politeness	0.15	0.21	.28**	.27**	.09**	-.19**	1				
6. Constructiveness	0.01	0.07	-.07*	-.09**	.01	-.05	.24**	1			
7. Justification	0.02	0.01	-.14**	-.15**	.02	-.09**	.12**	-.04	1		
8. Relevance	0.02	0.01	-.09**	-.10**	.01	-.06	.06*	.16**	-.15**	1	
9. Reciprocity	0.01	0.01	.09**	.16**	.04	-.15**	.27**	.04	-.12**	-.07*	1
10. Emp.-Respect	0.01	0.01	.29**	.31**	.12**	-.45**	.84**	.23**	-.16**	.16**	.27**
11. Incivility	0.01	0.01	-.08*	-.07*	-.02	.04	-.12**	-.06	.06*	-.15**	-.16**

Table 5: Correlation Matrix Between Connectivity and "Bridging" Attributes

Variable	M	SD	1	2	3	4	5	6	7	8
1. Conn. (BERT)	0.48	0.50	1							
2. Conn. (Human)	0.41	0.49	.73**	1						
3. Conn. (GPT)	0.50	0.50	.48	.50	1					
4. Affinity	0.38	0.42	.09**	.06*	.25**	1				
5. Compassion	0.42	0.21	.21**	.27**	.09**	.21**	1			
6. Curiosity	0.40	0.22	.11**	.11**	.05	.11**	.32**	1		
7. Nuance	0.36	0.21	-.06	-.11**	-.06*	.09**	.38**	.34**	1	
8. Personal_Story	0.44	0.29	-.07*	-.06	-.11**	.14**	.45**	.24**	.46**	1
9. Reasoning	0.36	0.23	.14**	.21**	-.03	.30**	.24**	.62**	.22**	.54**
10. Respect	0.45	0.26	.30**	.38**	.06	.23**	.45**	.24**	.54**	.23**

interaction patterns across social media platforms and over time. *Nature*, 628(8008):582–589.

[3] Nikolay Babakov, Varvara Logacheva, and Alexander Panchenko. 2024. Beyond plain toxic: building datasets for detection of flammable topics and inappropriate statements. *Language Resources and Evaluation*, 58(2):459–504.

[4] V. S. Bakkialakshmi and T. Sudalaimuthu. 2022. Anomaly Detection in Social Media Using Text-Mining and Emotion Classification with Emotion Detection. In *Cognition and Recognition*, pages 67–78. Springer Nature Switzerland.

[5] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

[6] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

[7] Joseph N Cappella, Vincent Price, and Lilach Nir. 2002. Argument repertoire and patterns of participation in the 2000 presidential campaign. *Political Communication*, 19(1):73–93.

[8] Paulo R. Cavalin, Luis G. Moyano, and Pedro P. Miranda. 2015. A Multiple Classifier System for Classifying Life Events on Social Media. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1332–1335.

[9] Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4):658–679.

[10] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

[11] Marc Esteve Del Valle, Rimmert Satsma, Hanne Stegeman, and Rosa Borge. 2020. Online deliberation and the public sphere: Developing a coding manual to assess deliberation in twitter political networks. *Javnost-The Public*, 27(3):211–229.

[12] Michael X. Delli Carpini, Fay Lomax Cook, and Lawrence R Jacobs. 2004. Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature. *Annual Review of Political Science*, 7(1):315–344.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[14] Giovanna Maria Dora Dore, Arya D McCarthy, and James A Scharf. 2023. A Free Press, If you Can Keep It: What Natural Language Processing Reveals About Freedom of the Press in Hong Kong. Springer Nature.

[15] Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3):321–342.

[16] Katharina Esau, Lena Wilms, Janine Baleis, and Birte Keller. 2023. For deliberation sake, show some constructive emotion! how different types of emotions affect the deliberative quality of interactive user comments. *Javnost-The Public*, 30(4):472–495.

[17] Neele Falk and Gabriella Lapesa. 2022. Scaling up discourse quality annotation for political science. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3301–3318.

[18] Charles Feltman. 2011. *The Art of Waking People Up: Cultivating Awareness and Authenticity at Work*. Jossey-Bass.

[19] Eleonore Fournier-Tombs and Michael K MacKenzie. 2021. Big data and democratic speech: predicting deliberative quality using machine learning techniques. *Methodological Innovations*, 14(2):20597991211010416.

[20] Deen Freelon. 2015. Discourse architecture, ideology, and democratic norms in online political discussion. *New media & society*, 17(5):772–791.

[21] Pranjul Garg and Nancy Girdhar. 2021. A Systematic Review on Spam Filtering Techniques based on Natural Language Processing Framework. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 30–35.

[22] Anusha Garlapati, Neeraj Malisetty, and Gayathri Narayanan. 2022. Classification of Toxicity in Comments using NLP and LSTM. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 16–21.

[23] Sagar Gharge and Manik Chavan. 2017. An integrated approach for malicious tweets detection using NLP. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 435–438.

[24] Valentin Gold, Mennatallah El-Assady, Annette Hautli-Janisz, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2017. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158.

[25] Lucas Gover. 2023. Political bias in large language models. *The Commons: Puget Sound Journal of Politics*, 4(1):2.

[26] Jürgen Habermas. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT Press.

[27] Daniel Halpern and Jennifer Gibbs. 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.

[28] Seyed Habib Hosseini-Saravani, Sara Besharati, Hiram Calvo, and Alexander Gelbukh. 2020. Depression Detection in Social Media Using a psychoanalytical Technique for Feature Extraction and a Cognitive Based Classifier. In *Advances in Computational Intelligence*, pages 282–292. Springer International Publishing.

[29] Maria Iranzo-Cabrera and Andreu Casero-Ripollés. 2023. Political entrepreneurs in social media: Self-monitoring, authenticity and connective democracy. The case of Íñigo Erlijón. *Heliyon*, 9(2):e13262.

[30] Kokil Jaidka. 2022. Developing a multilabel corpus for the quality assessment of online political talk. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5503–5510.

[31] Chenyan Jia, Michelle S Lam, Minh Chau Mai, Jeffrey T Hancock, and Michael S Bernstein. 2024. Embedding democratic values into social media ais via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–36.

[32] Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

[33] P Latha, V Sumitra, V Sasikala, J Arunarasi, AR Rayini, and N Nithiya. 2022. Fake profile identification in social network using machine learning and nlp. In *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 1–4. IEEE.

[34] Abdul Majeed and Seong Oun Hwang. 2024. Reliability issues of llms: Chatgpt a case study. *IEEE Reliability Magazine*.

[35] Ilaria Mariani, Maryam Karimi, Grazia Concilio, Giuseppe Rizzo, and Alberto Benincasa. 2022. Improving public services accessibility through natural language processing: Challenges, opportunities and obstacles. In *Proceedings of SAI Intelligent Systems Conference*, pages 272–289. Springer.

[36] Sanju Menon, Weiyu Zhang, and Simon T Perrault. 2020. Nudge for deliberativeness: How interface features influence online discourse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

[37] Lara Souto Moreira, Gabriel Machado Lunardi, Matheus de Oliveira Ribeiro, Williamson Silva, and Fabio Paulo Basso. 2023. A study of algorithm-based detection of fake news in brazilian election: Is bert the best. *IEEE Latin America Transactions*, 21(8):897–903.

[38] Caroline Murray, Marley Duchovnay, and NJ Stroud. 2021. Making your political point online without driving people away. Online report, Center for Media Engagement.

[39] Caroline Murray, Martin J Riedl, and Natalie J Stroud. 2023. Using facebook messenger versus groups for news engagement. *Digital Journalism*, pages 1–19.

[40] Diana C Mutz. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

[41] Aviv Ovadya and Luke Thorburn. 2023. Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance. *arXiv preprint arXiv:2301.09976*.

[42] Christian Staal Bruun Overgaard, Anthony Dudo, Matthew Lease, Gina M. Masullo, Natalie Jomini Stroud, Scott R. Stroud, and Samuel C. Woolley. 2021. Building connective democracy: Interdisciplinary solutions to the problem of polarisation. In *The Routledge Companion to Media Disinformation and Populism*. Routledge.

[43] Christian Staal Bruun Overgaard, Gina M. Masullo, Marley Duchovnay, and Casey Moore. 2022. Theorizing Connective Democracy: A New Way to Bridge Political Divides. *Mass Communication and Society*, 25(6):861–885.

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

[45] Thomas F Pettigrew and Linda R Tropp. 2013. *When groups meet: The dynamics of intergroup contact*. Psychology Press.

[46] Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2024. Computational politeness in natural language processing: A survey. *ACM Computing Surveys*, 56(9):142.

[47] Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter van Atteveldt. 2021. Are we human, or are we users? the role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *1st workshop on NLP for positive impact*, pages 47–59. Association for Computational Linguistics.

[48] Gabriel Roccabruna, Steve Azzolin, Giuseppe Riccardi, et al. 2022. Multisource multi-domain sentiment analysis with bert-based models. In *European Language Resources Association*, pages 581–589.

[49] Hannes Rosenbusch, Claire E Stevenson, and Han LJ van der Maas. 2023. How accurate are gpt-3's hypotheses about social science phenomena? *Digital Society*, 2(2):26.

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

[51] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

[52] Indira Sen, Fabian Flöck, and Claudia Wagner. 2020. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1413–1426.

[53] Yifan Shen and Jiahao Liu. 2021. Comparison of text sentiment analysis based on bert and word2vec. In *2021 IEEE 3rd international conference on frontiers technology of information and computer (ICFTIC)*, pages 144–147. IEEE.

[54] Sarah Shugars. 2020. *Reasoning Together: Network Methods for Political Talk and Normative Reasoning*. Ph.D. thesis, Northeastern University.

[55] Elena Shushkevich, Mikhail Alexandrov, and John Cardiff. 2022. Bert-based classifiers for fake news detection on short and long texts with noisy A comparative analysis. In *International Conference on Text, Speech, and Dialogue*, pages 263–274. Springer.

[56] [ILLEGIBLE]

[57] [ILLEGIBLE]

[58] J. Srinivas, K. Venkata Subba Reddy, G. J. Sunny Deol, and P. VaraPrasada Rao. 2021. Automatic Fake News Detector in Social Media Using Machine Learning and Natural Language Processing Approaches. In *Smart Computing Techniques and Applications*, pages 295–305. Springer.

[59] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

[60] Jennifer Stromer-Galley. 2007. Measuring deliberation's content: A coding scheme. *Journal of Deliberative Democracy*, 3(1).

[61] Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. *Natural Language Processing Research*, 1(1-2):1–13.

[62] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*.

[63] David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54(4):851–874.

[64] Jan G Voelkel, Michael Stagnaro, James Chu, Sophia Pink, Joseph Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi Allen, et al. 2023. Megastudy identifying effective interventions to strengthen americans' democratic attitudes. Working paper.

[65] Jenq-Haur Wang, Ting-Wei Liu, Xiong Luo, and Long Wang. 2018. An LSTM Approach to Short Text Sentiment Classification with Word Embeddings. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, pages 214–223.

[66] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

[67] Mike Yeomans, Alejandro Kantor, and Dustin Tingley. 2023. *politeness: Detecting Politeness Features in Text*.

[68] Marc Ziegele, Oliver Quiring, Katharina Esau, and Dennis Friess. 2020. Linking news value theory with online deliberation: How news factors and illustration factors in news articles affect the deliberative quality of user discussions in sns' comment sections. *Communication Research*, 47(6):860–890.

[69] Ke Zhou, Luca Maria Aiello, Sanja Scepanovic, Daniele Quercia, and Sara Konrath. 2021. The language of situational empathy. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–19.

Appendix

6.1 Concept Validation

To assess the conceptualization and operationalization of connective language, we conducted an online survey with 621 individuals varying in gender, race/ethnicity, and political beliefs. Initially, 977 people participated in the survey, but data were not used for those who may have taken the survey more than once ($n = 233$), failed a validation check within the survey ($n = 88$), failed one or more attention checks ($n = 7$), did not indicate they were at least 18 years old ($n = 6$), or did not indicate they were a U.S. resident ($n = 5$). Participants were recruited using CloudResearch, an online platform that draws participants from Amazon Mechanical Turk (MTurk). CloudResearch screens out MTurker participants who may be bots, based on inconsistent answers to demographic questions and/or

suspicious geolocations [**?**]. We set quotas for gender, race, and political beliefs to ensure that we would get suitable diversity for comparisons.

Participants were first invited to rate four posts—two rated as "connective" and two rated as "not connective" by our undergraduate coders—but the participants were not told of these undergraduates' ratings. They rated how much they disagreed or agreed on a 1 to 5 scale with each of the following statements for each validation comment they viewed: "The person who wrote this posts seems open to understanding the views of someone who might disagree," "The post might help someone with a different viewpoint to understand this person's beliefs," "This post has the potential to build connections with people who disagree with it," and "Someone who disagrees with the views expressed in this post would likely find this post respectful." Responses were averaged together for each validation comment, and only data for those participants who answered all the validation questions correctly were used to actually rate the comments.

Then participants were randomly assigned to answer the same questions about five additional comments out of 40 total possible comments (20 that undergraduates had rated as "connective," and 20 that they had rated as "not connective.") These 40 comments were randomly selected out of the dataset.

After averaging together ratings for each of the 40 comments, we conducted a series of chi square tests of independence that examined whether there was a relationship between gender, race, or political beliefs, and whether people rated the comments as "connective" or "not connective."

Only two comments of 40 comments were rated differently based on demographics. In one case, women and men differed in their ratings: "Um, if every square inch of a park has smokers, honestly it may be on the family to find a less crowded park and clearly the smokers have a bigger interest than the family since they would outnumber the family. Cars really dont have that much benefit besides they destroyed the public transit system and we waste a shit ton of resources on them. We also are unhealthier, waste money, and waste land because of them. Smoking in general seems to be associated with lower income." Women interpreted this post to be connective, whereas men interpreted this post to be non-connective. In another, Black Indigenous People of Color (BIPOC) people disagreed with white people: "I understand it's not polite to try to talk with random strangers while they are trying to shop. *You* understand that. Kids don't. They'll go up to any interesting person and yammer on unless you teach them not to. This is one way to teach them not to." White people perceived this as slightly more connective, whereas BIPOC people did not.

Both of these comments had been rated as non-connective by our trained undergraduate coders. Given that only two analyses out of the 120 chi squares showed any relationships between demographics and how people answered, we are confident that our operationalization of connective posts resonates across various groups.

6.2 Prompt Engineering

To develop the final prompt we used, we tried two alternatives and tuned them to improve on the classification task for the third and final prompt.

First Prompt:

> Please perform a text annotation task: I will provide you with the definition of 'connectivity' and several example posts which demonstrate "connectivity". Then, I will show you some unlabeled posts. Your task is to classify the post based on whether it demonstrates connectivity or not. Label 1 if yes, 0 otherwise.
>
> Here is the definition of connectivity. "Connectivity" reflects the tone of a message. A post is connective if it expresses a willingness to engage in conversation with others that they disagree with, includes a hedge, or is tonally polite when sharing an opinion or fact. For example, expressing honesty, such as "in my honest opinion," is a connective language marker.
>
> Here are 5 example posts that demonstrate "connectivity": [1] "I hear you there Roger.....Miss this girl every day" [2] "I love how Cake's friends had Eiw's back when Cake was away, and continued to so in times like this by showing up, Fee too. The siblings would need all the support they can get, killing off a character wasn't necessary in my opinion." [3] "Our fren got bounced off here last night—same night he debuted his newest (and best yet IMHO) vidya, Ey..." [4] "So....documents were found in the VP office that belonged to President Biden. Correct me if I'm wrong but isn't that the..." [5] "No, that's a dangerous practice in a relationship and certainly not very smart or cool imho."
>
> Please label the following posts as 1 = connective, 0 = non-connective

Second Prompt:

> Please perform a text annotation task: I will provide you with the definition of 'connective democracy', some human-labeled social media posts, and some posts to be coded. Your task is

to classify the unlabeled posts based on whether it demonstrates connective democracy or not.

Here is the definition of 'connective democracy': Connective democracy seeks to build bridges between divided groups so that they can hear each other in a deliberative manner. "Connectivity" refers to a willingness to prioritize relationships over competitiveness and engage in conversation with one's political adversaries to genuinely understand their viewpoints.

6.3 Examples

Examples of messages that were coded as connective but not polite include:

- wasn't sure wether to comment on this publicly you couldve removed that shit of the tweet my g but i hear you.
- rt ninoboxer folks this isnt over the shit show is just beginning in my opinion
- im not going to contest that the anime community isnt toxic because i have no idea but im gonna be honest with you i watch most of the mainstream anime and i have no fuckin clue whats going on in those communities and i really dont care quite frankly you dont have to be apart of the community to enjoy these shows

6.4 Replication Files

The labeled dataset, codebook, and BERT model can be found here: `https://osf.io/lxrkva/`.