Does BERT Learn as Humans Perceive?

Understanding Linguistic Styles through Lexica



Shirley Anugrah Hayati
Dongyeop Kang
Lyle Ungar

Penn University of Pennsylvania
University of Minnesota
Georgia Tech

Motivation

I will understand if you decline, but would very much like you to accept. May I nominate you?

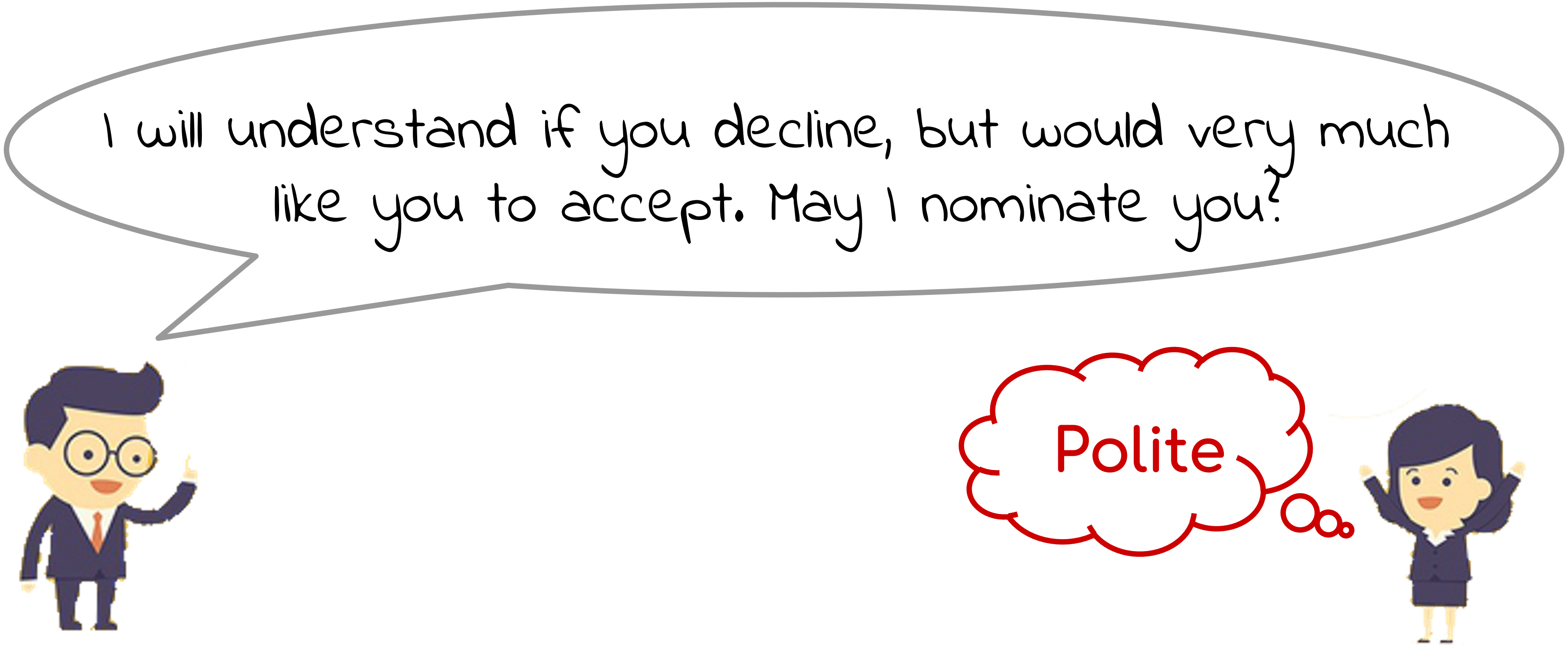
Polite ✅

Offensive ❌

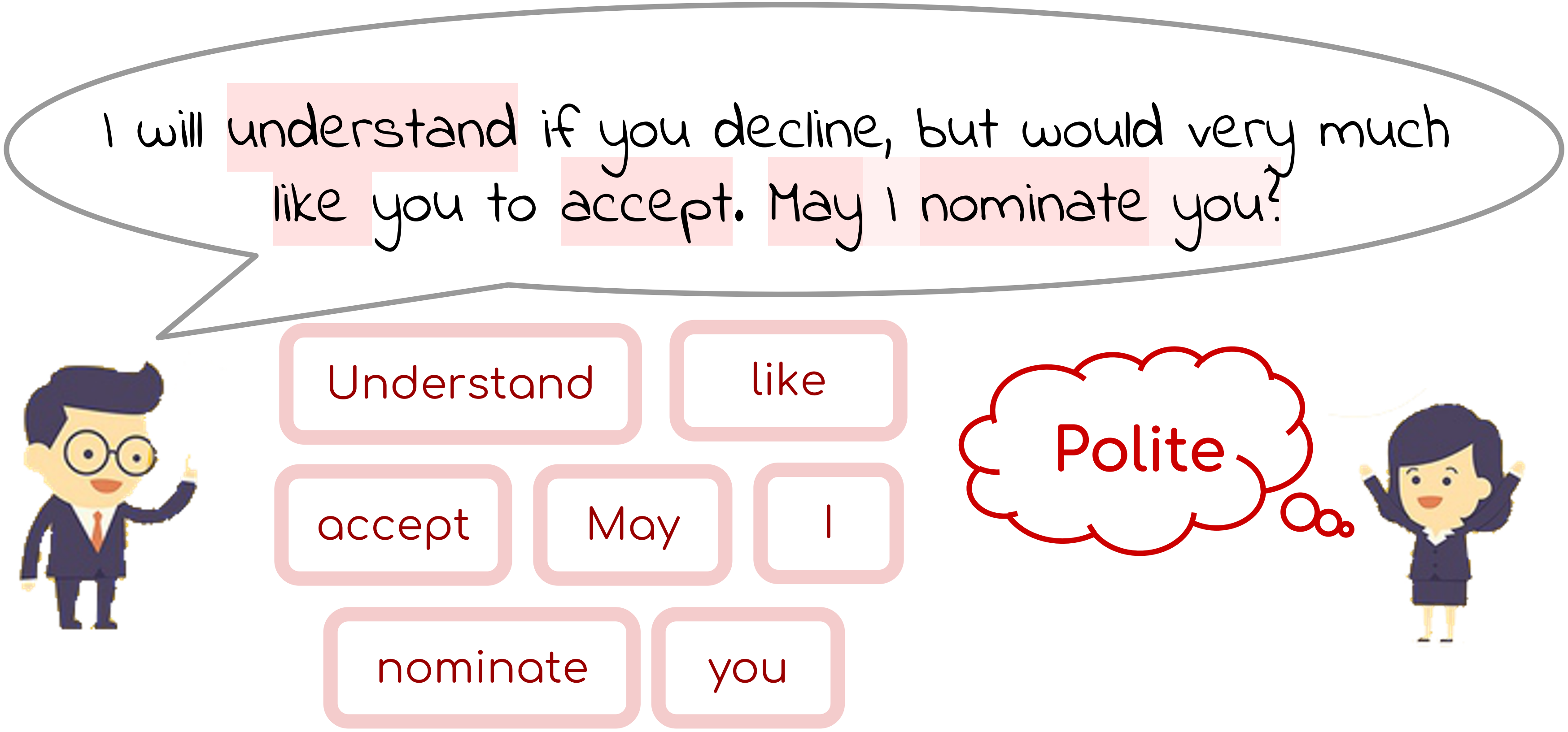
Positive ✅

Joyful ✅

Motivation

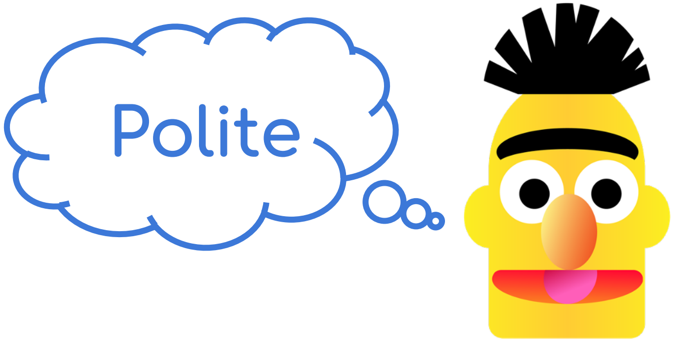

I will understand if you decline, but would very much like you to accept. May I nominate you?
Polite

Human vs. BERT in Styles

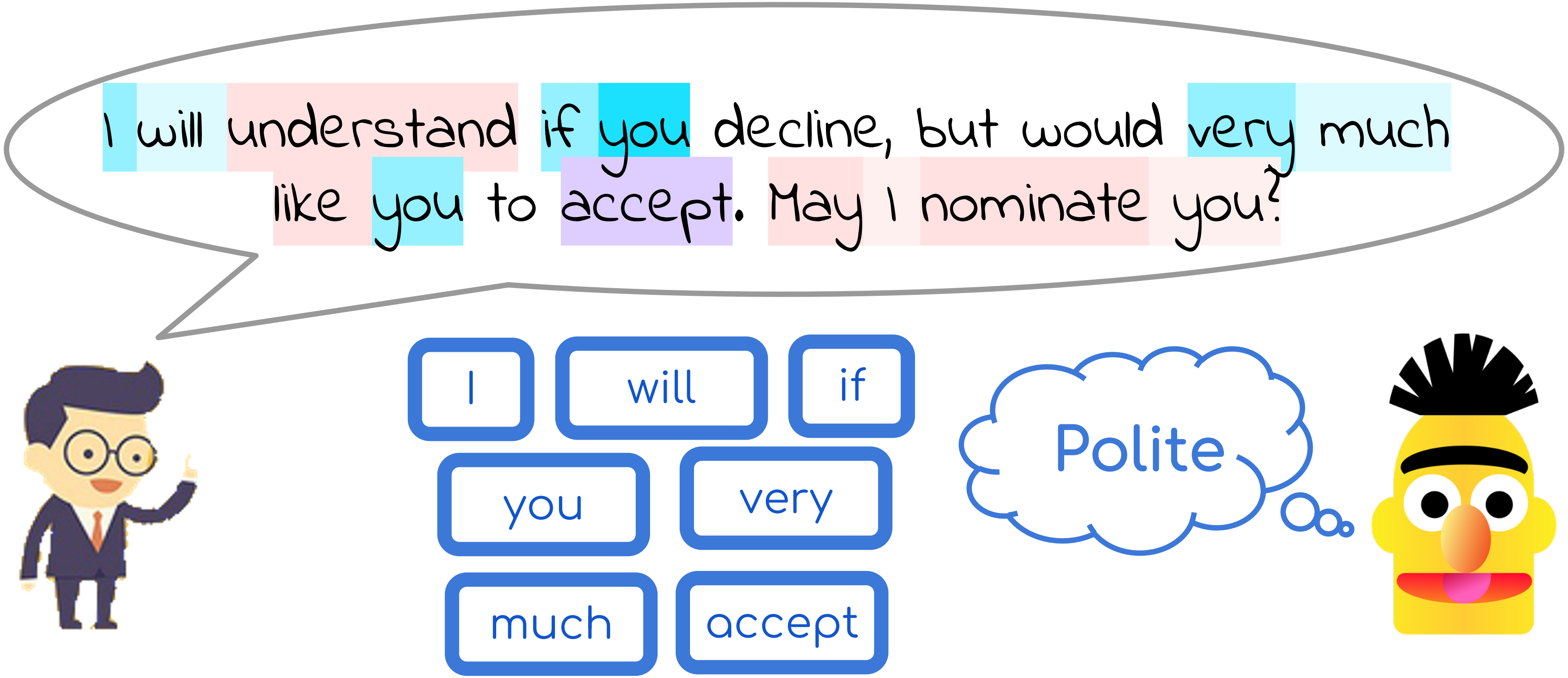


Human vs. BERT in Styles

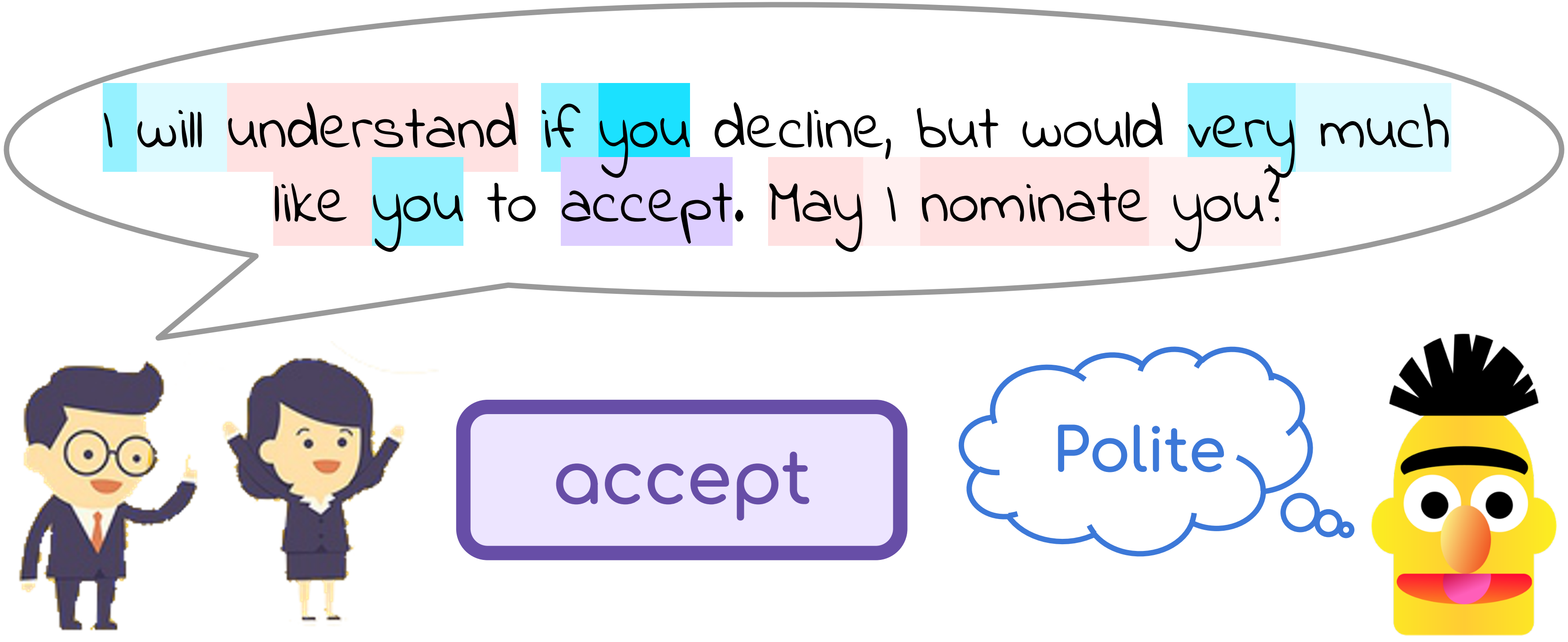
I will understand if you decline, but would very much like you to accept. May I nominate you?



Human vs. BERT in Styles

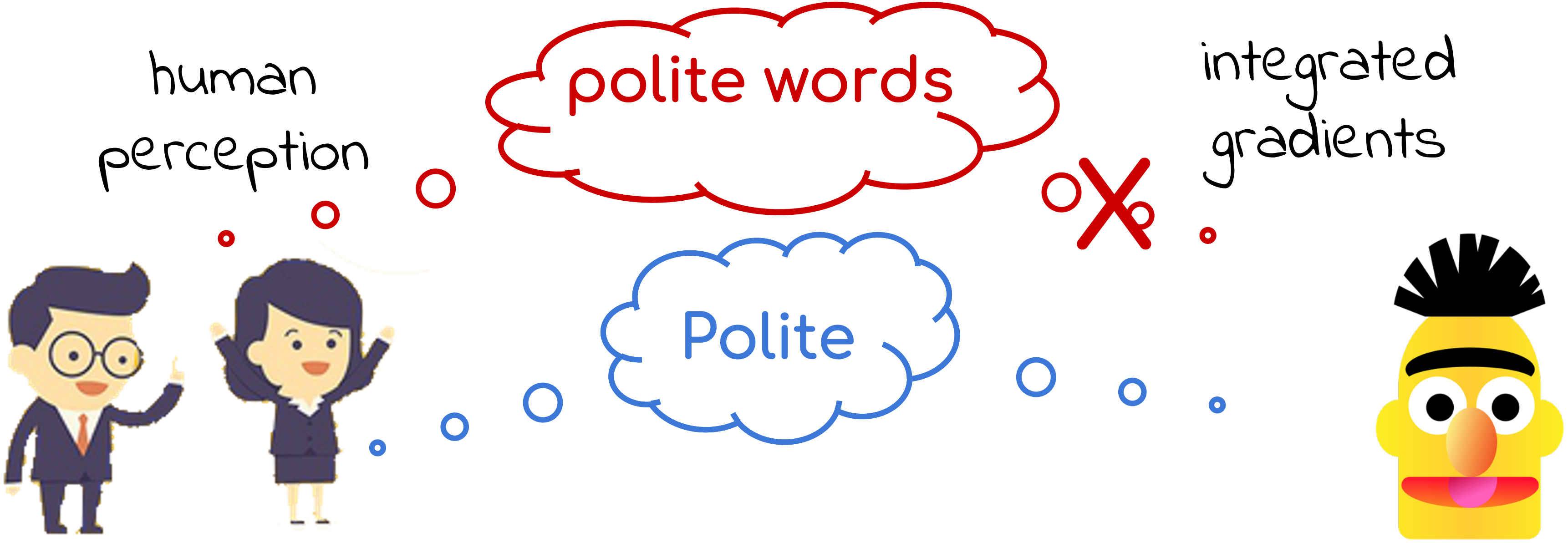


Human vs. BERT in Styles



Human vs. BERT in Styles

Words that BERT thinks as important != humans perceive



Main Question

To what extent does BERT's word importance align with human perception?

8 Linguistic Style Datasets (Kang and Hovy, 2021)

Politeness (Danescu-Niculescu- Mizil et al., 2013)

Polite | Impolite

Sentiment Treebank (Socher et al., 2013)

Positive | Negative

Hate and Offensive Tweets (Davidson et al., 2017)

Offensive | Not Offensive

SemEval 2018: Affect in Tweets (Mohammad et al., 2018)

Anger | Disgust | Fear

Joy | Sadness

Hummingbird Dataset Collection

Existing Style Dataset

BERT

500 Stylistically-diverse texts

ranked by avg and std of probability scores

Style	F1 (%)
Politeness	69.4
Sentiment	96.5
Offensiveness	98.0
Anger	82.0
Joy	86.5

*please refer to the paper for the full result

Hummingbird Dataset Collection

500 Stylistically-diverse texts

Crowd workers

Hummingbird

Average
Inter-annotator agreement:
Sent: 73.2%
Word: 27.7%

Human Perception Score

$$H(w_i) = \frac{\sum_{j=1}^{\#\text{annotators}} h_j(w_i)}{\#\text{annotators}}$$

$h_j \in \{-1, 0, 1\}$ given by j^{th} annotator

#annotator = 3

BERT's Word Importance:

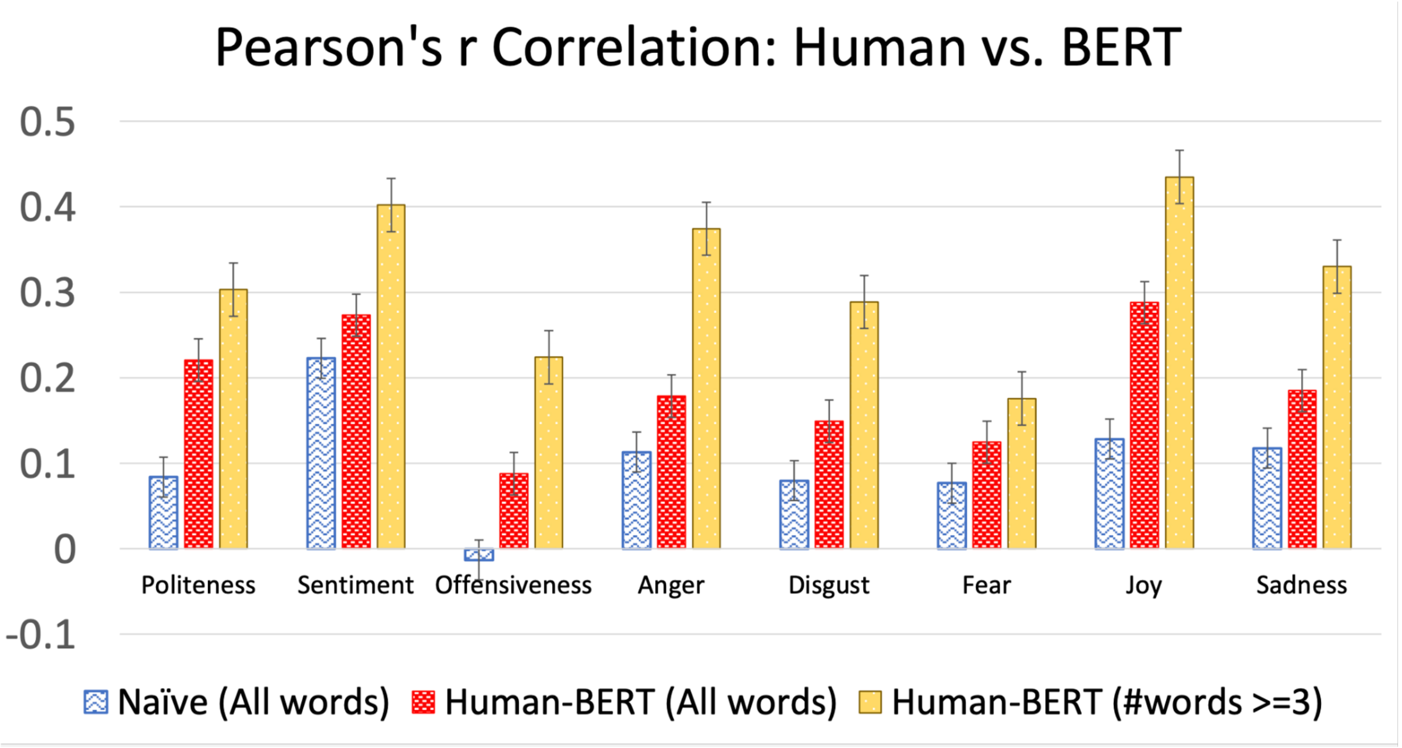
Integrated Gradients

(Sundaranjan et al., 2017; Mudrakarta et al., 2018)

$$\mathrm{IG}_i(x, x') ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

x = input word piece
x' = baseline input
dF/dx = the gradient of neural network F
IG (x, x') ∈ [-1, 1]

Intra-Stylistic Analyses



Intra-Stylistic Analyses

Joy		
🙋🤖	🙋	🤖
excited	moved	movies
love	share	managing
entertaining	performances	referring
great	congrats	documentary
perfect	smile	baseball

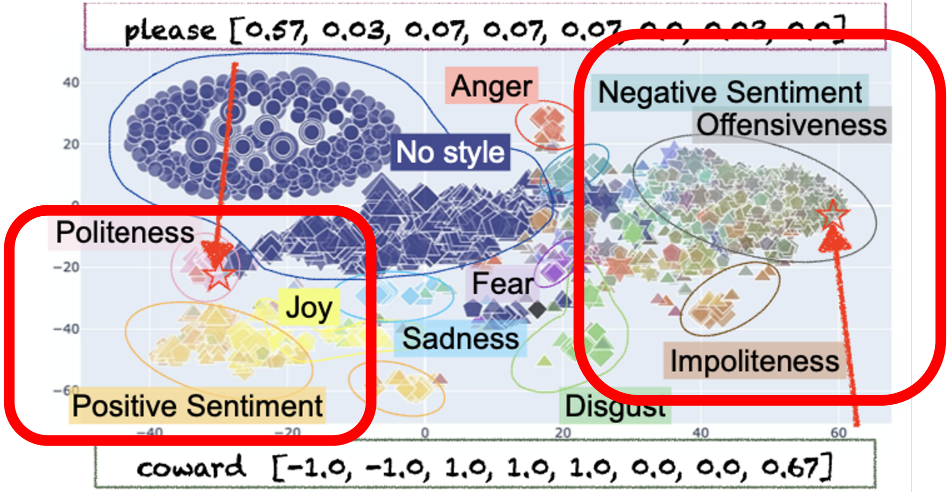
Multi-stylistic Analyses

human

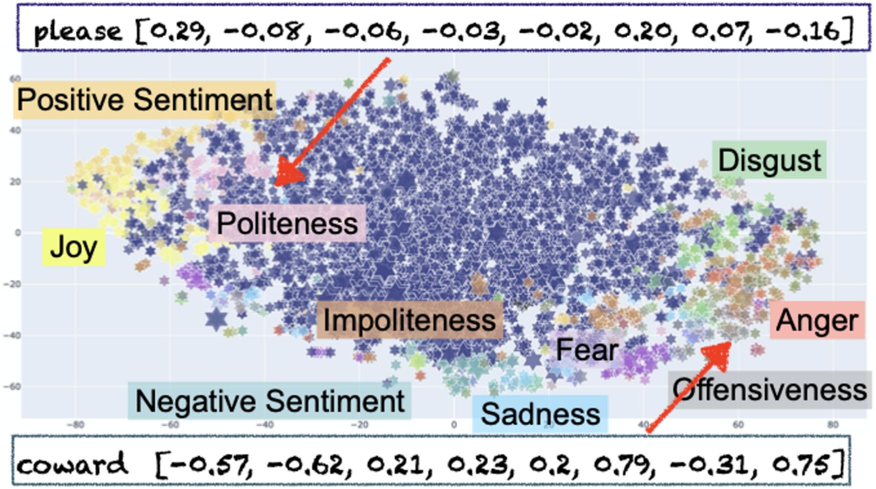
	politeness	sentiment	offensive	anger	disgust	fear	joy	sadness
politeness	1	0.58	-0.49	-0.46	-0.49	-0.17	0.35	-0.18
sentiment	0.17	1	-0.41	-0.53	-0.55	-0.36	0.45	-0.43
offensive	-0.099	-0.15	1	0.46	0.48	0.21	-0.079	0.2
anger	-0.27	-0.19	0.19	1	0.56	0.28	-0.1	0.31
disgust	-0.22	-0.25	0.21	0.46	1	0.32	-0.098	0.35
fear	-0.068	-0.12	-0.041	0.11	0.079	1	-0.072	0.36
joy	0.23	0.24	-0.12	-0.32	-0.37	-0.17	1	-0.083
sadness	-0.0097	-0.19	-0.01	0.14	0.16	0.21	-0.17	1

machine

Multi-stylistic Analyses



Human



Machine

Takeaways

1. Word-importances tend to be noisy for rare words
2. BERT takes more context;
humans intuitively choose the most obvious "stylistic" words
3. Styles are subjective, so humans may have different perception towards them

Future Work

1. Scaling up the data size for more styles
2. Informing BERT with human perceptions for explaining styles and generalizability

Thank you! 😄

https://github.com/sweetpeach/hummingbird/