Estimating Conversational Styles in Conversational Microtask Crowdsourcing

Supplemental Material

CSCW 2020

1 Features of Style Prediction

Linguistic features extracted from the conversation, and used for machine learning model training are listed as follows:

pp : the percentage of personal pronouns, i.e. the fraction of first-person and second-person pronouns that the worker used among the total words sent to the conversational agent. Third-person pronouns are not considered in this feature as they are not relevant to the worker's personal topics and interactions between the worker and the agent. This feature refers to the *1) personal focus of topic* of the conversation.

rep : mean repeat times (frequency) of words. This feature refers to *topic cohesion* of the conversation since High-Involvement workers tend to keep reintroducing the topic. It also refers to *use of repetition* (although it is mostly absent in text-based chat).

pse : the percentage of self-editing. This feature refers to the *4) tolerance of silence* of the conversation. High-Involvement speakers like to introduce topics without hesitation, while High-Considerateness speakers can tolerate long pauses. To measure *pse*, we record all the characters that the worker has typed and then compare them with final sentences.

rtype : rate of typing, i.e. mean frequency of key pressing (per second). This feature refers to the *3) pacing* of the conversation.

pause : mean pause length, i.e. mean length of time interval between each key pressing event (unit: ms). This feature refers to the *pacing* and *tolerance of silence* of the conversation.

tt : mean length of turn-taking pause, i.e. mean length of pause between each conversation turn (we only count turn-takings from agent to worker, unit: ms). This feature also refers to the *pacing* and *4) tolerance of silence* of the conversation.

wpu : mean number of words per utterance. An utterance is the text shown in a "dialogue bubble" that worker sends to the conversational agent after the worker presses `Enter/Return` key or clicks `SEND` button. This feature refers to the *enthusiasm* of the conversation.

wps : mean number of words per sentence. A sentence is a sequence of words ending with a full stop ("."), a question mark ("?"), an exclamation mark ("!"), or the end of an utterance (pressing `Enter/Return` key or clicking `SEND` button). This feature is also about the *2) enthusiasm* of the conversation.

nint : number of interruptions. An interruption is counted if the worker starts typing before the conversational agent send the message. This feature is included because Tannen's research shows that speakers of High-Involvement style like to interrupt the conversation. It also refers to *3) pacing* of the conversation.

nque : number of question marks used by the worker. According to Tannen's conclusion of conversational styles, the linguistic device called machine-gun question is often used by High-Involvement speakers, meaning those speakers somewhat enjoy debating and usually immediately ask questions when they are in doubt. This feature is about *5) use of questions*.

nexcl : number of exclamation marks used by the worker. Speakers of High-Involvement style usually show enthusiasm during the conversation, therefore we count the number of exclamation marks as a sign of *3) enthusiasm*.

heu : the total score derived from heuristic assessment. It is calculated by $heu = h1 + h2 + h3$, representing the superimposition of answers from the first, second and third interactions respectively (h_i, $1 \leq i \leq 3$, is assigned to 1 if the Involvement answer is selected by the worker, otherwise it is assigned to -1).

sa_h : the value of feature *humor* derived from self-assessment.

sa_p : the value of feature *pace* derived from self-assessment.

sa_n : the value of feature *narrative* derived from self-assessment.

2 Binary Classifiers

Following machine learning models (binary classifiers) are used for evaluation:

Logistic Regression : Logistic regression is widely used in various fields for addressing binary classification problems.

Naive Bayes : Naive Bayes classifier is a probabilistic classifier based on Bayes' theorem. In our experiments we use a Gaussian naive Bayes classifier.

k-Nearest Neighbors : k-Nearest Neighbors (k-NN) is a simple classification algorithm. In our experiments, k value (number of neighbors) was set to 5, and all points are weighted equally.

Decision Tree : As one of the most common classifier, the decision (classification) tree is also applied in our experiments with Gini impurity as the criteria to measure the how "good" the set is split.

Gradient Boosting : Boosting algorithms are extensively used in the machine learning field. We choose Gradient Boosting as the representative of boosting algorithm family. The base estimator used in Gradient Boosting is decision tree, and the number of which is 100. The learning rate of the gradient algorithm is 1.0.

3 Evaluation Metrics

Performance of conversational style prediction is measured by comparing prediction results of classifiers with the labeled ground truth. We first categorize all the workers into four classes, as shown in Table 1, where TI ("True Involvement") and TC ("True Considerateness") represent correctly predicted workers whose predicted (also actual) styles are Involvement and Considerateness respectively, while FI ("False Involvement") and FC ("False Considerateness") mean incorrectly predicted workers whose predicted styles are Involvement and Considerateness respectively.

Table 1: Classes of workers

		Actual style	
		Involvement	Considerateness
Predicted style	Involvement	TI	FI
	Considerateness	FC	TC

To this end, measurements we used for evaluating prediction performance are listed as follows.

Overall Accuracy: the overall accuracy, noted as *accuracy*, is the fraction of correctly predicted workers among all the workers, which is calculated by:

$$accuracy = \frac{TI + TC}{TI + FI + TC + FC}. \quad (1)$$

Precision of a conversational style is the fraction of correctly predicted workers among all the workers predicted as this style. The precision of Involvement style ($precision_i$) and Considerateness style ($precision_c$) can be calculated by:

$$precision_i = \frac{TI}{TI + FI}, \tag{2}$$

$$precision_c = \frac{TC}{TC + FC}. \tag{3}$$

Recall of a conversational style is the fraction of correctly predicted workers among all the workers actually of this style. Therefore, equations for calculating the recalls of Involvement style ($recall_i$) and Considerateness style ($recall_c$) are listed as follows:

$$recall_i = \frac{TI}{TI + FC}, \tag{4}$$

$$recall_c = \frac{TC}{TC + FI}. \tag{5}$$

F1 score is the harmonic mean of the precision and recall, which is a measure of performance considering both precision and recall. The F1 scores of Involvement style ($F1_i$) and Considerateness style ($F1_c$) can be calculated by:

$$F1_i = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i}, \tag{6}$$

$$F1_c = 2 \cdot \frac{precision_c \cdot recall_c}{precision_c + recall_c}. \tag{7}$$

We can evaluate the performance of a classifier according to its F1 scores on both Involvement style and Considerateness style. A classifier is "good" only if it obtains high F1 scores on both conversational styles (both $F1_i$ and $F1_c$).

3.1 Results

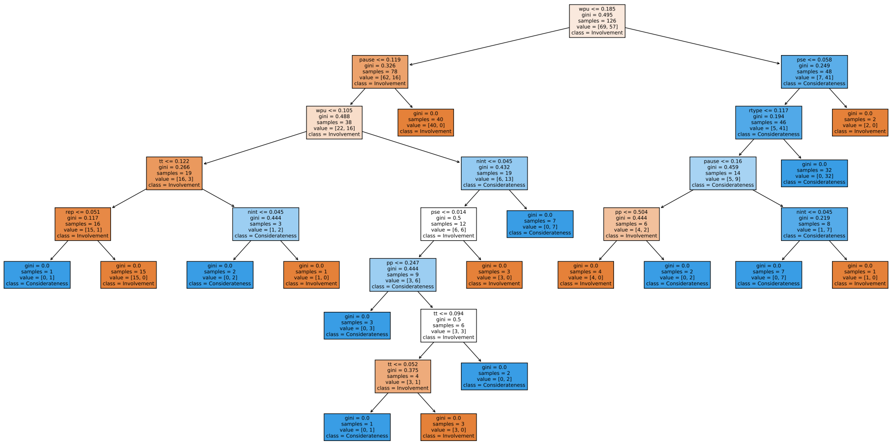
3.2 Prediction with rich data

The prediction performance is shown in Table 2. As we described before, to evaluate the overall performance of the classifier, its F1 scores on both Involvement and Considerateness styles should be high, so that we not only compare the overall accuracy, but also look at the F1 scores of both styles. Gradient Boosting can obtain the highest performance scores by using **Prediction with rich data**, whose overall accuracy and F1 scores on both styles are around 80%. The performance of Decision Tree is also high, whose overall accuracy and F1 scores on both styles are 77.78%. We show the graphic representation of a trained Decision Tree in Figure 1.

Table 2: The performance (accuracy, precision, recall and F1, unit: percent %) of conversational style prediction with rich data by common classifiers, where the classifiers with the highest *accuracy* are highlighted.

Classifiers	***accuracy***	$\boldsymbol{precision_i}$	$\boldsymbol{recall_i}$	$\boldsymbol{F1_i}$	$\boldsymbol{precision_c}$	$\boldsymbol{recall_c}$	$\boldsymbol{F1_c}$
Logistic Regression	75.93	90.00	62.07	73.47	67.65	92.00	77.97
Naive Bayes	75.93	83.33	68.97	75.47	70.00	84.00	76.36
k-NN	74.07	82.61	65.52	73.08	67.74	84.00	75.00
Decision Tree	77.78	84.00	72.41	77.78	72.41	84.00	77.78
Gradient Boosting	79.63	78.12	86.21	81.97	81.82	72.00	76.60

Figure 1: The graphic representation of Decision Tree.


wpu <= 0.185 gini = 0.495 samples = 126 value = [69, 57] class = Involvement
pause <= 0.119 gini = 0.326 samples = 78 value = [62, 16] class = Involvement
pse <= 0.058 gini = 0.249 samples = 48 value = [7, 41] class = Considerateness
wpu <= 0.105 gini = 0.488 samples = 38 value = [22, 16] class = Involvement
gini = 0.0 samples = 40 value = [40, 0] class = Involvement
rtype <= 0.117 gini = 0.194 samples = 46 value = [5, 41] class = Considerateness
gini = 0.0 samples = 2 value = [2, 0] class = Involvement
tt <= 0.122 gini = 0.266 samples = 19 value = [16, 3] class = Involvement
nint <= 0.045 gini = 0.432 samples = 19 value = [6, 13] class = Considerateness
pause <= 0.16 gini = 0.459 samples = 14 value = [5, 9] class = Considerateness
gini = 0.0 samples = 32 value = [0, 32] class = Considerateness
rep <= 0.051 gini = 0.117 samples = 16 value = [15, 1] class = Involvement
nint <= 0.045 gini = 0.444 samples = 3 value = [1, 2] class = Considerateness
pse <= 0.014 gini = 0.5 samples = 12 value = [6, 6] class = Involvement
gini = 0.0 samples = 7 value = [0, 7] class = Considerateness
pp <= 0.504 gini = 0.444 samples = 6 value = [4, 2] class = Involvement
nint <= 0.045 gini = 0.219 samples = 8 value = [1, 7] class = Considerateness
gini = 0.0 samples = 1 value = [0, 1] class = Considerateness
gini = 0.0 samples = 15 value = [15, 0] class = Involvement
gini = 0.0 samples = 2 value = [0, 2] class = Considerateness
gini = 0.0 samples = 1 value = [1, 0] class = Involvement
pp <= 0.247 gini = 0.444 samples = 9 value = [3, 6] class = Considerateness
gini = 0.0 samples = 3 value = [3, 0] class = Involvement
gini = 0.0 samples = 4 value = [4, 0] class = Involvement
gini = 0.0 samples = 2 value = [0, 2] class = Considerateness
gini = 0.0 samples = 7 value = [0, 7] class = Considerateness
gini = 0.0 samples = 1 value = [1, 0] class = Involvement
gini = 0.0 samples = 3 value = [0, 3] class = Considerateness
tt <= 0.094 gini = 0.5 samples = 6 value = [3, 3] class = Involvement
tt <= 0.052 gini = 0.375 samples = 4 value = [3, 1] class = Involvement
gini = 0.0 samples = 2 value = [0, 2] class = Considerateness
gini = 0.0 samples = 1 value = [0, 1] class = Considerateness
gini = 0.0 samples = 3 value = [3, 0] class = Involvement

3.3 Prediction with limited data

The prediction performance is shown in Table 3, where the highest accuracy value is highlighted. Because more discrete (and binary) data are used in features of limited conversational data for training, we replace the Gaussian naive Bayes classifier with Bernoulli naive Bayes classifier (for binary variables). Prediction results illustrate that the performance of **Prediction with limited data** is generally lower than **Prediction with rich data**. The overall accuracy of Gradient reaches 66.67% and its F1 scores on both styles are also the highest (around 65%) among other selected classifiers.

We believe a fine-tuned deep-learning-based model has great potential to accurately estimate the conversational style, even with limited conversational data.

Table 3: The performance (accuracy, precision, recall and F1, unit: percent %) of conversational style prediction with limited data by common classifiers, where the classifier with the highest *accuracy* is highlighted.

Classifiers	$\boldsymbol{accuracy}$	$\boldsymbol{precision_i}$	$\boldsymbol{recall_i}$	$\boldsymbol{F1_i}$	$\boldsymbol{precision_c}$	$\boldsymbol{recall_c}$	$\boldsymbol{F1_c}$
Logistic Regression	57.41	62.50	51.72	56.60	53.33	64.00	58.18
Naive Bayes	55.56	61.90	44.83	52.00	51.52	68.00	58.62
k-NN	48.15	51.35	65.52	57.58	41.18	28.00	33.33
Decision Tree	59.26	60.00	72.41	65.62	57.89	44.00	50.00
Gradient Boosting	66.67	70.37	65.52	67.86	62.96	68.00	65.38