Neural Computing and Applications (2023) 35:21935–21947
https://doi.org/10.1007/s00521-023-08819-z

ORIGINAL ARTICLE

Neural nets for sustainability conversations: modeling discussion disciplines and their impacts

Katrina Pugh[1] · Mohamad Musavi[2] · Teresa Johnson[2] · Christopher Burke[2] · Erez Yoeli[3] · Emily Currie[2] · Benjamin Pugh[4]

Received: 31 January 2023 / Accepted: 28 June 2023 / Published online: 1 September 2023

Abstract

We live in the age polarization, where conversations on matters of sustainability more often produce acrimony or stalemate than productive action. Better understanding conversation features and their impacts may lead to better innovation, solution-design, and ongoing collaboration. We describe a study to test alternate machine learning models for classifying six "discussion disciplines", which are conversation features associated with rhetorical intent. The model providing the best outcome used the Bi-directional Encoder Representations from Transformers (BERT) layered with a Residual Network (ResNet). The training data were 1135 utterances from Maine aquaculture town hall-like meetings and similar conversations, which had been hand-coded for the discussion disciplines. In addition, we generated 300 phrases corresponding to three conversation outcomes: Intent-to-Act, Options-Generation, and Relationship-Building. We then used the trained model and information retrieval to classify a large corpus of 591 open-source transcripts, containing over 21,000 utterances. A binary logistic regression analysis showed that two discussion disciplines, "*Inclusion*" and "*Courtesy*," had positive, statistically significant, impacts on Intent-to-act: a 10 percentage point increase in the share of the *Inclusion* or *Courtesy* yielded a 45% or 34% increase, respectively, in the likelihood of *Intent-to-Act*. This study shows the applicability of neural networks in modeling conversations and identifying the dialog acts that can provide measurable and predictable impact on conversation outcomes. Conversational intelligence can support a variety of human interactions, such as town halls, policy-deliberations, private–public partnerships, and sustainability teamwork.

Keywords Aquaculture · Conversational intelligence · Neural networks · Natural language processing · Modeling

1 Introduction

Managers, researchers and policymakers have long sought to better understand the friction in, productivity of, and durability of interacting groups (e.g. [1, 9]. For several decades, natural language processing (NLP) researchers have attempted to quantify the relationships between rhetorical intents in conversations, and their outcomes (collectively, "conversation features") [29, 22]. While "conversational AI" (chat bots) is well established, rapidly interpreting human-to-human conversation in a manner to improve innovation, motivation, and belonging ("AI for conversation") is in its infancy [4, 24].

Yet, there is a growing need: there is polarization and manipulation in social media, public meetings, and policy tables. These are having negative effects on the environment and human health, and are spreading anti-democratic behaviors [19, 26]. Sitting at the confluence of humans with their environment, sustainability discourse is vulnerable to misinformation and conflict [2, 10]. At risk in all these conversations are participants' accountability, reciprocity, and innovativeness.

The discussion disciplines are rhetorical intents that characterize the dialog acts that make up speech [27]. Our research hypothesis was that the discussion disciplines'

✉ Katrina Pugh
kp2462@columbia.edu

[1] Columbia University, New York, USA

[2] University of Maine, Orono, USA

[3] Massachusetts Institute of Technology, Sloan School of Management, Cambridge, USA

[4] Olin College, Needham, USA

shares would correlate to specific outcomes. The six discussion disciplines[1] that our research explored were Integrity, Integrity-q, Courtesy, Inclusion, Translation, and Snarky [21]. Discussion disciplines derive from MIT's four dialog practices—voice, respect, listening, and suspension—which have been shown to improve collaboration [12, 13]. To accommodate everyday speech, we augmented the dialog practices with classifications of Translation-related, Inclusion-related, and "Snarky" rhetorical intents [22]. We also studied three conversation outcomes: Intent-to-Act, Relationship-Building, and Options-Generation, as shown in Fig. 1.

We started with town hall-like meetings that were required in the Maine aquaculture lease scoping process [23]. We used these to train a neural net that would describe sustainability-related conversation. Using a large open-source corpus, we used the trained neural net model to classify the discussion disciplines. We then accurately correlated discussion discipline shares to conversations' affective and motivational outcomes. Our novel neural network model layered the Bidirectional Encoder Representations from Transformers (BERT) [6] with Residual Network (ResNet) [11]. We have shown that the BERT-ResNet model outperforms BERT alone, as well as the Term Frequency-Indirect Document Frequency (TF*IDF) [28] and [29].

In the following sections, we describe our method and its performance. The "Conversation Data Preparation" section describes the aquaculture town-hall like meetings that we hand-coded and the large corpus of open-source data. The "Modeling and Simulation" section explains the TF*IDF and BERT models, as well as our additional ResNet layer. In the "Model Application" section, we apply the BERT-ResNet model to a large corpus of utterances, and use a binary logistic regression to evaluate the impacts of the discussion disciplines' percentages on conversation outcomes. The "Discussion" section suggests new areas for neural network research to inform and improve conversations at work and in society. Finally, we conclude our research in the last section of this paper.

2 Conversation data preparation

We began our research by attending aquaculture lease "scoping session" meetings (LSMs) which are town hall-like gatherings of Maine aquaculture stakeholders, such as riparian landowners, harbor masters, boaters, and aquaculture farmers. LSMs are part of Maine's aquaculture lease-approval process governed by the Maine Department of Marine Resources [16]. In each LSM, participants debate the costs and benefits of a new lease, of lease expansion or of lease renewal, e.g., for a scallop, oyster or kelp farm. Participants discuss boat traffic, biodiversity, noise pollution, esthetics, marine navigation, livelihoods, and food security, to name a few topics. To collect transcript data, we attended seven LSMs over Zoom in Fall 2020 and Spring 2021. We manually recorded each utterance, speaker, gender, and role in the conversation. Conversation utterances were then hand-parsed into distinct "moves" (dialog acts of one or more sentences with observable, individual rhetorical intents). Single moves were then hand-coded for each of the six discussion disciplines, for a total of 728 moves.[2] These transcripts statistics are shown in row No. 1 of Table 1. Interviews with aquaculture farmers and other researchers helped validate the coding of the transcripts.

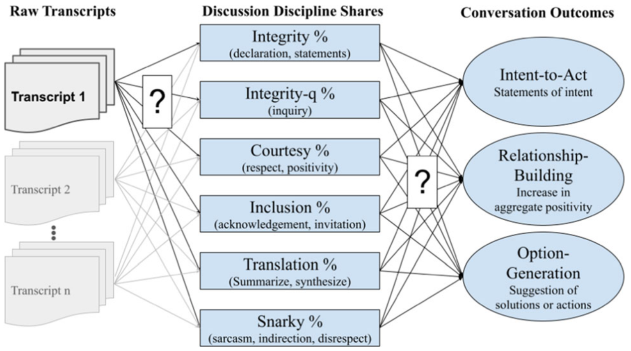


Fig. 1 Logic model for sustainability conversations modeling. *Note*: Raw transcripts for conversations contain dialog acts (utterances or moves), which can be coded as discussion disciplines. The shares of the discussion disciplines, by transcript, are mapped to specific conversation outcomes. We sought to model the classification of speech into discussion disciplines and, using a large corpus, to assess the relationship of their shares in a conversation to that conversation's outcomes

After coding transcripts for discussion disciplines, we manually observed the relationships between the discussion discipline percentages and conversation outcomes: Intent-to-Act, Relationship-Building, and Options-Generation. In our manual analysis, we found correlations between Inclusion and Intent-to-Act; between Integrity-Q and Options-Generation; between Translation and Options-Generation; and between Courtesy and Relationship-Building, as presented in Fig. 2. For each quadrant, the discussion discipline percentage in the transcript is sorted left to right, least to greatest. The outcomes, listed on the horizontal axis of each graph, are Relationship-Building (RB); Options-Generation (OG), and Intent-to-Act (ITA).

[1] Definitions for much of the vocabulary in this article are provided in the Appendix.

[2] The term "move" applies to hand coded data, which have been parsed down to single dialog acts. We use the term "utterances" to refer to the smallest unit available with the open data.

Table 1 Transcript and utterance counts for open-source and hand-coded aquaculture LSM (and similar) transcripts

Source	Number of transcripts	Number of utterances
1. Hand-coded Aquaculture Lease Scoping Meeting (LSM) Transcripts	7	728
2. Other hand-coded transcripts	4	410
Total Hand-coded	***11***	***1138***
3. Open-source coarse discourse corpus	122	4373
4. Open-source friends corpus	49	1439
5. Open-source GAP corpus	28	8009
6. Open-source movie corpus	103	3475
7. Open-source persuasion corpus	135	2793
8. Open-source tennis corpus	80	160
9. Open-source the argument podcast	74	802
Total open-source	*591*	*21,051*
Total	*602*	*22,189*

Numbers of transcripts and utterances. Open corpus transcripts (#s 3–9) can be found at Chang et al. [3] and Cornell's ConvoKit https://doi.org/10.48550/arXiv.2005.04246. These transcripts are divided into utterances. The utterances from the seven aquaculture lease scoping meetings (LSM) transcripts' and other hand-coded transcripts' (Row No.s 1 and 2) were further divided into moves (dialog acts). Each utterance or move (whichever was the smallest unit) contained one discussion discipline

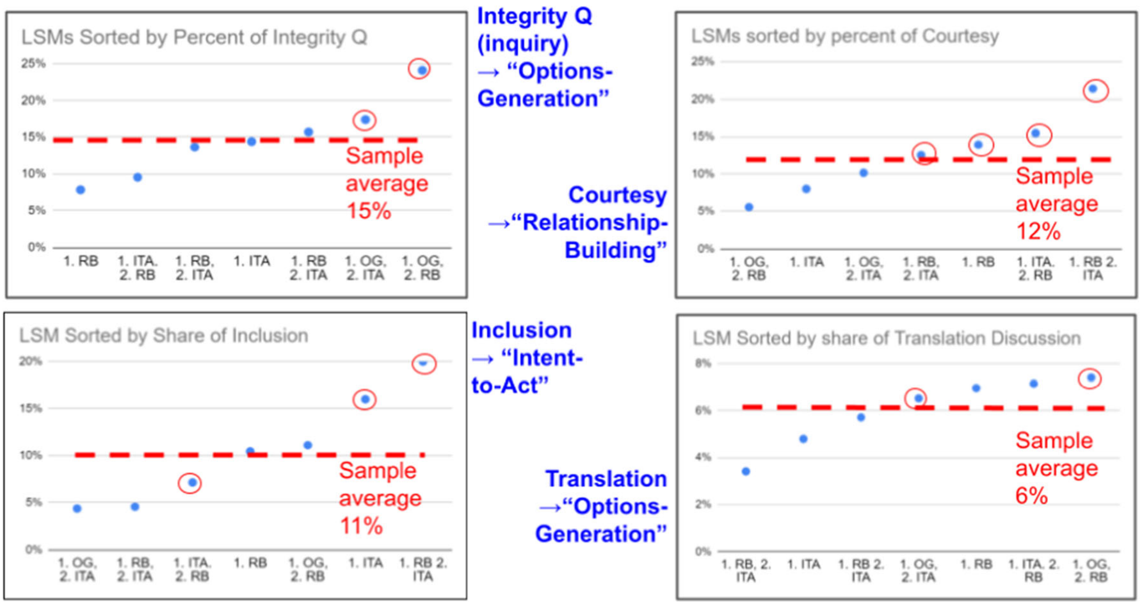


Fig. 2 Mapping between hand-coded discussion disciplines and outcomes: relationship-building (RB); Options-Generation (OG), and Intent-to-Act (ITA)

Snarky tended to reduce all outcomes. Integrity moves, at over 50% of the samples, tended to be spoken by the aquaculture farmer (who convened the LSM), and were associated more with information exchange than with collective outcome, so this is not shown in Fig. 2.

In order to prepare more training data for the neural network model to identify the discussion disciplines, we added some additional hand-coded training data. We raised the observations to 1,138 utterances by adding to the 728 aquaculture LSM data another 410 utterances from US National Archive transcripts, student online discussions, and professional community online discussions, as shown in row No. 2 of Table 1.

In order to measure patterns between the discussion disciplines and the outcomes of transcripts in the open-source data, we regressed outcomes on discussion discipline shares at the transcripts-level. For statistical significance with our six-independent variable model, we would need to classify tens of thousands of utterances for discussion disciplines, and calculate percentages inside transcripts, and then relate those percentages to at least four hundred transcript-level outcomes. We obtained these utterances from 591 open data transcripts from Cornell University's ConvoKit [3], as shown in rows No. 3 to 9 in Table 1. Approximately 400 transcript observations would be sufficient to find statistical significance. Our open-source transcripts were used first to train the TF*IDF model and then to test the hypothesis that discussion discipline percentages impact the outcomes of the conversations. These transcripts contributed 21,051 utterances. We detected the transcript-level outcomes in the open data by using information retrieval from a lookup table, which was created manually from the outcomes of the LSMs. This lookup table contained 300 phrase examples correlated to two of the outcomes: Intent-to-Act (ITA) and Options-Generation (OG) (Table 2). We used lemmatization to expand phrase examples prior to building the lookup table.

In each transcript the third outcome, Relationship-Building, was the percentage change in "net positivity," in other words, Courtesy counts plus Inclusion counts, minus Snarky counts. Thus Relationship-Building was calculated as net positivity for the second half, less net positivity for the first half, divided by net positivity in the first half.

Figure 3 is a visualization of the data and modeling pipeline, as explained below.

3 Modeling and simulation

Based on our previous research into the discussion disciplines [22], we theorized that all conversations could be classified into the six discussion disciplines. Our initial goal was to generate a model that would categorize utterances into clusters which would align with the discussion disciplines. To do this, we used three different models, as briefed below.

Table 2 Outcomes look up table illustration

	Phrase present?	Assigned outcome
Transcript 1	"I will…" (and variants)	Intent-to-Act (ITA)
Transcript 2	"Let's try…" (and variants)	Options-Generation (OG)
Transcript n	"Have we tried…?" (and variants)	Options-Generation (OG)

Using a 300-phrase lookup table, outcomes of intent-to-Act (ITA) or Options-Generation (OG) were assigned at the transcript level

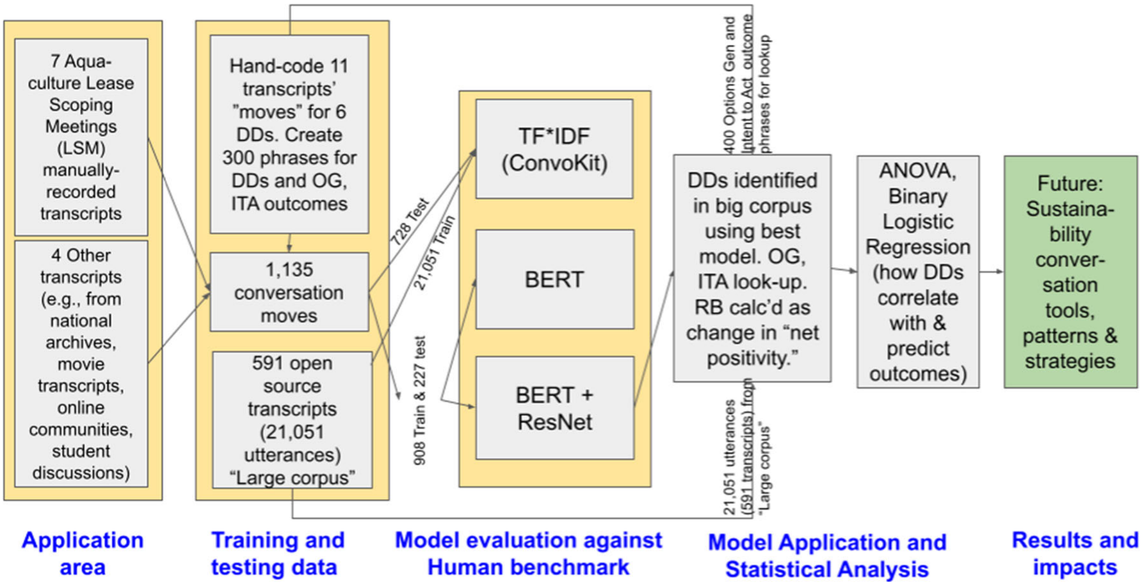

7 Aqua-culture Lease Scoping Meetings (LSM) manually-recorded transcripts
4 Other transcripts (e.g., from national archives, movie transcripts, online communities, student discussions)
Hand-code 11 transcripts' "moves" for 6 DDs. Create 300 phrases for DDs and OG, ITA outcomes
1,135 conversation moves
591 open source transcripts (21,051 utterances) "Large corpus"
728 Test
21,051 Train
908 Train & 227 test
TF*IDF (ConvoKit)
BERT
BERT + ResNet
400 Options Gen and Intent to Act outcome phrases for lookup
21,051 utterances (591 transcripts) from "Large corpus"
DDs identified in big corpus using best model. OG, ITA look-up. RB calc'd as change in "net positivity."
ANOVA, Binary Logistic Regression (how DDs correlate with & predict outcomes)
Future: Sustaina-bility conver-sation tools, patterns & strategies
Application area
Training and testing data
Model evaluation against Human benchmark
Model Application and Statistical Analysis
Results and impacts

Fig. 3 Pipeline for the Sustainability Conversation for Impact project. Note: Figure includes the data preparation, count of utterances or "moves," three modeling approaches (TF*IDF, BERT and BERT + ResNet), model evaluation against manually coded data, and statistical analysis (outcomes regressed on DD percentages, by transcript). "DD" = Discussion Discipline; "OG" = Options-Generation, "ITA" = Intent-to-Act; "RB" = Relationship-Building

3.1 TF*IDF-based classification model

The Term Frequency-Indirect Document Frequency (TF*IDF) transformer model [8] was chosen for its computational transparency, as it derives from the search engine optimization process. TF*IDF begins with word embedding, which is a transformation of text into a numerical vector of *m* dimensions, where *m* is the number of unique elements, or "tokens" (word, phrase, word-group, phrase-group) for each utterance. Therefore, for *n* utterances, we will have a numerical matrix of *mxn* where *n* is the number of rows and *m* is the number of columns. Table 3 shows a simple word embedding of two utterances adapted from the aquaculture transcripts: "I'm hearing no consensus, so I recommend we bring out further options." and "It is obvious that we need further information to reach consensus." In the example, after stripping out punctuation, each word is associated with a unique category ("index"), the cell taking the value of the frequency with which the word is present in the sentence [18]. This simple embedding does not capture context, and it becomes computationally inefficient for large vocabularies. To address this, we used the TF*IDF model that relies on a fixed number of tokens (*m*) derived from the English dictionary.

TF*IDF in our context uses "Terms" which are tokens and "Documents" which are utterances, as described above. TF*IDF starts with vectors of token counts for each utterance. For a token *t* in utterance *d*, the weight ($W_{t,d}$) is given by:

$$W_{t,d} = TF_{t,d} \log\left(n/DF_t\right)$$

Where $TF_{t,d}$ is the number of occurrences of *t* in utterance *d*, DF_t is the number of utterances containing token t, *n* is the total number of utterances in the corpus, *m* is the total number of tokens.

DF_t in the denominator, reduces the size of the natural log. Not surprisingly, as the TF*IDF model comes from search engine optimization (SEO), one would want a high-occurrence, or salience, of the term in each utterance (numerator), and low occurrence (rarity in the corpus) (denominator) [8]. What this inverse relationship means is that TF*IDF finds the term to be *dominating* with respect to the utterance, but *rare or salient* with respect to the corpus.

Our TF*IDF based model used Cornell's Convokit "PromptTypesWrapper" [3] to find token similarities across utterances in the corpus. We computed in-utterance term frequency (TF) relative to the inverse of corpus-based term frequency (IDF). In our case, our tokens were "phrasing motifs": We started with pairs of dependency-related words (or "bi-grams"). Where pairs of bi-grams were frequently observed, they are called "phrasing motifs" [14].

Table 3 Illustration of word embedding for two sentences

Word	Am	Bring	Consensus	Further	Hearing	I	Information	is	Need	No	Obvious	Options	Out	Reach	Recommend	So	That	To	We
S1	1	1	1	1	1	2				1		1	1		1	1			1
S2			1	1			1	1	1		1			1			1	1	1

We generated an *mxn* matrix, where *m* (columns) was the number of unique phrasing motifs for an utterance and *n* (rows) was the number of unique utterances. (When phrasing motifs are used as the basis for counting the unique element, we add back a certain degree of rhetorical context, leveraging the dependency-related nature of our bi-gram tokens, that, in turn, combine into phrasing motifs. Frequently occurring phrasing motifs used as the columns in the TF*IDF matrix decreased the vocabulary and, thus, sparsity of the matrix.)

We fed into TF*IDF the 21,051 utterances from the open-source transcripts (See Table 1). Transcripts of multiple utterances were selected based on their corpus-similarity to the LSM conversations [17]. For our corpus, we reduced the matrix using the Singular Value Decomposition (SVD) process and then used a K-means clustering process to arrive at six clusters with common meaning for the six discussion disciplines.

To make the unsupervised K-means clusters into a supervised classifier, we fed the TF*IDF model the 728 hand-coded LSM utterances. The clusters were each assigned to a discussion discipline by looking at the largest percent of any discussion discipline that was close to the sigmoid of the cluster. Then, those cluster and discussion disciplines were set aside. The next largest percent of a discipline in a cluster was identified, and that cluster and that discussion discipline were set aside. This continued until all of the clusters were labeled with a discussion discipline. Where there were conflicts, the more frequently occurring discussion disciplines were favored. For the LSM transcripts, the average percentages were Integrity (51%), Integrity-Q (15%), Courtesy (12%), Inclusion (11%), Translation (6%) and Snarky (5%). Thus, if both Integrity and Courtesy had the majority of their utterances in cluster 1, Integrity would be the assigned cluster. This maintained a conservative approach to the scorecard, as described below. The hand-coded LSM transcripts were then used to test the model by running the LSM moves through the model and validating the coding match. This entailed starting with the reduced matrix and then re-running the k-means clustering.

Based on low results (42%), we added a lookup (information retrieval) process. Using the 300-phrase dictionary of discussion disciplines we used simple information retrieval (lookup) to locate and append discussion disciplines before running the TF*IDF process. (The phrase match and append occurred for approximately 20% of the utterances.) We then parsed the utterances into phrasing motifs as usual. The TF*IDF matrix was then generated, and then the SVD and clustering were performed.

Next, to accommodate the asymmetrical distribution of the discussion disciplines, we evaluated the use of a Poisson normalization within the SVD process. This involved taking all nonzero column values in the TF*IDF matrix and dividing them by the square root of (cell value + 1) and then subtracting the mean of the related column. Then, SVD was performed, and the Poisson/mean step was repeated in reverse. This was followed by the K-Means clustering as usual.

The scorecard is shown in Table 4. Ultimately, we found that appending the discussion disciplines in 20% of the cases helped, but the Poisson normalization did not improve the overall performance. The TF*IDF variants' overall accuracy did not surpass 45.2%.

3.2 BERT-based classification model

The Bi-directional Encoder Representations from Transformers (BERT) is a neural network open-sourced by Google in 2018, described by Devlin and Chan (2017). Built on top of another of Google's open-sourced applications, TensorFlow, BERT was trained on Google Search and Wikipedia, and was intended for classifying speech (e.g., for sentiment analysis). The ancestor of

Table 4 Scorecard for model performance against hand-coded utterances data: TF*IDF, BERT, and BERT + ResNet

NLP model	Lookup/ append discussion disciplines [1]	Poisson normalization [2]	Percent of all moves correctly categorized	Integrity	Integrity-Q	Courtesy	Inclusion	Translation	"Anti" (Snarky)
TF*IDF	No	No/Yes	42.0%/39.3%	74.8%/ 73.5%	9.3%/ 9.1%	12.9%/–%	–%/6.2%	9.1%/8.1%	5.4%/–%
	Yes	No/Yes	**45.2%/44.5%**	**84.9%/85.1%**	6.5%/5.1%	4.3%/–%	–%/7.3%	–%/–%	5.4%/ 7.1%
BERT [3]	No	NA	85%	**99%**	**100%**	**98%**	**93%**	0%	0%
BERT + ResNet [4]	No	NA	**95.2%**	**98.4%**	**100.0%**	**91.7%**	**95.8%**	**100.0%**	64.7%

Model performance for TF*IDF and neural net BERT or BERT + ResNet. [1] Appended metadata. [2] Poisson normalization. [3] BERT alone. [4] BERT with a ResNet layer with random node exclusion during iterations to reduce overfitting the ChatGPT large language model, BERT's transfer learning leverages a pre-trained general-purpose model, which can be used to train on new, labeled data. BERT uses neural network layers that are derived from self-attention in the sentence or utterance ("contextual" self-attention), combined with look-ups ("non-contextual" self-attention). For example, contextual and non-contextual elements allow BERT to recognize paraphrases [7]. Devlin and Chan [6] enumerate BERT's capabilities: Word sense disambiguation, polysemy resolution (e.g., "river bank," "rob a bank"), named entity determination, textual entailment / next sentence prediction, coreference resolution, question answering, and automatic summarization.

For our initial BERT-only model, we used 80%/20% data distribution for training and validation using the 1138 hand coded utterances. The model generated a base-BERT 768-dimension word embedding, which is the standard for the model. This resulted in 85% accuracy, with poor performance on the less-abundant disciplines of Translation and Snarky (Table 4).

3.3 BERT-ResNet based classification model

To improve upon the discussion discipline classification performance for TF*IDF and BERT, we enhanced the BERT model with ResNet (Residual Neural Network), which is a deep neural network developed in 2015 for the purpose of image classification [11]. Being highly convolutional makes ResNet valuable for feature extraction where the training data are limited for any variable. As ResNet uses transfer learning, like BERT, it starts with basic knowledge that could be fine-tuned for the discussion disciplines. Figure 4 contains our BERT-ResNet model sequence.

With ResNet, we initially took the BERT embeddings and transformed them into a 30 × 30 matrix, as shown in the left-hand side of Fig. 4. We used overlapping (stacked) sections of the embedding as rows in the output matrix.

A typical padding process of a ResNet model is shown in Fig. 5.

Given that ResNet expects a three-dimensional tensor, i.e., red–green–blue (RGB) images, the matrix was stacked on itself three times. After the stacking, a three-dimensional average pooling layer was used to downsample the input. An example of how to generate a two-dimensional average pooling layer with a 2 × 2 filter and a (2,2) stride is shown in Fig. 6.

The input was down-sampled to (3, 28, 28) (Fig. 4). This was then passed to the ResNet model. The output of ResNet has a shape of (5, 5, 2048). This output was passed to a 2-dimensional average pooling layer to down-sample the input to have dimensionality of (1, 1, 2048). This output was then flattened to a dense layer of just 2048 nodes. Subsequent steps reduced the size of the layer until ultimately reaching the classification layer with six sigmoid nodes. Rectified Linear Unit (ReLU) activation functions and two dropout layers were used to reduce the computational intensity of the backpropagation algorithm and to protect against overfitting, respectively. Finally, the sigmoid layer was used to detect the presence of the six discussion disciplines. Each node had a sigmoid activation function for identifying the six discussion disciplines,

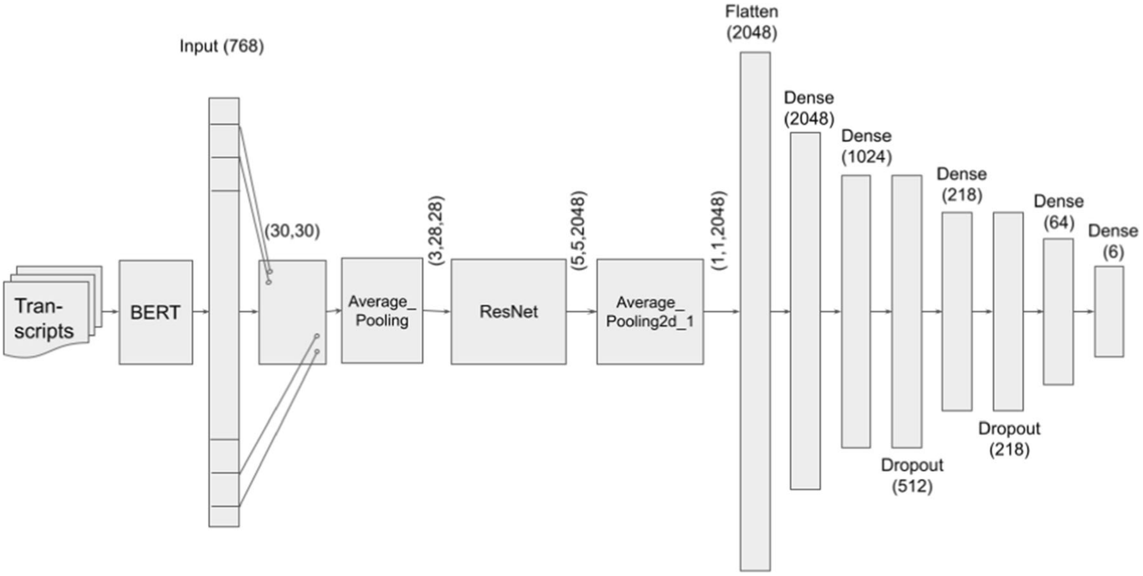


Fig. 4 BERT-ResNet model Sequence. Note: Shows matrix dimensions, and drop-out processes. Rectified Linear Unit (ReLU) activation functions reduced computational intensity of the back-propagation algorithm, and two dropout layers protected against overfitting

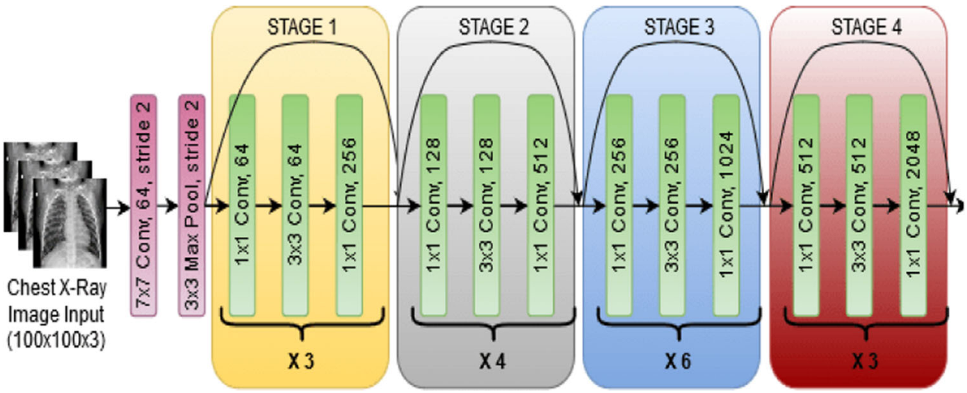


Fig. 5 Four stage padding process for ResNet model used for image recognition

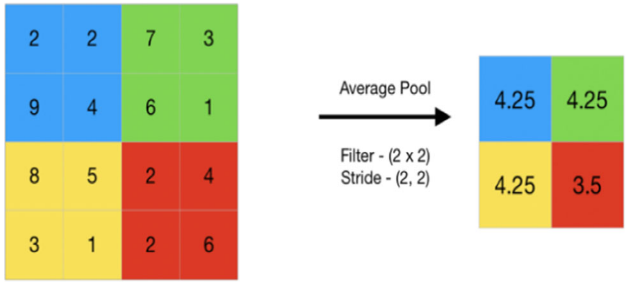


Fig. 6 Illustration of pooling with ResNet. *Note*: Pooling using 2 × 2 filter and a (2,2) stride

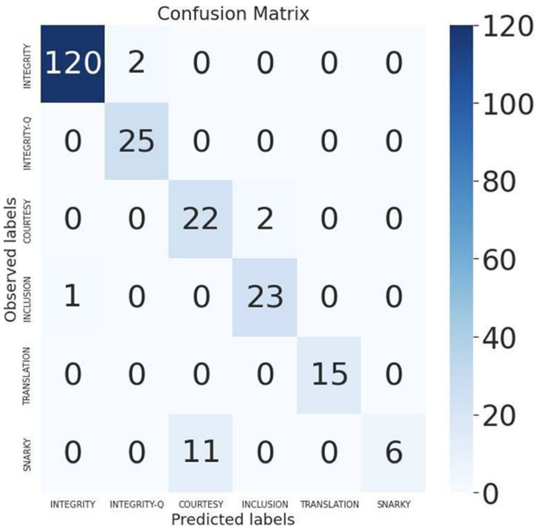
which varied between 0 and 1. The nodes with the highest values were chosen as the "winning" discussion discipline.

The hand-coded 1135 training data samples were randomly broken up into two groups for training and validation, at 80%-20% proportions (908 training and 227 validation). The distribution was checked to make sure that the model would see all types of the various classes in both datasets.

Initially, all layers were made trainable for the first five epochs (learning cycle sweep through training data), after which all of the ResNet layers were frozen (no longer trainable), to prevent the ResNet model from overfitting. The remaining 45 epochs trained the average pooling and dense layers in the model that were not part of ResNet.

Compared to BERT-alone, at 85%, the BERT + ResNet combination improved accuracy substantially to 95% as shown in Table 5. The discussion discipline with the highest misclassification was Snarky, with an accuracy of 45%. This could be due to the complex nature of Snarkiness, such as sarcasm (e.g., criticism masked as positivity, as in [15]), indirect speech, or innuendo, which can be hard to detect, even for humans. In fact, in our manual analysis, indirection was a factor in hand-coding several LSMs, where participants in the conversation made oblique references or used sarcasm.

Table 5 Confusion matrix for the final BERT + ResNet model.



Confusion Matrix

Observed labels \ Predicted labels	INTEGRITY	INTEGRITY-Q	COURTESY	INCLUSION	TRANSLATION	SNARKY
INTEGRITY	120	2	0	0	0	0
INTEGRITY-Q	0	25	0	0	0	0
COURTESY	0	0	22	2	0	0
INCLUSION	1	0	0	23	0	0
TRANSLATION	0	0	0	0	15	0
SNARKY	0	0	11	0	0	6

120
100
80
60
40
20
0

After using 908 hand coded moves for training, we used 227 validation. 216 out of 227 moves (95%) were correctly classified

4 Model application

Using the BERT-ResNet model, which yielded the best result, we ingested 21,051 utterances from 591 open-source transcripts, as shown in Table 6. Table 6 shows that the open-transcripts' outcomes distribution generally matches the hand-coded data, except for Relationship-Building, which was higher in the LSM data. (Recall that the data and modeling pipeline, including our data and open-source data, was presented in Fig. 3.)

With a binary logistic regression statistical model, using the open-source data processed by the BERT-ResNet model, we regressed outcomes of each transcript on the percentages of each of the discussion disciplines, by transcript. Correlations between several of the discussion disciplines, and between the discussion disciplines with the

Table 6 Transcript counts, outcomes, and gender, open-source v. hand-coded

Source	Number of transcripts	Number of utterances	Female utterances	Transcripts with intent-to-act outcome	Transcripts with relationship- building outcome	Transcripts with options-generation outcome
1. Coarse Discourse Corpus	122	4,373	NA	113	42	39
2. Friends Corpus	49	1,439	NA	44	24	9
3. GAP corpus	28	8,009	NA	19	10	5
4. Movie Corpus	103	3,475	960	87	42	48
5. Persuasion Corpus	135	2,793	1,388	91	86	49
6. Tennis Corpus	80	160	NA	6	51	1
7. The Argument Podcast	74	802	0	70	22	16
Total	591	21,051	2,348	430	277	167
			11% of utterances	73% of transcripts	47% of transcripts	28% of transcripts
8. LSM transcripts	7	728	152	5	5	2
			21% of utterances	71% of transcripts	71% of transcripts	29%of transcripts
9. Other hand-coded transcripts	4	410	NA	NA	NA	NA

Numbers of transcripts, utterances, female shares (where available), and outcomes. Large open corpus transcripts can be found at Chang et al. [3] and ConvoKit https://doi.org/10.48550/arXiv.2005.04246

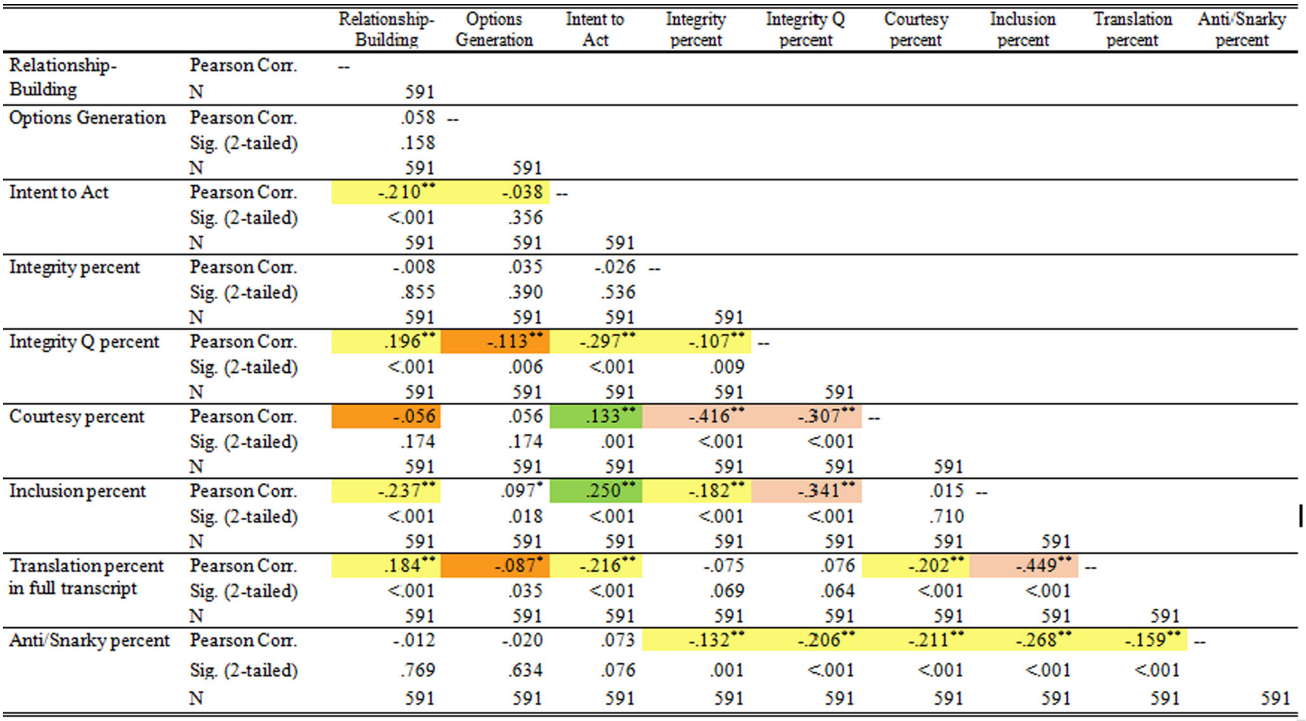
outcomes showed a meaningful relationship between Inclusion and Courtesy and Intent-to-Act, as was seen in the manual data. This is shown in the Pearson Correlation matrix, Table 7.

Inclusion and courtesy are both correlated to Intent-to-Act, and Integrity and Translation are both negatively correlated with Intent-to-Act. Negative correlations between Integrity Q and Translation with Options-Generation are unusual. and may be due to Translation's negative correlations with Courtesy and Inclusion. Naturally, Snarky is negatively correlated with the five other discussion disciplines.

Due to Collinearity, we evaluated combinations of the variables (discussion discipline percentages, and outcomes detected), to determine the Binary Logistic Regression experiment with the most explanatory power. Table 8 and 9 show the most successful experiment. Table 8 indicates that the discussion disciplines can predict the presence of an Intent-to-Act 98.6% of the time and the absence of Intent-to-Act at 29.8%, with a cut value of 0.5 for the function, for an overall classification accuracy of 79.9%. This result presents an increase of 7.1% over the base case of 72.8% (a guess "yes" for Intent-to-Act, the actual overall share in the open-source transcripts in the third-to-last column in Table 5, above row No. 9). The 29.8% is the "sensitivity," and the 98.6% is the "specificity." Table 9 presents coefficients of the binary logistic regression of Intent-to-Act on the discussion disciplines. We see a positive statistically significant explanatory power of Inclusion and Courtesy. Columns indicate the standardization process. The Betas in the first column are unstandardized. The "Wald Statistic" is the quotient of Beta divided by Standard Error, and then squared. The Expected (B) (or "Exp(B)") is the odds ratio. Each odds ratio in this table indicates the multiplicative change in the odds (of a case falling into the Intent-to-Act target, an output of 1), per unit increase on a given predictor, controlling for the other predictors in the model. If the odds ratio, Exp(B), is 1, it indicates that there is no change in the impact to the dependent variable per unit impact in the predictor. If the odds ratio, Exp(B), is greater than 1, then it indicates that the odds associated with target group (Intent-to-Act) membership are increasing. If it is less than 1, then it indicates that the odds associated with the target group (Intent-to-Act) membership are decreasing.

Using the Exp(B) values, we can interpret the table. When the Inclusion percentage increases by 10 percentage points, we increase the odds of Intent-to-Act by 45% (= 10*(1.045–1)). When the Courtesy increases by 10 percentage points, we increase the odds of Intent-to-Act by

Table 7 Pearson correlation for five discussion disciplines (plus snarky) and three outcomes



		Relationship-Building	Options Generation	Intent to Act	Integrity percent	Integrity Q percent	Courtesy percent	Inclusion percent	Translation percent	Anti/Snarky percent
Relationship-Building	Pearson Corr.	--								
	N	591								
Options Generation	Pearson Corr.	.058	--							
	Sig. (2-tailed)	.158								
	N	591	591							
Intent to Act	Pearson Corr.	-.210**	-.038	--						
	Sig. (2-tailed)	<.001	.356							
	N	591	591	591						
Integrity percent	Pearson Corr.	-.008	.035	-.026	--					
	Sig. (2-tailed)	.855	.390	.536						
	N	591	591	591	591					
Integrity Q percent	Pearson Corr.	.196**	-.113**	-.297**	-.107**	--				
	Sig. (2-tailed)	<.001	.006	<.001	.009					
	N	591	591	591	591	591				
Courtesy percent	Pearson Corr.	-.056	.056	.133**	-.416**	-.307**	--			
	Sig. (2-tailed)	.174	.174	.001	<.001	<.001				
	N	591	591	591	591	591	591			
Inclusion percent	Pearson Corr.	-.237**	.097*	.250**	-.182**	-.341**	.015	--		
	Sig. (2-tailed)	<.001	.018	<.001	<.001	<.001	.710			
	N	591	591	591	591	591	591	591		
Translation percent in full transcript	Pearson Corr.	.184**	-.087*	-.216**	-.075	.076	-.202**	-.449**	--	
	Sig. (2-tailed)	<.001	.035	<.001	.069	.064	<.001	<.001		
	N	591	591	591	591	591	591	591	591	
Anti/Snarky percent	Pearson Corr.	-.012	-.020	.073	-.132**	-.206**	-.211**	-.268**	-.159**	--
	Sig. (2-tailed)	.769	.634	.076	.001	<.001	<.001	<.001	<.001	
	N	591	591	591	591	591	591	591	591	591

Analysis of Variance (ANOVA) generated in IBM SPSS. Intent-to-act and Courtesy have a positive correlation (with 99% confidence level) with Inclusion, and a negative correlation with Snarky. Colors represent outcomes consistent with (Green) or inconsistent with (Orange) the manual LSM analysis in Fig. 3. Pink are strong correlations

Table 8 Classification accuracy of binary logistic regression of intent-to-act on discussion disciplines

Classification table				
Observed		Predicted		
		Intent-to-act		Percentage Correct
		0	1	
Intent-to-act	0	48	113	29.8
	1	6	424	98.6
Overall percentage				79.9

34% (= 10*(1.034–1)). Snarky had a surprising large positive coefficient, which suggests it may be picking up other omitted collinear variables, as suggested in Table 7.

As with the manual analysis between discussion disciplines and outcomes (Fig. 3), our results suggest that Inclusion (acknowledgement) has the biggest impact on Intent-to-Act. Being included or acknowledged may arouse a sense of being recognized, and thus a desire to be accountable and/or to take action. In our aquaculture data, Intent-to-Act appeared in a number of ways, such as a statement by the aquaculture farmer about an intent to move their scallop or oyster lease coordinates to reduce

Table 9 Binary logistic regression of Intent-to-Act conversation outcome on discussion disciplines

	B	S.E	Wald	df	Sig	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Integrity percent	.025	.008	9.105	1	.003	1.025	1.009	1.041
Courtesy percent	.033	.008	19.184	1	< .001	1.034	1.018	1.049
Inclusion percent	.044	.007	43.803	1	< .001	1.045	1.031	1.058
Snarky percent	.034	.008	19.328	1	< .001	1.034	1.019	1.050
Constant	− 1.856	.445	17.368	1	< .001	.156		

Data generated in SPSS for best-performing binary logistic regression model navigation-obstruction risks. Intent-to-Act could also be other participants' statements that they would share information, such as their fishing methods, research outcomes, or land investment plans. Courtesy (positivity, pro-sociality) may have added to the overall sense of mutuality and conscientiousness.

5 Discussion

Our findings suggest that using a combination of BERT and ResNet for the discussion discipline detection, and using a rules-based process for outcomes detection, can profile conversations accurately, and that there is some evidence relating Inclusion and Courtesy to Intent-to-Act. For example, in our aquaculture domain, we can predict how Inclusion statements like, "Please tell us your concern about protecting navigation channels," could yield an Intent-to-Act by some participant in the conversation, such as the farmer's intent to relocate their aquaculture lease, or community members' intent to participate in a future lease siting study. In interviews with the aquaculture farmers, they expressed surprise (and some comfort) about the impact of conversation rhetoric, independent of professional facilitation. This was an empowering outcome for them. They felt that combining qualitative (hand-coded conversation) and these neural net insights could improve community members', farmers' and policymakers' toolkits for reducing conflict, and thus improve the outcomes of similar conversations.

While these results are promising, we see four areas for future research. First, differences in the appearance of conversation features across domains and cultures can yield different model formations. For example, the precise language of courtesy (positivity) may differ in a business versus social community. Those differences necessitate new training data and require another test of corpus similarity for the large-corpus statistical analysis [17].

Second, our model detects the *presence* of any of the six discussion disciplines but does not assess their *magnitude*. That is to say, the model has no sense of "how" snarky, inclusive, or inquiring an utterance is. It would be beneficial to look at the boundary conditions for discussion discipline labels. The weak instances of the discussion disciplines were not measured in our study, and just as a Snarky move could have a more negative conversation outcome than witty sarcasm, a vehement Inclusion move might improve Intent-to-Act more than a weak Inclusion move.

Third, our open data transcripts were divided into utterances, whereas the utterances of our training data (LSM transcripts and our other hand-coded transcripts) were further divided into "moves" (a single dialog act, containing one discussion discipline). The BERT-ResNet model while working with the open-source transcripts took the most likely discussion discipline contained in each utterance. It is possible that singularly labeling a compound utterance (for example, an utterance containing both Courtesy and Translation) could cause us to miss the nuanced impacts of each discussion discipline on outcomes. Żelasko et al. [27] suggesting that utterances consist of multiple dialog acts, recommends coding for punctuation, such as commas and colons.

Finally, context-dependence may change discussion disciplines' meaning. For example, different language may be used with larger groups, with more or less familiarity, or more or less hierarchical social-cultural contexts. Context also supports double-entendre: Signaling [25] like gestures of generosity, or indirect speech [20] like vague accusations of "invasive species" were mentioned occasionally in the LSMs. Some approaches, such as prompt-response designations as inputs [28] and the use of the self-attention layers [27], are attempting to address such context.

6 Conclusion

In this paper, we tested neural net models to classify six rhetorical intents in conversations, called discussion disciplines. By combining BERT and ResNet, we achieved a 95.2% accuracy rate relevant to human-coded data, surpassing pure BERT by over 10 percentage points. We applied the best model to a large, open-source corpus of transcripts to explore the relationship between discussion disciplines and outcomes, and found Inclusion and Courtesy to be significant determinant of Intent-to-Act. We suggest that incorporating discussion discipline intensity, splitting utterances into "moves," and incorporating measures of context may improve both the accuracy of the classification, as well as the usefulness of the model across settings and cultures. We also suggest applying our model pipeline with new training data for unique language-cultures.

There is great opportunity to train a similar neural network model with different sustainability conversation scenarios, such as climate change, PFAS contamination, and water scarcity. So often well-intended policymakers, citizens, managers and scientists are held back by an unawareness of their rhetorical impacts in conversations. Our hope is to assist sustainability teams, policymakers, and citizens toward conversations with productive outcomes.

Appendix: Definitions

Arc: Mathematically generated dependency pair of words (bi-grams) in a language. Arcs are similar to grammatical forms, but may not rely on word sequence.

BERT: (Bi-Directional Encoder Representations from Transformers) Google's open-sourced NLP modeling tool using neural network layering and transfer learning to compute word meaning in context.

ConvoKit PromptTypes Wrapper ("ConvoKit"): A set of transformers (programs) and conversation text corpuses open sourced by Cornell University in 2020 to enable conversation-based NLP processing.

Corpus: Collection of transcripts containing utterances, which, in turn, may each contain multiple moves.

Dialog Act: A gesture inside of an utterance that expresses a single rhetorical intent, such as an opinion, statement, question, or invitation. (Several dialog acts may make up an utterance.)

Discussion Discipline: Rhetorical intents that characterize the dialog acts that make up speech. The six discussion disciplines are Integrity (declaration), Integrity-Q (question), Courtesy (positivity), Inclusion (acknowledgement), Translation (synthesis, summary) and Snarky (behaviors contrary to the first five)..

Embedding: Transformation of text into a numerical vector of m dimensions, where dimensions is the number of unique elements, or "tokens" (word, phrase, word-group, phrase-group) for which each of n documents (or, in our case, utterance) is represented as rows or observations. Embeddings make it easier to do machine learning on large inputs like sparse vectors representing words.

Lease Scoping Meeting (LSM): Also called lease scoping "sessions." Town hall-like gatherings of aquaculture stakeholders, such as riparian landowners, harbor masters, boaters, and aquaculture farmers. LSMs are part of Maine's aquaculture lease-approval process governed by the Maine Department of Marine Resources.

Move: Dialog act with a single rhetorical purpose. One or more moves makes up utterances. For example, "I am going to the store for you. Do you have your wallet on you?" is two moves that can be classified with discussion disciplines: Integrity (statement) and Integrity-Q (question).

Phrasing Motif: Commonly occurring pair of arcs used in the Convokit PromptTypesWrapper.

PCA: Principal Component Analysis.

ResNet (Residual Network): A neural network using transfer learning based on graphical image processing. Being highly convolutional makes ResNet valuable for feature extraction where the training data are limited for any variable.

Rhetorical Intent: The goal of a sentence or phrase in a conversation, such as to express an opinion, to provide information, to ask a question, to make an invitation, or to evoke emotion.

Term Frequency, Indirect Document Frequency (TF*IDF): A transformation method for finding terms in content by creating an embedding, and then modeling terms based on their frequency in a document, and their scarcity in a corpus.

Transformer: Program that manipulates (e.g., parses, combines, counts) text and applies metadata.

Token: The smallest fragment of conversation used for computing the NLP model.

Utterance: One speaker's statement in a transcript, similar to one reply in a theatrical dialog.

Author contributions KP, MM, TJ, and EY contributed to conceptualization; KP, MM, and EY contributed to methodology; KP, EC, CB, and BP contributed to formal analysis and investigation; KP and CB contributed to writing—original draft preparation; MM, KP, and EC contributed to writing—review and editing.: MM contributed to resources; MM and KP contributed to supervision.

Funding Partial funding for the research was supported by the National Oceanographic and Atmospheric Association (NOAA) National Sea Grant Award #NA18OAR4170103 and the USDA National Institute of Food and Agriculture, Hatch project number #ME022012 through the Maine Agricultural and Forest Experiment Station. Additional support came from the University of Maine School of Marine Sciences and Maine College of Engineering and Computing.

Data availability The open data sets (transcripts) used during the current study are available at ConvoKit, Chang et al. [3] and the Cornell ConvoKit site. See https://doi.org/10.48550/arXiv.2005.04246. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest Authors declare that they do not have any known financial, editorial, or personal conflict of interest.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-023-08819-z.

References

1. Almaatouq A, Alsobay M, Yin M, Watts DJ (2021) Task complexity moderates group synergy. Proc Natl Acad Sci USA 118(36):1–9. https://doi.org/10.1073/pnas.2101062118
2. Bago B, Rand D, Pennycook G (2021) Reasoning about climate change. PsyArXiv 1–42. https://psyarxiv.com/vcpkb/
3. Chang J, Chiam C, Fu L, Wang A, Zhang J, Danescu-Niculescu-Mizil C (2020) ConvoKit: a toolkit for the analysis of conversations. ArXiv. http://arxiv.org/abs/2005.04246 https://doi.org/10.48550/arXiv.2005.0424
4. Chang J, Schluger C, Danescu-Niculescu-Mizil C (2022) Thread with caution: proactively helping users assess and deescalate tension in their online discussions. In: Proceedings of CSCW 2022
5. Danescu-Niculescu-Mizil C, Lee L, Pang B, Kleinberg J (2012) Echoes of power: language effects and power differences in social interaction. In: WWW'12—Proceedings of the 21st annual conference on world wide web, pp 699–708. https://doi.org/10.1145/2187836.2187931 and https://arxiv.org/pdf/1112.3670.pdf
6. Devlin J, Chan M (2018) Open sourcing BERT: state-of-the-art pre-training for natural language processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
7. Google (2021) Transfer learning and transformer models. Google ML Tech Talks. Retrieved March 4, 2022 from https://www.youtube.com/watch?v=LE3NfEULV6k.
8. Goralewicz B (2021) The TF*IDF algorithm explained, Onely, Retrieved June 13, 2021 from https://www.onely.com/blog/what-is-tf-idf/
9. Hansen M (2009) Collaboration: how leaders avoid the traps, build common ground, and reap big results. Harvard Business Press, Boston
10. Hart D (2018) Teamwork is the new leadership. In: Maine policy review, vol 27. https://digitalcommons.library.umaine.edu/mpr/vol27/iss1/10.
11. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://www.semanticscholar.org/paper/Deep-Residual-Learning-for-Image-Recognition-He-Zhang/
12. Isaacs W (1999) Dialogue and the art of thinking together. Princeton Press, Princeton
13. Isaacs W (2016) Trim-tab dialogues: transformative vision and action in South Asia. In: The World Needs Dialogue! Issue February, pp 1–19
14. Jurafsky D, Martin J (2014) Dependency parsing. Cognit Technol 9783642414633:403–437. https://doi.org/10.1007/978-3-642-41464-0_13
15. Kumar A, Narapareddy VT, Srikanth VA, Malapati A, Neti LBM (2020) Sarcasm detection using multi-head attention based bidirectional lSTM. IEEE Access 8:6388–6397. https://doi.org/10.1109/ACCESS.2019.2963630
16. Maine Department of Marine Resources (2023) Chapter 2: Aquaculture lease regulations. https://www.maine.gov/dmr/rules-enforcement/regulations-rules. Retrieved on May 20, 2023.
17. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. Proc Natl Conf Artif Intell 1:775–780
18. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Demeester T, Rocktäschel T, Riedel S (2016) Lifted rule injection for relation embeddings. EMNLP 2016—Conference on empirical methods in natural language processing, proceedings, pp 1389–1399. https://doi.org/10.18653/v1/d16-1146
19. Pennycook G, Rand D (2021) The psychology of fake news. Trends Cognit Sci. https://doi.org/10.1016/j.tics.2021.02.007
20. Pinker S, Nowak M, Lee J (2008) The logic of indirect speech. Proc Natl Acad Sci USA 105(3):833–838
21. Pugh K (2022) Sustainability conversation for impact: transdisciplinarity on four scales. Electronic Theses and Dissertations 3608. https://digitalcommons.library.umaine.edu/etd/3608
22. Skifstad S, Pugh K (2014) Beyond netiquette: discussion discipline drives innovation (Chapter 8). In Pugh K (eds) Smarter innovation: using interactive processes to drive better business results. Ark Group, Wilmington
23. Sadusky H, Brayden G, Zydlewski, Belle S (2022) Maine Aquaculture Roadmap 2022–2032. Maine Sea Grant. Retrieved from https://seagrant.umaine.edu/wp-content/uploads/sites/467/2022/01/Maine-Aquaculture-Roadmap-2022.pdf on March 13, 2022.
24. See A, Roller S, Kiela D, Weston J (2019) What makes a good conversation? How controllable attributes affect human judgments. Facebook Research and Stanford University. https://parl.ai/projects/
25. Spence M (1973) Job market signaling. Quart J Econ 87:355–374. https://doi.org/10.2307/1882010
26. Voelkel et al (2022) Megastudy identifying successful interventions to strengthen Americans' democratic attitudes. In prep. Stanford University, Stanford. Contact: willer@stanford.edu and jvoelkel@stanford.edu.
27. Żelasko P, Pappagari R, Dehak N (2021) What helps transformers recognize conversational structure? Importance of context, punctuation, and labels in dialog act recognition. Trans Assoc Comput Linguist 9:1179–1195. https://doi.org/10.1162/tacl_a_00420
28. Zhang A, Culbertson B, Paritosh P (2017) Characterizing online discussion using coarse discourse sequences. In: Proceedings of the 11th international conference on web and social media, ICWSM 2017, pp 357–366. https://research.google/pubs/pub46055/
29. Zhang J (2021) Toward actionable understandings of conversations: a computational approach. Cornell University. August, 2021. https://tisjune.github.io/papers/phd-thesis.pdf
30. Zhang J, Mullainathan S, Danescu-Niculescu-Mizil C (2020) Quantifying the causal effects of conversational tendencies. ArXiv 4(October). https://dl.acm.org/doi/abs/https://doi.org/10.1145/3415202