StylEx: Explaining Style Using Human Lexical Annotations

Shirley Anugrah Hayati Kyumin Park Dheeraj Rajagopal
Lyle Ungar Dongyeop Kang

University of Minnesota
KAIST
Carnegie Mellon University
Penn University of Pennsylvania

1

What is StyLEx?

Motivation (Hayati et al., 2021)

all the performances are top notch and once you get through the accents all or nothing becomes an emotional though still positive wrench of a sit

Motivation

all the performances are top notch and once you get through the accents all or nothing becomes an emotional though still positive wrench of a sit

Motivation

all the performances are top notch and once you get through the accents all or nothing becomes an emotional though still positive wrench of a sit

Motivation

all the performances are top notch and once you get through the accents all or nothing becomes an emotional though still positive wrench of a sit

Positive

Motivation

all the performances are top notch and once you get through the accents all or nothing becomes an emotional though still positive wrench of a sit

Motivation

all the performances are top notch and once you get through the accents all or nothing becomes an emotional though still positive wrench of a sit

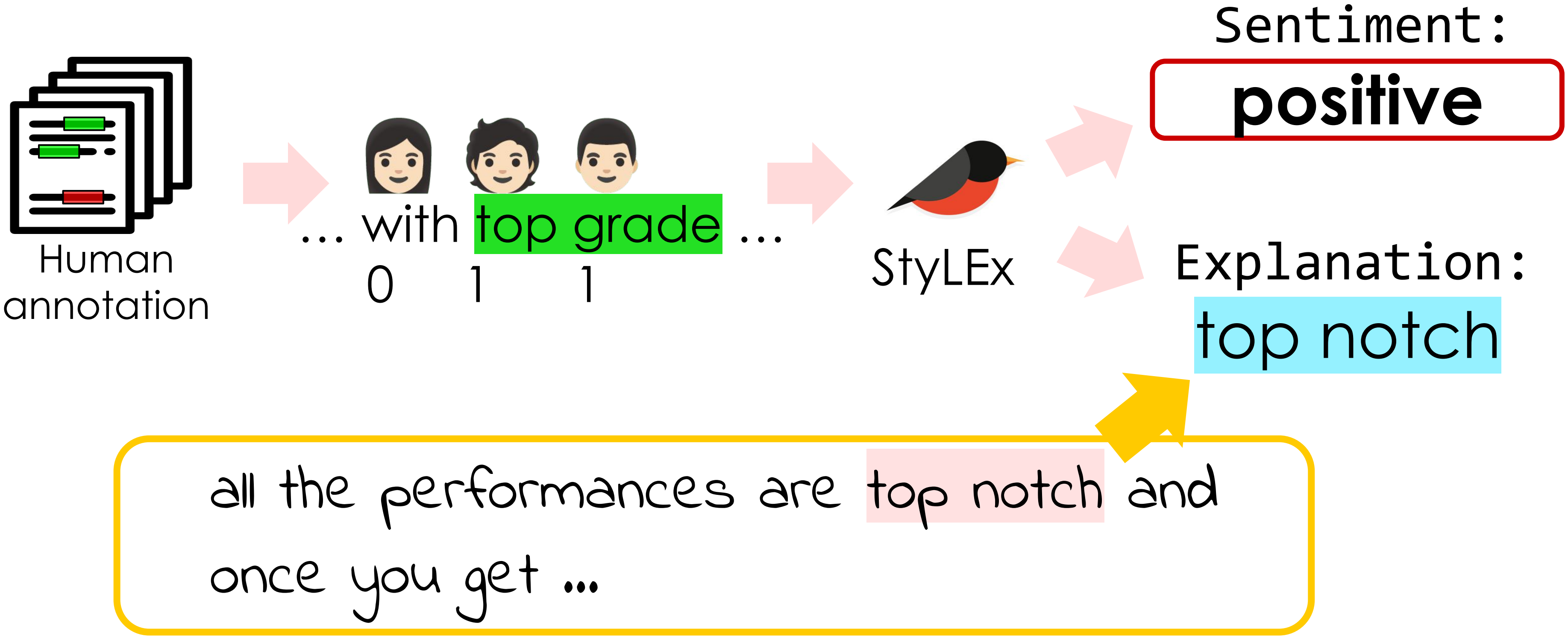
emotional

positive

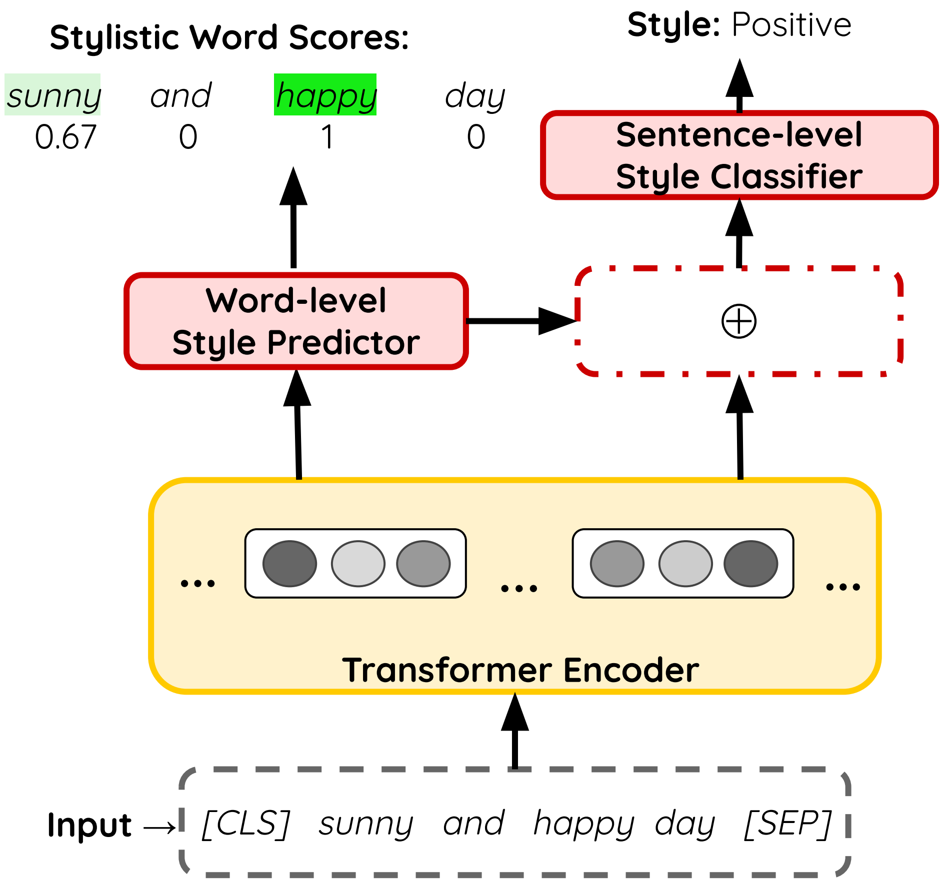
top notch

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

StyLEx


Sentiment:
positive
Human annotation
… with top grade …
0 1 1
StyLEx
Explanation:
top notch
all the performances are top notch and once you get …

StyLEx



$$\mathcal{L} = \mathcal{L}_{style} + \alpha \times \mathcal{L}_{word}$$

8 Linguistic Styles (Hayati et al., 2021)

Polite | Impolite

Positive | Negative

Offensive | Not Offensive

Anger | Disgust | Fear

Joy | Sadness

Datasets

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
- **Emotions:** Twitter

#Instances: 6.8k - 238k

Out-of-Domain (OoD)

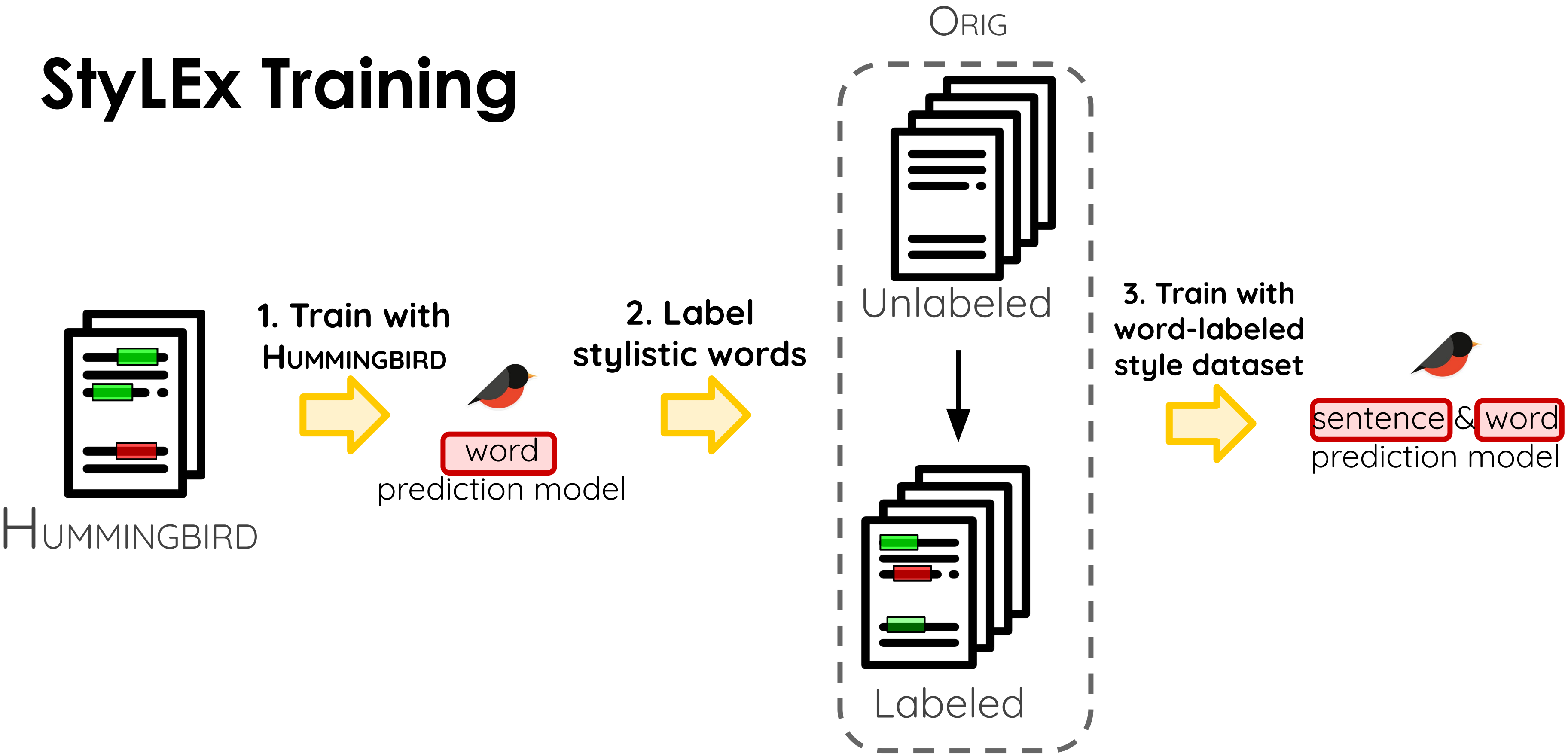
❌ Lexical Annotation

Sources:

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

#Instances: 1k - 16k

StyLEx Training



2

Experiment & Discussion

Experiment Setup

1 **Generalizability** → Model performance (F1-score)

Baseline: Fine-tuned BERT models (ORIG + HUMMINGBIRD)

2 **Explainability** →
- Sufficiency
- Plausibility
- Understandability

Baseline: Integrated Gradient (Sundaranjan et al., 2017; Mudrakarta et al., 2018; and used by Hayati et al., 2021)

Style Classification Results

Style	Original		OOD	
	Baseline	**StyLEx**	**Baseline**	**StyLEx**
Politeness	<u>67.96%</u>	65.84%	71.45%	<u>74.18%</u>
Sentiment	96.52%	<u>96.59%</u>	85.45%	<u>86.18%</u>
Offensiveness	97.75%	<u>97.81%</u>	88.62%	<u>88.98%</u>
Disgust	86.50%	<u>86.90%</u>	74.06%	<u>74.63%</u>
Sadness	88.38%	<u>88.41%</u>	78.37%	<u>78.71%</u>

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

* Please refer to the paper for full result

Example

... because I'm gonna add insult to injury

Disgust ✅

... because I'm gonna add insult to injury

Not Disgust ❌

... please put them all back are you on dsl

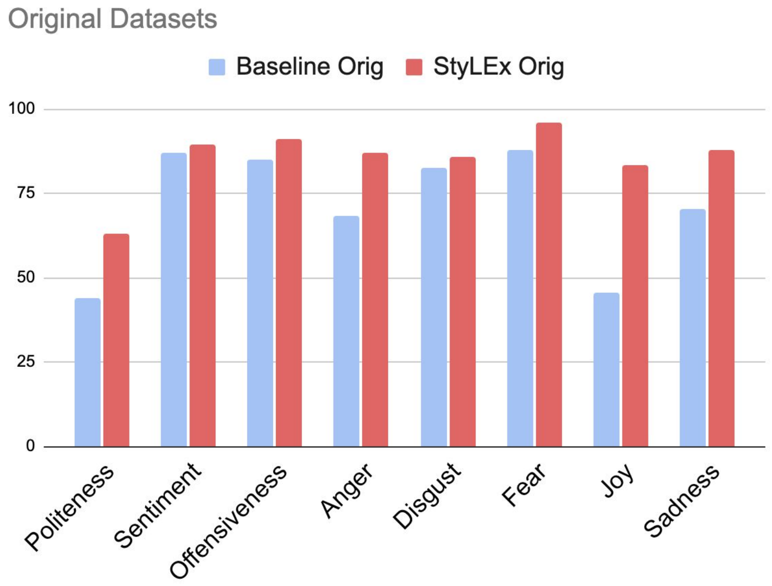
Polite ❌

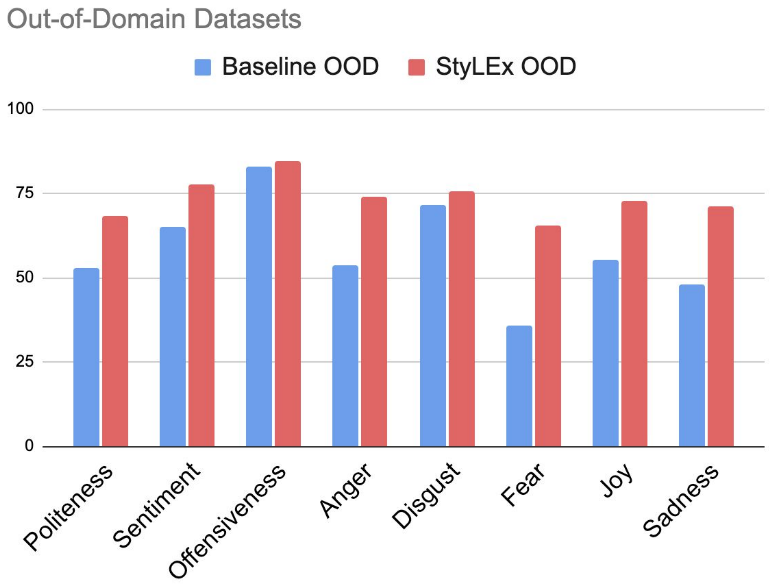
... please put them all back are you on dsl

Impolite ✅

Sufficiency

F1 scores for fine-tuning BERT with top-*k* important words



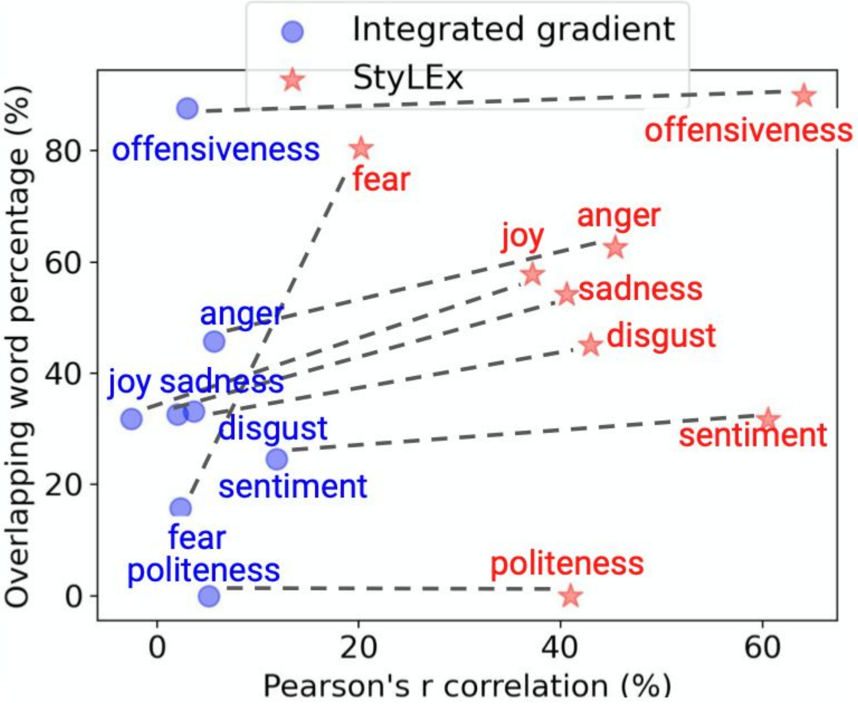


Baseline: Integrated Gradient (Sundaranjan et al., 2017; Mudrakarta et al., 2018)

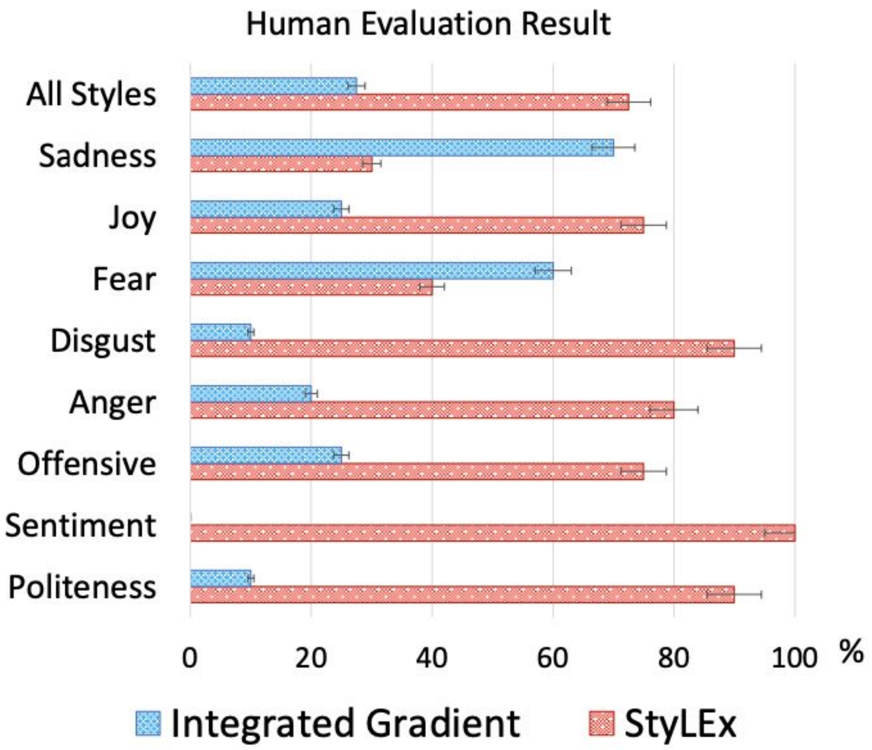
Plausibility

1. Correlation with **human perception**

2. Comparison with **stylistic lexicon dictionary**



Understandability



Takeaways

1. StyLEx provides explanation and doesn't hurt performance
2. StyLEx's explanations are sufficient for model prediction and more preferred by humans
3. StyLEx is more generalized than the baseline (Out-of-Domain results)

Limitations and Future Work

1. Increasing the dataset size and including more styles, e.g., formality, humor, etc., and phrase-level explanation.

2. Capturing subtle stylistic words and handling sparsity in stylistic words.

3. Applying to style-content disentanglement for stylistic text generation.

Thank you! 😊

https://github.com/minnesotanlp/stylex