Does BERT Learn as Humans Perceive: Understanding Linguistic Styles Through Lexica

Shirley Anugrah Hayati — shirley@gatech.edu
Dongyeop Kang — dongyeop@umn.edu
Lyle Ungar — ungar@cis.upenn.edu

BACKGROUND

(a) **Human:** *Polite* **BERT:** *Polite*

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

(b) **Human:** *Anger* **BERT:** *Not Anger*

a nightmare date with a half-formed wit done a great disservice by a lack of critical distance and a sad trust in liberal arts college bumper sticker platitudes .

Human BERT Both

OUR PROBLEM

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

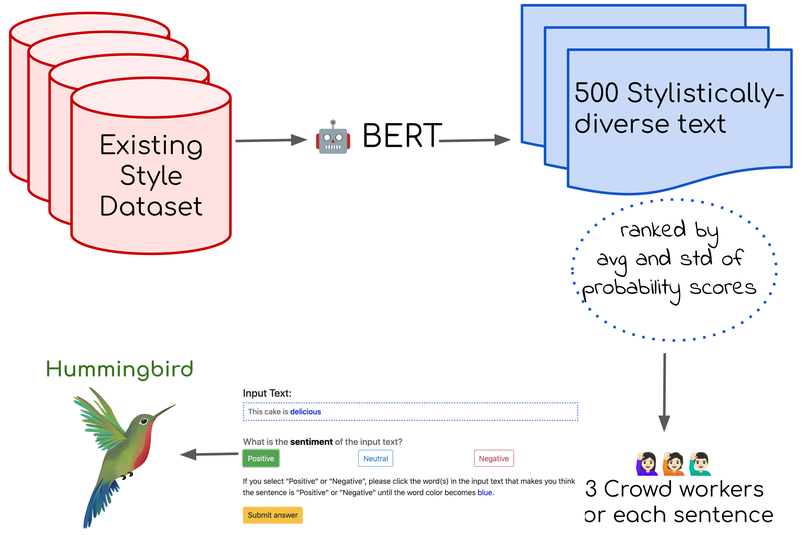
OUR CONTRIBUTIONS

- The first comparative study to examine stylistic lexical cues from human perception and BERT.
- A new dataset, called **Hummingbird**, where crowd-workers relabeled bench marking datasets for eight style classification tasks: politeness, sentiment, offensiveness, 5 emotions
- BERT pays more attention to content words

FUTURE WORK

- Scaling up the dataset
- Explaining styles with human perception
- Building a more generalized model

HUMMINGBIRD DATA COLLECTION



STATISTICS

Style	Label Distribution	Inter-annotator agreement	F1 (%)
Politeness	22.8% polite 41.2% impolite	62.8	69.4
Sentiment	24.6% positive 54.6% negative	71.1	96.5
Offensiveness	33.6%	75.7	98.0
Anger	35.0%	73.5	82.0
Disgust	41.6%	71.2	80.7
Fear	16.4%	76.1	84.6
Joy	22.6%	82.7	86.5
Sadness	26.4%	72.4	78.2

METHODS

Human Perception Scores

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

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

#annotator = 3

Integrated Gradients

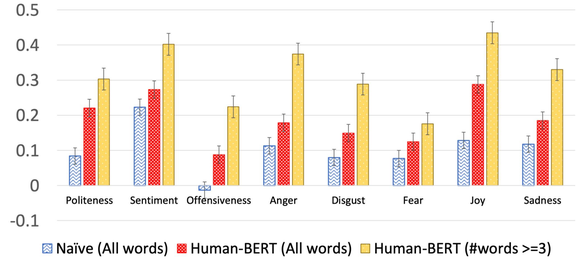
(Sundararajan et al., 2017; Mudrakarta et al., 2018)

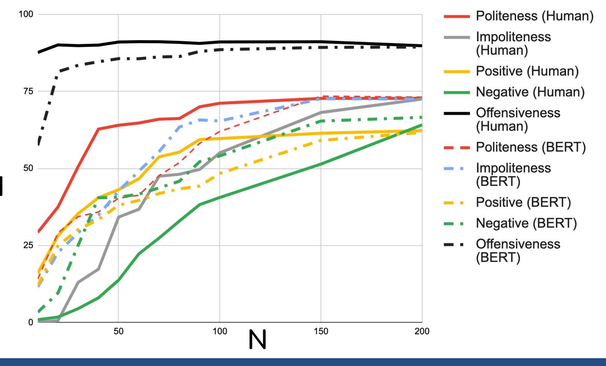
$$\text{IG}_i(x, x') ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

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

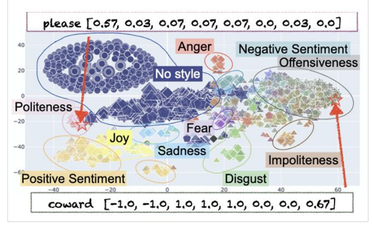
INTRA-STYLISTIC ANALYSES

Pearson's r Correlation: Human vs. BERT

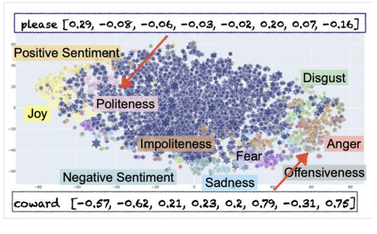




MULTI-STYLISTIC ANALYSES



Human



Machine