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 University of Pennsylvania
hayat023@umn.edu pkm9403@kaist.ac.kr dheeraj@cs.cmu.edu
ungar@cis.upenn.edu dongyeop@umn.edu

Abstract

Large pre-trained language models have achieved impressive results on various style classification tasks, but they often learn spurious domain-specific words to make predictions (Hayati et al., 2021). While human explanation highlights stylistic tokens as important features for this task, we observe that model explanations often do not align with them. To tackle this issue, we introduce **StyLEx**, a model that learns from human annotated explanations of stylistic features and jointly learns to perform the task and predict these features as model explanations. Our experiments show that StyLEx can provide human-like stylistic lexical explanations without sacrificing the performance of sentence-level style prediction on both in-domain and out-of-domain datasets. Explanations from StyLEx show significant improvements in explanation metrics (sufficiency, plausibility) and when evaluated with human annotations. They are also more understandable by human judges compared to the widely-used saliency-based explanation baseline.[1]

1 Introduction

People use style as a strategic choice for their personal or social goals in communications, making style analysis a long-studied field in NLP (Hovy, 1987; Kabbara and Cheung, 2016; Kang and Hovy, 2021). While large language models have obtained state-of-the-art results on many NLP tasks, they have been shown to overfit to spurious correlations in data across several datasets (Sen et al., 2021; Schlangen, 2021; Bras et al., 2020). Hayati et al. (2021) found a phenomenon in style classification tasks where the model's word-level explanation do not align with human's stylistic cues (stylistic cues are words that signify the style of a text). For instance, words such as "performances" and "wrench" in Figure 1 are marked as important cues for sentiment by a saliency method. However, they are different from words that humans perceive as essential features for predicting the style ("top," "notch").

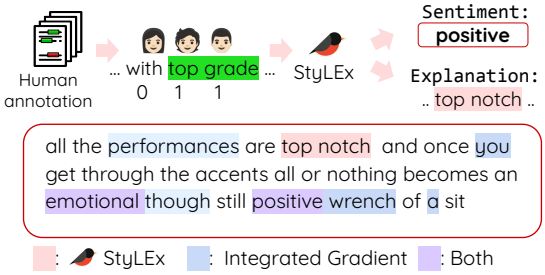

Figure 1: StyLEx classifies the input sentence's style and provides lexical explanation. Compared to explanations computed by the integrated gradient method (Mudrakarta et al., 2018), StyLEx can find more accurate stylistic words. Green highlight refers to human's annotated positive word, pink for StyLEx, blue for baseline, and purple for both StyLEx and the baseline.

Prior research in style have developed stylistic lexicon dictionary to identify the style of a text, such as sentiment or emotion, and in-turn incorporated them for style classification tasks (Mohammad and Turney, 2010; Hutto and Gilbert, 2014; Tausczik and Pennebaker, 2010). While lexicon-based matching methods (Taboada et al., 2011; Eisenstein, 2017) provide interpretability for the task, they lack coverage and do not incorporate the context for prediction. On the other hand, current large scale models like BERT (Devlin et al., 2019) are effective at style classification. However, their explanations often reveal that the model do not rely on the stylistic words to make the prediction. In this work, we hypothesize that leveraging stylistic lexica along with the effectiveness of a large scale model like BERT can not only predict style but also provide meaningful explanations that align with human style explanations.

Towards this, we introduce **StyLEx**, a style classification model that jointly learns to align human-annotated stylistic cues as explanations and then

* currently at Google
[1] Code and data are publicly available at https://github.com/minnesotanlp/stylex

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2843–2856

predict the style of the overall sentence based on the cues (Figure 1). StyLEx uses a semi-supervised approach to expand the stylistic words from a handful number of human-annotated stylistics words. First, we train StyLEx using existing small human stylistic word annotation from Hayati et al. (2021). Then, we obtain predicted stylistic words on the larger benchmarking style datasets and retrain StyLEx on this expanded data to predict both the sentence's style label and the stylistic lexical explanation.

In this study, we show that for both in-domain and out-of-domain data, StyLEx not only shows competitive classification performance with BERT-based model, but also generate stylistic lexical explanations that have higher alignment with human explanations. In terms of explanation quality, StyLEx surpasses the baseline method across multiple explanation metrics. For the sufficiency metric, we improve upon the baseline by 14.12% on the average. For plausibility, StyLEx's lexical explanations correlate highly with human lexical annotation ground truth with an average Pearson's r correlation score of 44% compared to the baseline's correlation score of 3.9%. Finally, we found that 72.5% of StyLEx's explanations are more preferred by human judges compared to the baseline.

2 StyLEx: Style Classification with Human Lexical Annotations

2.1 StyLEx Model Architecture

StyLEx is a joint model for word-level and sentence-level style prediction. Unlike a multi-task learning approach where tasks are independent of each other, StyLEx exploits these stylistic word scores obtained from human annotation and then helps predict the sentence's styles. As displayed in Figure 2 (left), StyLEx involves three modules: a transformer-based (Vaswani et al., 2017; Devlin et al., 2019) encoder, a word-level style predictor and a sentence-level style predictor. This work is based on BERT although the encoder can be applied to any transformer architecture.

Given an input of token sequence $\mathbf{x} = \{x_1, ..., x_n\}$ and its corresponding set of stylistic word scores $\{s_1, ..., s_n\}$, we encode $\mathbf{x}$ using a pre-trained transformer model. We extract the final layer output as $\mathbf{h} = \{\mathbf{h}_1, ..., \mathbf{h}_n\}$ and feed $\mathbf{h}$ to the word-level style prediction layer which is a neural classifier that outputs stylistic word logits for each word l_{word_i} computed as follows:

$$l_{word_i} = \mathbf{W}_{word}\mathbf{h}_i + \mathbf{b}_{word}$$

where $i \in \{1, .., n\}$, $\mathbf{W}_{word}$ is a matrix with the size $H \times d_{l_{word}}$, and $\mathbf{b}_{word} \in \mathbb{R}^{d_{l_{word}}}$ is the bias term. H is the size of the default hidden layer in BERT which is 768 and $d_{l_{word}}$ denotes the number of classes of each style (e.g., positive or negative word in a sentiment classification task).

For the sentence-level style classification, we first take both the encoded representation $\mathbf{h}$ and stylistic word logits l_{word}. We then apply max pooling on the aggregation of $\mathbf{h} \oplus l_{word}$ along the sequence, resulting in vector $\mathbf{v} \in \mathbb{R}^{H+d_{l_{word}}}$ consisting of important logits. Finally, we input $\mathbf{v}$ into the sentence-level style classifier defined as follows:

$$l_{sentence} = \texttt{softmax}(\mathbf{W}_{sentence}\mathbf{v} + \mathbf{b}_{sentence})$$
$$P_{sentence} = \arg\max(l_{sentence})$$

where $l_{sentence} \in \mathbb{R}^2$ denotes sentence-level style logits, $\mathbf{W}_{sentence}$ is a matrix with the size $(H + d_{l_{word}}) \times 2$, and $P_{sentence}$ is the index of the predicted sentence-level style.

During training, StyLEx's objective is to maximize the probability of the sentence's style and stylistic word scores. The loss for both sentence-level style predictor and word-level style predictor is computed using binary cross entropy loss function. To jointly train the model, we optimize the following loss:

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

where α is a regularization hyperparameter.

2.2 StyLEx Model Training

To train StyLEx, we need a dataset of stylistic sentences along with their corresponding stylistic words (§2.2.1). We use the HUMMINGBIRD dataset (Hayati et al., 2021) that contains 500 sentences with word-level style annotation for obtaining the stylistic lexical explanation. Due to HUMMINGBIRD's small size, we first train StyLEx on HUMMINGBIRD and then predict pseudo stylistic words on larger benchmark style datasets (> 6.8k sentences in the training sets) for training the final StyLEx model for both sentence classification and lexical explanation.

2.2.1 Datasets

Following Hayati et al. (2021), we explore the same set of eight styles used in the dataset: politeness, sentiment, offensiveness and five emotions (anger, disgust, fear, joy, and sadness) for style classification tasks. We use three sets of publicly available style datasets for our experiments as follows.

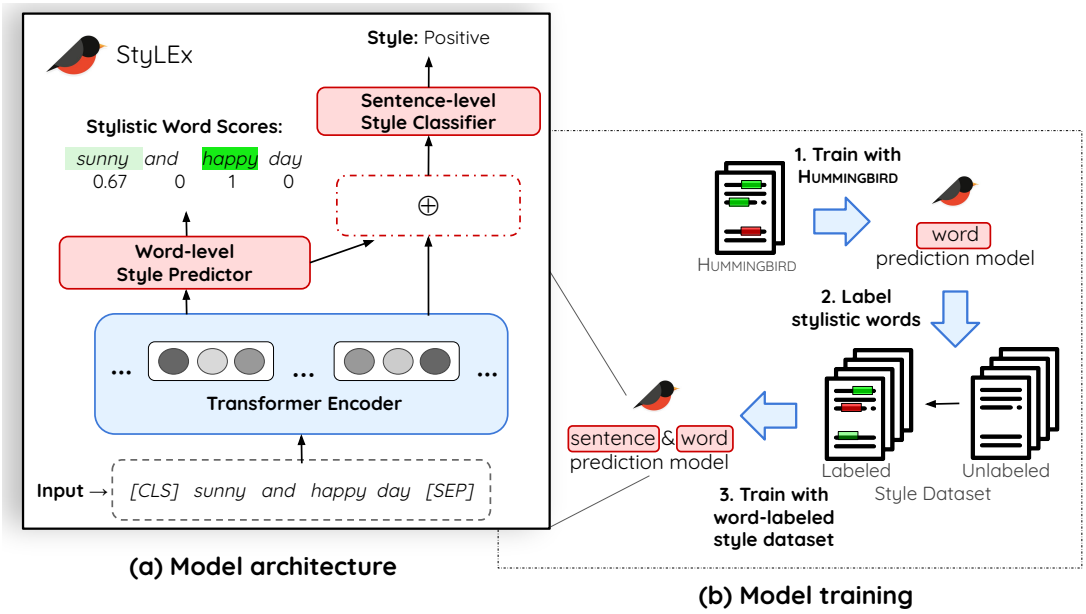

Figure 2: **(a) StyLEx model architecture** (left). Our model has two new modules: a word-level style predictor and a sentence-level style classifier. An aggregator appends the word-level style logit for each word to the hidden layer representations of each word and takes the max pooling of this aggregation. **(b) Model training** (right). Human labels come from HUMMINGBIRD (Hayati et al., 2021) for stylistic word scores and from ORIGINAL datasets from sentence-level style classification. (1) We train a stylistic word-level prediction model on HUMMINGBIRD dataset in order to (2) obtain *pseudo-stylistic words* of sentences in the ORIGINAL datasets. (3) Then we train another stylistic word and sentence prediction model on this ORIGINAL sentences, now labeled with stylistic words.

HUMMINGBIRD is a multi-style dataset annotated with human perception scores on its important stylistic lexicons (Hayati et al., 2021). HUMMINGBIRD contains 500 sentences based on eight style datasets: politeness, sentiment, offensiveness, and five emotions (anger, disgust, fear, joy, and sadness). Three different crowd workers annotate each word in a sentence with 1 if they perceive the word as stylistic and 0 if not. The human perception score for a word is the average score of these annotators' labels. This perception score is what we call as stylistic word score and it is within the range [-1, 1]. We use HUMMINGBIRD for training StyLEx's word-level style predictor.

ORIGINAL datasets are used by Hayati et al. (2021) to curate HUMMINGBIRD.[2] Since some style labels in ORIGINAL may contain continuous numbers rather than binary labels, we follow the same setting of Hayati et al. (2021) which only uses binary labels: polite or impolite, positive or negative, offensive or not offensive, anger or not anger, and so on. The politeness dataset comes from StackExchange and Wikipedia requests (Danescu et al., 2013) (9.8k training instances). The sentiment dataset is a collection of movie review texts (Socher et al., 2013) (117k training instances). The offensiveness dataset is from Twitter (Davidson et al., 2017) (20k training instances). The emotions dataset (Mohammad et al., 2018) is collected from tweets (6.8k training instances). For all these ORIGINAL datasets, we use the default train/dev/split as explained in their papers.

Out-of-Domain (OOD) datasets are used to evaluate StyLEx's performance on different domains. For each style, we use data from different sources or topics, but their style labels are in HUMMINGBIRD and ORIGINAL datasets. For politeness, we use the polite and impolite sentences from the Enron email corpus (Klimt and Yang, 2004; Madaan et al., 2020). For sentiment, we test StyLEx on 5-core reviews from Amazon review dataset (Ni et al., 2019) for each product categories, except for movie reviews. We exclude movie reviews because it would be similar to the domain of the ORIGINAL's sentiment dataset. We convert ratings of 4-5 to positive labels and ratings of 1-2 as negative labels. For offensiveness, we use OffensEval (Zampieri et al., 2019) dataset for offensiveness. For five emotions, we collect Reddit comments

[2]We will refer to these individual datasets as "ORIGINAL"

Style	ORIGINAL (%)		OOD (%)	
	BERT	StyLEx	BERT	StyLEx
Politeness	67.96	65.84	71.45	74.18
Sentiment	96.52	96.59	86.70	86.99
Offensiveness	97.75	97.81	88.62	89.00
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

Table 1: For the sentence-level style classification task, StyLEx does not sacrifice the task performance (F1-scores) of the **BERT** model across all of the style tasks across both **ORIGINAL** and **OOD** settings.

from GoEmotions corpus (Demszky et al., 2020).[3]

2.2.2 Training

The whole pipeline of StyLEx model training is in Figure 2 (right). First, we train a stylistic word score prediction model with the same StyLEx architecture in Figure 2 (left). We do this since the sentences in the benchmarking style datasets do not have human annotations of stylistic word scores. We then use a semi-supervised learning approach called, pseudo-labeling (Lee et al., 2013; Rizve et al., 2020), to label the stylistic words. Now the sentences in ORIGINAL contain stylistic word scores which are output by the stylistic word predictor. Finally, we use both HUMMINGBIRD and ORIGINAL for training another model of StyLEx which predicts sentence-level binary style labels (polite and impolite, positive and negative etc.,) and provides lexical explanation scores within the range of [0, 1].[4]

3 Evaluation on Style Classification

3.1 Baseline

To assess StyLEx's classification performance, we compare it with a fine-tuned BERT-based classifier as a baseline. The training data for the baseline is also a combination of HUMMINGBIRD and ORIGINAL. For explanation evaluation, we compare StyLEx's explanation with the commonly-used explanation method called integrated gradients (Mudrakarta et al., 2018; Sundararajan et al., 2017), implemented in Captum[5]. Integrated gradient, which can be viewed as an approximate method of estimating Shapley values, is defined as follows. For the input sequence of words x and a neural network function F, an attribution score (the explanation) for each word is defined as the gradient between the input x and baseline x' of the function F where x' is a zero scalar.

3.2 Results

In our experiment, we have eight StyLEx models for each style: politeness, sentiment, offensiveness, and five emotions. For each style, we run StyLEx on the ORIGINAL test sets and OOD datasets for five times with different seeds and report the average of F1-scores in Table 1. For ORIGINAL datasets, StyLEx achieves higher F1 scores compared to the fine-tuned BERT model on sentiment, offensiveness, disgust, joy, and sadness. Overall, we observe that StyLEx does not sacrifice task performance of the state-of-the-art classifiers while predicting stylistic word scores. When tested on the OOD test sets, StyLEx achieves higher F1 score against the fine-tuned BERT model for all styles. Politeness has the greatest improvement from 71.48% to 74.18% since we observe that the ORIGINAL dataset of politeness contains many spurious content words. When we use bigger language models such as RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2020) for StyLEx, StyLEx still has better results than the baseline for five styles: sentiment, offensiveness, anger, and disgust.

From example sentences from the test sets in Table 2, we can see how StyLEx helps task performance. ORIG-1 and ORIG-2 sentences show how StyLEx can capture stylistic words and correct the sentence's style label to *disgust*. For example, "insult" and "injury" in ORIG-1 are initially labeled by the integrated gradient method as unimportant for identifying *disgust*, but StyLEx identifies the words as stylistic cues. Similarly, for the word "downfall" in ORIG-2, StyLEx finds it as offensive, but the baseline does not. StyLEx also has a higher stylistic word score for indicating "stogie" as an offensive word.

As we look at the politeness classification results, we find that most of the incorrect cases are when StyLEx mislabels subtle *impolite* sentences as *polite*. As we observe in Table 2 ORIG-3, StyLEx finds the word "please" as a polite cue, but the ground truth label of the sentence is *impolite*. We

[3] More details on the datasets are in Appendix A.1.
[4] Other implementation details are in Appendix A.2.
[5] https://captum.ai

ID	Model	Sentence with Predicted Stylistic Word Scores	Sentence Style
Incorrect Baseline Prediction → Correct StyLEx Prediction			
ORIG-1	StyLEx	... because i'm gonna ' add insult to injury	Disgust ✓
	Baseline	... because i'm gonna ' add insult to injury	Not Disgust ✗
ORIG-2	StyLEx	... yet you say i'm a stogie you're your own downfall	Offensive ✓
	Baseline	... yet you say i'm a stogie you're your own downfall	Not Offensive ✗
Correct Baseline Prediction → Incorrect StyLEx Prediction			
ORIG-3	StyLEx	please put them all back are you on dsl	Polite ✗
	Baseline	please put them all back are you on dsl	Impolite ✓
ORIG-4	StyLEx	... can't just be mean and do horrid things busy without paying the price	Not Fear ✗
	Baseline	... can't just be mean and do horrid things busy without paying the price	Fear ✓

Table 2: Error analysis on prediction-flipped sentences. Baseline refers to sentence-level style prediction result by fine-tuned BERT model and highlights are stylistic words found by the integrated gradient method. Green highlights on the words mean that the model predicts high positive word-level stylistic scores; red for the opposite label (e.g., *impolite* or *negative*). Sentence Style is a model's sentence-level style prediction ✓ marks correct prediction and ✗ denotes incorrect prediction.

then inspect its continuous score from the original politeness dataset by Danescu et al. (2013). It turns out that its politeness score is -0.38 in the range of [-2, 2] as -2 being the most impolite sentence. This shows that this sentence score is closer to neutral than impolite. This finding also reflects how HUMMINGBIRD dataset has been collected: as mentioned in Hayati et al. (2021), words from the offensive dataset (mostly swear words) are often labeled as impolite by human annotators. Thus, it may bias the annotators' view that sentences with impolite labels are not as bad as offensive sentences, making them not mark offensive sentences as *impolite*. Therefore, for such subtle impolite sentences, human annotators in HUMMINGBIRD may not label the sentence and words as impolite.

In contrast, StyLEx misclassifies *anger* sentences as *not anger* and *fear* as *not fear*. As we look at ORIG-4 in Table 2, StyLEx weakly finds *fear* cues ("horrid", "things") but they do not help in boosting the model to predict the sentence as *fear*. We conjecture that this is because there are very few training samples labeled with *fear* and *fear* has quite low word-level inter-annotator agreement as reported in Hayati et al. (2021).

4 Style Explanation Evaluation

We investigate StyLEx's explanations if they are sufficient, plausible, and understandable following previous works (DeYoung et al., 2020; Jacovi and Goldberg, 2020; Wiegreffe and Marasovic, 2021; Rajagopal et al., 2021). Jacovi and Goldberg (2020) define that a faithful interpretation represents a model's reasoning process. To evaluate whether StyLEx's explanations are faithful, we run a sufficiency test that evaluates whether the model explanations alone are highly relevant for predicting the label (Jacovi et al., 2018). Meanwhile, we measure plausibility to examine whether the explanation is agreeable to humans (DeYoung et al., 2020). Finally, understandability measures if a user is able to understand model explanations (Rajagopal et al., 2021). For investigating sufficiency and plausibility, we run automatic metrics. To assess understandability, we ask human judges to choose the explanation that can be better understood by a non-expert between StyLEx and the integrated gradients (IG) method.

4.1 Sufficiency

Following Jain et al. (2020); Rajagopal et al. (2021)'s sufficiency test, we fine-tune a BERT model on the top-k words as explanation instead of the whole sentence. We limit an explanation to contain 30% words of the average sentence length for each of the style datasets. These words are ranked based on their importance score by the baseline integrated gradient method and StyLEx for all the positive stylistic words (polite words, positive words, offensive words, angry words).

In Table 3, we can see that explanations from StyLEx show much higher predictive performance

Style	ORIGINAL		OOD	
	IG	StyLEx	IG	StyLEx
Politeness	43.92	**63.08**	52.89	**8.19**
Sentiment	87.18	**89.39**	64.93	**77.52**
Offensiveness	84.87	**91.26**	82.93	**84.75**
Anger	68.36	**86.90**	53.76	**73.99**
Disgust	82.54	**85.91**	71.36	**75.76**
Fear	87.82	**96.10**	35.85	**65.26**
Joy	45.54	**83.16**	55.49	**72.71**
Sadness	70.49	**87.94**	47.83	**70.95**

Table 3: The results for sufficiency test on the **ORIGINAL** and **OOD** data show the F1 scores on top-k words. IG stands for integrated gradient.

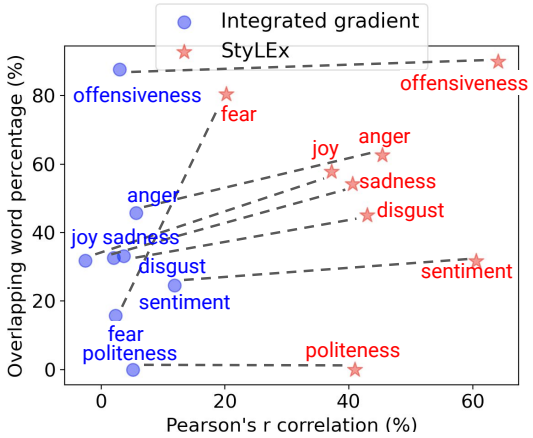

Figure 3: Plausibility experiment result. There are two points for each style in this plot. A blue circle point is for the baseline IG method and a red star point for StyLEx. X-axis is Pearson's r correlation score for each style. Y-axis is the percentage of stylistic sentences with style words appearing in the existing style lexicon dictionary.

compared to explanations extracted by the integrated gradient method for all styles in both datasets, ORIGINAL and OOD. This result suggests that human-like stylistic words are much more strongly predictive of a sentence's style compared to the gradient-based explanation methods that often rely on content words as an explanation. This indicates that StyLEx's explanations are relatively more faithful compared to the integrated gradients based method.

4.2 Plausibility

We use two approaches to measure the agreement between StyLEx's lexical explanations and stylistic words perceived by humans to assess the plausibility of StyLEx's explanations. In the first approach, we compare StyLEx's stylistic words on the HUMMINGBIRD test set and compare it with the ground truth human perception scores in HUMMINGBIRD. Second, we compare StyLEx's top-k stylistic words with existing expert-curated stylistic lexicon dictionaries. The domain of the existing expert-curated stylistic lexicon dictionaries could be different from the domain of the datasets in our study which range from social media texts to Wikipedia. However, it is still useful to compare how much StyLEx and the baseline agree human experts on identifying stylistic words.

Figure 3 shows a scatterplot of StyLEx vs. the integrated gradient. X-axis represents the Pearson's r correlation score on the HUMMINGBIRD test set. Y-axis is the percentage of overlapping words between the important words found by StyLEx and the baseline compared with the human-curated style lexicon dictionary. We calculate the overlapping word percentage as follows. We compute how many of the top 30% of the stylistic words in the ORIGINAL datasets found by StyLEx or baseline appear in human-curated dictionaries for the emotion/sentiment/offensive lexicons.

In Figure 3, the higher the Pearson's correlation score is (to the right), the better the explanation words produced by the model (StyLEx or baseline) are aligned with human perception ground truth from HUMMINGBIRD. The dashed lines show how much StyLEx's generated stylistic words align more with human annotations for stylistic words from both HUMMINGBIRD and human-curated stylistic lexicon dictionaries.

(1) Correlation with human perception. We investigate how similar StyLEx's explanations are with human perceptions. To do so, we compute the Pearson's correlation r between stylistic word scores predicted by StyLEx and annotated by humans from HUMMINGBIRD annotations for each word by concatenating all the predicted stylistic word scores. Our correlation score is different from Hayati et al. (2021) because we split the Hummingbird dataset into a training set and test set even though we use the same human perception scores and the integrated gradient method. The correlation score we reported in Figure 3 is from the HUMMINGBIRD test set.

In Figure 3 (vertical trend), we can see that StyLEx explanations correlate more with ground truth human perception for all styles, as red stars are stretched to the right. Sentiment and offen-

Style	Both	StyLEx	Integrated gradient
Positive	good, fun, love	associate, develop, instruct	deserve, endure, football
Negative	bad, horror, silly	mess, chaos, disappoint	maternal, banger, yell
Offensive	bitch, bitches, pussy	blind, racist, panties	fairy, amateur, fisting
Anger	angry, anger, awful	frowning, scare, lose	belt, campaigning, destroying*
Disgust	awful, terrible, angry	dismal, frowning, animosity	congress, finally, sentence*
Fear	fear, anxiety, nervous	horrid, war, threaten	rejects, mum, beating
Joy	happy, love, good	faith, sing, succeed	deal, independence, football
Sadness	depression, sadness, lost	bad, offended, leave	funeral, bloody, case*

Table 4: Three important words found by StyLEx ( ) and the integrated gradient method ( ) that appear in the stylistic lexicon dictionary.* = words only appear one time in the test data.

siveness are styles that have the highest correlation scores (60.53% and 64.09%) while fear is the lowest (20.17%)). Explanations from integrated gradient correlate very loosely with human perception ground truth with sentiment as the highest (11.89%) and joy negatively correlates with human perceptions (-2.55%).

(2) Comparison with stylistic lexicon dictionaries. We then investigate how similar the stylistic words found by StyLEx are to the stylistic words curated by humans in the existing lexicon dictionary. We use sentiment emotion lexicons from Mohammad and Turney (2010) and offensive lexicons from von Ahn's research group (2021).[6] Using the same set up of sufficiency test, we select top-30% stylistic words from each sentence in ORIGINAL datasets with the positive style label. Then we check if at least one of these words appear in the existing lexicon dictionary and compute its average across all training samples.

In Figure 3 (horizontal trend), we see that StyLEx consistently has higher percentage of word occurrences in the lexicon dictionary compared to the integrated gradient method where fear has the highest percentage difference (15.78% → 80.43%) and offensiveness has the lowest percentage change (87.67% → 89.99%). Averaging across all styles, we find that 56.70% of the stylistic sentences with StyLEx stylistic words appear in the existing style lexicon dictionary while integrated gradient only identifies 37.01% of those words.

The score is higher for offensiveness than for sentiment or emotion. We observe that people use more offensive words in social media, which is the source for dataset collection. We also examine the lower occurrence for emotions. From our analysis, we found that the emotion lexicon dictionary contains several colloquially rare words "aberration" or "meritorious", leading to a very low overlap with the datasets that we used for the analysis.

[6]We couldn't find a publicly available politeness lexicon dictionary.

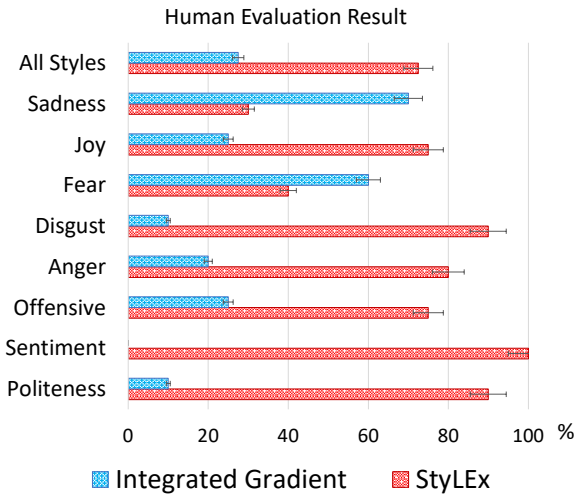

Figure 4: Human evaluation results. X-axis is the percentage of explanations preferred by human judges.

We also take a closer look at how many and the nature of important words are captured by StyLEx and/or the integrated gradient method as shown in Table 4. These word scores are obtained by averaging their scores and then we sort them based on these average scores. In general, we find that StyLEx can find more diverse stylistic words as defined in the existing lexicon dictionary for all styles except for positive sentiment. Some emotion words found by the integrated gradients only appear rarely in the data (mostly only once).

4.3 Understandability

To investigate the quality of StyLEx's explanation, we ask human judges to evaluate StyLEx's expla-

nations compared to baseline explanations. *Understandability* asks whether human judges understand our explanation better than the explanation computed using integrated gradients. For this study, we randomly select 20 stylistic sentences for each of the eight styles, resulting in total 160 sentences. These 20 sentences are constructed by 10 sentences from ORIGINAL test set and 10 sentences from OOD test set. We normalize the stylistic word scores for sentence length across all sentences.

A human judge is shown two versions of the same sentence with different anonymized highlights, as shown in Figure 1. We then ask three different human judges to select (through Amazon Mechanical Turk) the explanation that was more understandable. Each worker annotated 20 sentences of the same style. The order of explanations is randomized to remove bias. We say that an explanation by a method is preferred by human judges when the majority choose that method. If the majority chooses a method for all the 20 sentences, the X-axis score will be 100%. Results in Figure 4 show that across all styles (160 pairs of explanations), StyLEx gives an overall gain of 27.5%.

5 Related Work

Styles in NLP Research on style in NLP has addressed various tasks including style classification (Danescu et al., 2013; Socher et al., 2013), style transfer (Rao and Tetreault, 2018; Li et al., 2018), style and content disentanglement (John et al., 2018; Zhu et al., 2021), and multiple style analysis (Hayati et al., 2021; Kang and Hovy, 2021). This work focuses on understanding stylistic variation in style classification. Style classification models often produce spurious features (Sen et al., 2021; Schlangen, 2021; Bras et al., 2020), motivating us to leverage stylistic variation from human perspectives to distinguish between stylistic words and content words. Past work have used stylistic lexica for classification (Taboada et al., 2011; Eisenstein, 2017), but in this work, we fine-tune the language model to generate these lexica and use them to help style prediction. Our work is most closely related to Hayati et al. (2021), but they do not develop any new models to use the human perception scores as explanations. Moreover, while linguistics styles can cover an author's writing style or figurative language, we limit our study to high-level style as used in Kang and Hovy (2021); Hayati et al. (2021).

Explainable NLP Heat maps generated from attention values from the models (Bahdanau et al., 2014) are widely used as an interpretability tool, but these attention maps are often unfaithful and unreliable (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Zhong et al., 2019; Pruthi et al., 2020). Saliency maps computed via gradients offer an alternative (Sundararajan et al., 2017; Smilkov et al., 2017; Mudrakarta et al., 2018). Annotating explanations as *rationales* (part of input) (Lei et al., 2016) through expert annotations (Zaidan and Eisner, 2008) is widely used to model explanations in NLP when external annotations are available. Another class of inherently interpretable models aims to optimize model explanations without any external annotations (Card et al., 2019; Croce et al., 2019; Rajagopal et al., 2021). Our work is similar in spirit to the rationale approaches (Lei et al., 2016) but focuses on understanding style attributes in text and computationally modeling them based on human annotation of the important words.

6 Conclusion

We proposed StyLEx, a style classification model for learning stylistic variations through lexical explanation. With only 500 sentences with word-level style annotation, we find improvement in both classification and explanation. Compared to the commonly-used integrated gradient method, StyLEx's explanations are more accurate for model prediction, more consistent with human-found stylistic words from existing datasets and lexicon dictionaries, and better understandable by human judges, without sacrificing task performance on both in-domain and out-of-domain datasets.

Future Work Our approach opens up future work on human-centered lexical explanation for correcting the spurious behavior of NLP models and for better explaining linguistic styles. We plan to investigate collecting more human lexical annotations to more accurately model stylistic variation, especially with larger pretrained language models. Broader usage of StyLEx in providing stylistic cues will be applicable to lexical style and content disentanglement (Cheng et al., 2020; John et al., 2019), counterfactual data augmentation for style-related tasks (Sen et al., 2021), and stylistic paraphrasing (Pavlick and Nenkova, 2015).

Limitations Our work has some limitations, mostly stemming from the size and nature of the

human-annotated data. The training data (500 sentences) from HUMMINGBIRD is quite small to train deep learning models. However, our work shows that with just 500 sentences we could achieve a huge improvement in interpretability as well as a slight improvement in OOD performance, using semi-supervised training. Moreover, there is sparsity of stylistic words in the sentences. We also found that some stylistic words have various scores of human perception; capturing such subtle stylistic words is difficult. An interesting future work would be to handle these problems of sparsity and subtlety. We also notice that HUMMINGBIRD is annotated by people residing in the United States. Thus, their perception of styles may not reflect the perception of those with different cultural backgrounds. Nevertheless, StyLEx can be applied to any dataset training with similar human lexical annotations, and not limited to HUMMINGBIRD.

Ethical Considerations When collecting the explanation evaluation from human judges, we inform them that the content may contain offensive languages that could be upsetting.

Acknowledgments

We would like to thank Karin de Langis for her valuable feedback during the early version of writing and the anonymous reviewers for their thoughtful comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Dallas Card, Michael Zhang, and Noah A Smith. 2019. Deep weighted averaging classifiers. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 369–378.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541.

Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4028–4037.

Niculescu-Mizil Cristian Danescu, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jacob Eisenstein. 2017. Unsupervised learning for lexicon-based classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic transfer in natural language generation systems using recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47.

Dongyeop Kang and Eduard Hovy. 2021. Style is NOT a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387, Online. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. In *ACL*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. *arXiv preprint arXiv:2103.12279*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2020. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. How does counterfactually augmented data impact models for social computing constructs? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Luis von Ahn's research group (2021). Offensive/profane word list. Accessed: 2021-11-14.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from

annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*, pages 31–40.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2021. Neural stylistic response generation with disentangled latent variables. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4391–4401.

A Appendix

A.1 Sampling OOD Data and Data Statistics

- Politeness: We randomly sample 500 polite sentences and 500 impolite sentences from the Enron email corpus (Klimt and Yang, 2004; Madaan et al., 2020) since the size of entire corpus (>600k) is too large for inference.
- Sentiment: We test StyLEx on 5-core reviews from Amazon review dataset (Ni et al., 2019). For each category, we sample 100 positive sentences and 100 negative sentences from review categories, except for movie reviews which would be similar to the domain of the ORIGINAL dataset. We convert ratings of 4-5 to positive labels and ratings of 1-2 as negative labels.
- Offensiveness: We use OffensEval (Zampieri et al., 2019) dataset for offensiveness. We select all offensive tweets (3,002 instances) and all non-offensive tweets (2,991 instances) since OffensEval dataset is already nearly balanced.
- Emotions: For five emotions, we collect samples from GoEmotions corpus (Demszky et al., 2020) that contains Reddit comments labeled with 27 emotions, but we only select the five relevant emotions. For each emotion, we use all data for the positive emotion (e.g., joy) and undersample the negative emotion (e.g., not joy) data to equal the number of positive emotion samples.

The three dataset statistics are summarized in Table 6.

A.2 StyLEx Implementation Details

Throughout the experiment, we set $d_{l_{word}} = 2$ for politeness (polite, impolite) and sentiment (positive, negative) which have two style classes and $d_{l_{word}} = 1$ for the other styles. At the loss calculation step, we set the regularization hyperparameter α to 0.05 which gives the best style and perception prediction found searching the range [0.01, 100]. For the pseudo-labeling approach, we use the same architecture and hyperparameters with StyLEx model. We first train StyLEx with HUMMINGBIRD training set only to predict stylistic word scores for 50 epochs. Then we select the model with the best F1 score as a stylistic word score prediction to provide stylistic word scores for tokens in ORIGINAL training set. Then, we use both human-annotated perception score from HUMMINGBIRD and predicted stylistic word scores from ORIGINAL to train the sentence-level style prediction as in Figure 2. For the sentence-level model, we train the model for 5 epochs. For both stylistic word prediction and sentence-level style classification, we use BERT-base-uncased pretrained model. We set 0.1 dropout rate, 512 maximum sequence length, AdamW optimizer of learning rate $2e^{-5}$. For other hyper-parameters, we follow the default setting from HuggingFace's transformer library (Wolf et al., 2020).

Our interface for human evaluation is shown as in Figure 5. Table 7 shows results for other pretrained language models. We report word-level style predictor performance tested on HUMMINGBIRD test data in the form of Pearson's r correlation scores as follows in Table 5.

Style	**Word-level Pearson's r**
Politeness	0.41
Sentiment	0.61
Offensiveness	0.37
Anger	0.45
Disgust	0.43
Fear	0.20
Joy	0.37
Sadness	0.41

Table 5: Pearson's r scores from training the word-level predicion model and testing it on HUMMINGBIRD.

	HUMMINGBIRD		ORIGINAL			OOD
Styles↓	Train	Test	Train	Dev	Test	Test
Politeness	256 (38%)	64 (28%)	9,855 (55%)	530 (56%)	567 (57%)	1,000 (50%)
Sentiment	312 (30%)	79 (37%)	117,219 (55%)	825 (51%)	1,749 (50%)	5,200 (50%)
Offensiveness	400 (34%)	100 (32%)	20,680 (82%)	1,173 (82%)	1,159 (81%)	5,993 (50%)
Anger	400 (35%)	100 (34%)	6,838 (37%)	886 (36%)	3,259 (34%)	16,168 (50%)
Disgust	400 (43%)	100 (38%)	6,838 (38%)	886 (36%)	3,259 (34%)	10,602 (50%)
Fear	400 (17%)	100 (13%)	6,838 (18%)	886 (14%)	3,259 (15%)	6,394 (50%)
Joy	400 (24%)	100 (19%)	6,838 (36%)	886 (45%)	3,259 (44%)	15,966 (50%)
Sadness	400 (29%)	100 (17%)	6,838 (29%)	886 (30%)	3,259 (29%)	13,516 (50%)

Table 6: Dataset statistics in our experiments. Note that these datasets are preprocessed from existing datasets. For HUMMINGBIRD (Hayati et al., 2021) and ORIGINAL datasets, the train, dev, and test sets have the same size for all emotions. We do not report the training size of Out-of-Domain (OOD) datasets since we are not using them for training. The label distributions for positive labels are in the parentheses.

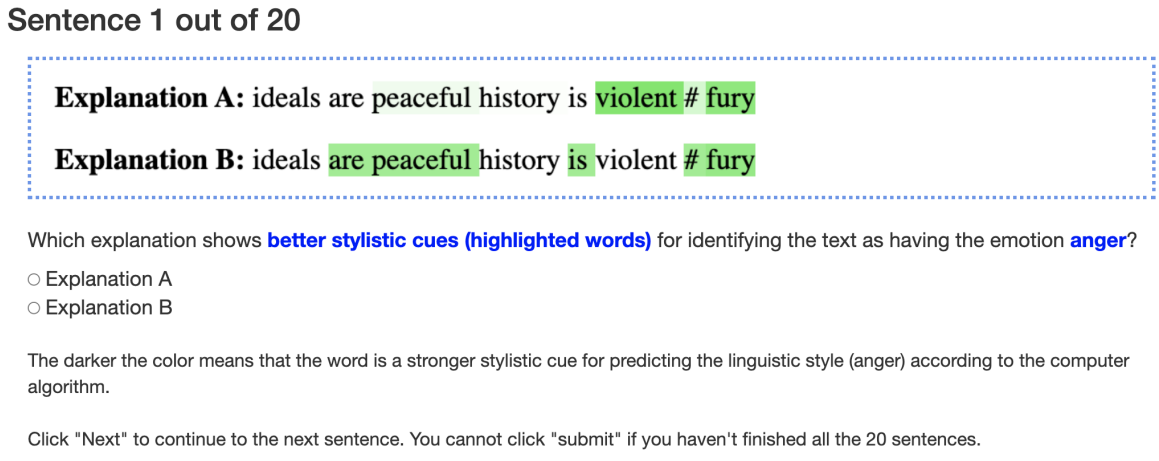

Figure 5: Interface for human evaluation

Model		F1 Score							
		Polite.	**Sent.**	**Offens.**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sad.**
ORIG									
BERT	Baseline	67.96	96.52	97.75	89.04	86.50	95.66	88.02	88.38
	StyLEx	65.84	96.59	97.81	89.01	86.90	95.63	88.14	88.41
RoBERTa	Baseline	65.83	96.94	96.40	89.56	87.17	95.68	88.32	88.52
	StyLEx	66.05	96.59	96.55	89.39	87.16	95.67	88.39	88.89
XLNet	Baseline	64.09	96.57	96.86	88.46	86.07	95.55	86.95	87.32
	StyLEx	63.69	96.54	96.38	88.20	86.32	95.44	87.33	87.90
T5	Baseline	65.75	97.13	97.21	88.79	86.29	95.39	88.18	87.68
	StyLEx	67.69	97.21	96.61	88.69	86.24	95.28	87.86	87.51
OOD									
BERT	Baseline	71.45	86.70	88.62	77.49	74.06	78.42	75.20	78.37
	StyLEx	74.18	86.99	88.98	77.51	74.63	78.48	74.26	78.71
RoBERTa	Baseline	72.08	90.41	87.48	77.01	74.56	79.95	74.63	80.24
	StyLEx	69.27	89.90	89.63	77.86	75.07	78.66	74.18	79.09
XLNet	Baseline	68.84	88.25	88.33	76.24	74.77	78.48	74.03	78.78
	StyLEx	67.28	89.65	88.56	76.41	74.71	78.92	74.09	78.28
T5	Baseline	70.76	91.72	88.14	75.74	73.83	79.58	73.76	78.37
	StyLEx	68.14	91.73	88.05	75.48	73.33	80.23	73.70	77.61

Table 7: More classification results with several language models.