StyLEx : Explaining Style Using Human Lexical Annotations

Shirley Anugrah Hayati
hayat023@umn.edu

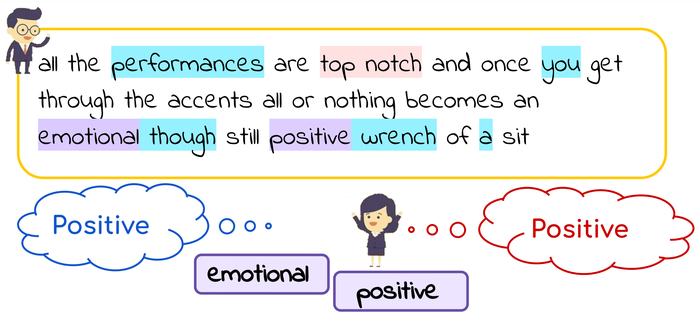
Kyumin Park
pkm9403@kaist.ac.kr

Dheeraj Rajagopal
dheeraj@cs.cmu.edu

Lyle Ungar
ungar@cis.upenn.edu

Dongyeop Kang
dongyeop@umn.edu

BACKGROUND



OUR PROBLEM

How can we incorporate **human perceptions** for improving **model explanation** on generating **stylistic lexica**?

DATASET

HUMMINGBIRD (Hayati et al., 2021)

✅ Lexical Annotation
Source: Original (ORIG)
#Instances: 500 for each style

Original (ORIG)

❌ Lexical Annotation

Sources:
- **Politeness:** Wikipedia, StackExchange
- **Sentiment:** Movie review
- **Offensiveness:** Twitter
- **5 Emotions:** Twitter

#Instances: 6.8k - 238k

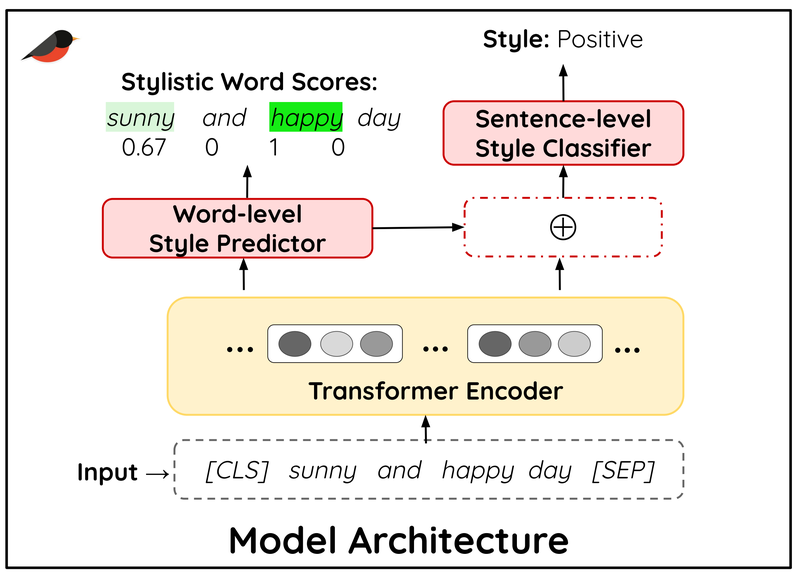
Out-of-Domain (OOD)

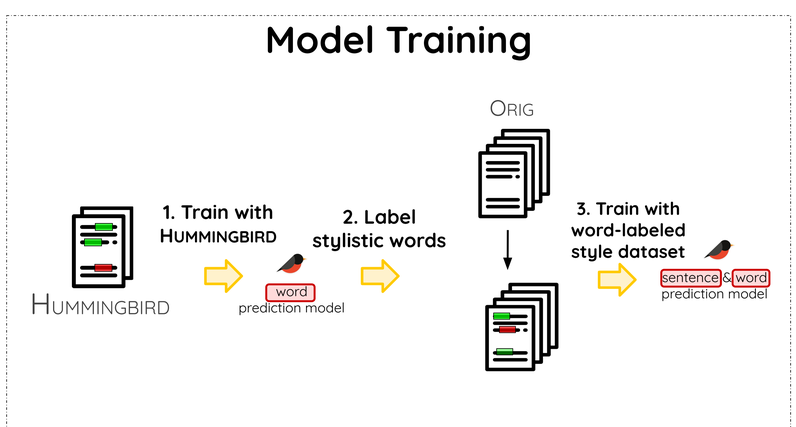
❌ Lexical Annotation

Sources:
- **Politeness:** Corporate email
- **Sentiment**: Product review
- **Offensiveness**: Twitter
- **5 Emotions**: Reddit posts

#Instances: 1k - 16k

STYLEX

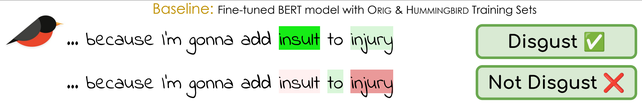




EXPERIMENT

Style	Original		OOD	
	Baseline	**StyLEx**	**Baseline**	**StyLEx**
Politeness	67.96%	65.84%	71.45%	74.18%
Sentiment	96.52%	96.59%	85.45%	86.18%
Offensiveness	97.75%	97.81%	88.62%	88.98%
Anger	89.04%	89.01%	77.49	77.51%
Disgust	86.50%	86.90%	74.06%	74.63%
Fear	95.66%	95.63%	78.42%	78.48%
Joy	88.02%	88.12%	75.20%	74.26%
Sadness	88.38%	88.41%	78.37%	78.71%

Baseline: Fine-tuned BERT model with ORIG & HUMMINGBIRD Training Sets



EXPLANATION EVALUATION