Style is NOT a single variable: Case Studies for Cross-Style Language Understanding

Dongyeop Kang[*]
dongyeopk@berkeley.edu
UC Berkeley

Eduard Hovy
hovy@cs.cmu.edu
Carnegie Mellon University

Abstract

Every natural text is written in some style. Style is formed by a complex combination of different stylistic factors, including formality markers, emotions, metaphors, etc. One cannot form a complete understanding of a text without considering these factors. The factors combine and co-vary in complex ways to form styles. Studying the nature of the co-varying combinations sheds light on stylistic language in general, sometimes called *cross-style language understanding*. This paper provides the benchmark corpus (XSLUE) that combines existing datasets and collects a new one for sentence-level cross-style language understanding and evaluation. The benchmark contains text in 15 different styles under the proposed four theoretical groupings: figurative, personal, affective, and interpersonal groups. For valid evaluation, we collect an additional diagnostic set by annotating all 15 styles on the same text. Using XSLUE, we propose three interesting cross-style applications in classification, correlation, and generation. First, our proposed cross-style classifier trained with multiple styles together helps improve overall classification performance against individually-trained style classifiers. Second, our study shows that some styles are highly dependent on each other in human-written text. Finally, we find that combinations of some contradictive styles likely generate stylistically less appropriate text. We believe our benchmark and case studies help explore interesting future directions for cross-style research. The preprocessed datasets and code are publicly available.[1]

1 Introduction

People often use style as a strategic choice for their personal or social goals in communication (Hovy, 1987; Silverstein, 2003; Jaffe et al., 2009; Kang, 2020). Some stylistic choices implicitly reflect the author's characteristics, like personality, demographic traits (Kang et al., 2019), and emotions (Buechel and Hahn, 2017), whereas others are explicitly controlled by the author's choices for their social goals like using polite language, for better relationship with the elder (Danescu et al., 2013). In this work, we broadly call each individual linguistic phenomena as one specific type of *style*.

Style is not a single variable, but multiple variables have their own degrees of freedom and they co-vary together. Imagine an orchestra, as a metaphor of style. What we hear from the orchestra is the harmonized sound of complex combinations of individual instruments played. A conductor, on top of it, controls their combinatory choices among them, such as tempo or score. Some instruments under the same category, such as violin and cello for bowed string type, make a similar pattern of sound. Similarly, text reflects complex combination of multiple styles. Each has its own lexical and syntactic features and some are dependent on each other. Consistent combination of them by the author will produce stylistically appropriate text.

To the best of our knowledge, only a few recent works have studied style inter-dependencies in a very limited range such across demographical traits (Nguyen et al., 2014; Preoţiuc-Pietro and Ungar, 2018), across emotions (Warriner et al., 2013), across lexical styles (Brooke and Hirst, 2013), across genres (Passonneau et al., 2014), or between metaphor and emotion (Dankers et al., 2019; Mohammad et al., 2016).

Unlike the prior works, this work proposes the first comprehensive understanding of cross-stylistic language variation, particularly focusing on how different styles co-vary together in written text, which styles are dependent on each other, and how they are systematically composed to generate text.

[*]This work was done while DK was at CMU.
[1]https://github.com/dykang/xslue

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2376–2387
August 1–6, 2021. ©2021 Association for Computational Linguistics

Our work has following contributions:

- Aggregate 15 different styles and 23 sentence-level classification tasks (§3). Based on their social goals, the styles are categorized into four groups (Table 1): figurative, affective, personal and interpersonal.
- Collect a cross-style set by annotating 15 styles on the same text for valid evaluation of cross-stylistic variation (§3.3).
- Study cross-style variations in classification (§4), correlation (§5), and generation (§6):
  - our jointly trained classifier on multiple styles shows better performance than individually-trained classifiers.
  - our correlation study finds statistically significant style inter-dependencies (e.g., impoliteness and offense) in written text.
  - our conditional stylistic generator shows that better style classifier enables stylistically better generation. Also, some styles (e.g., impoliteness and positive sentiment) are condtradictive in generation.

2 Related Work

Definition of style. People may have different definitions in what they call 'style'. Several sociolinguistic theories on styles have been developed focusing on their inter-personal perspectives, such as Halliday's systemic functional linguistics (Halliday, 2006) or Biber's theory on register, genre, and style (Biber and Conrad, 2019).

In sociolinguistics, indexicality (Silverstein, 2003; Coupland, 2007; Johnstone, 2010) is the phenomenon where a sign points to some object, but only in the context in which it occurs. Nonreferential indexicalities include the speaker's gender, affect (Besnier, 1990), power, solidarity (Brown et al., 1960), social class, and identity (Ochs, 1990).

Building on Silverstein's notion of indexical order, Eckert (2008) built the notion that linguistic variables index a social group, which leads to the indexing of certain traits stereotypically associated with members of that group. Eckert (2000, 2019) argued that style change creates a new persona, impacting a social landscape and presented the expression of social meaning as a continuum of decreasing reference and increasing performativity.

Despite the extensive theories, very little is known on extra-dependency across multiple styles. In this work, we empirically show evidence of extra-linguistic variations of styles, like a formality, politeness, etc, but limited to styles only if we can obtain *publicly available resources for computing*. We call the individual phenomena a specific type of "style" in this work. We admit that there are many other kinds of styles not covered in this work, such as inter-linguistic variables in grammars and phonology, or high-level style variations like individual's writing style or genres.

Groups	Styles
INTERPERSONAL	Formality, Politeness
FIGURATIVE	Humor, Sarcasm, Metaphor
AFFECTIVE	Emotion, Offense, Romance, Sentiment
PERSONAL	Age, Ethnicity, Gender, Education level, Country, Political view

Table 1: Style grouping in XSLUE.

Cross-style analysis. Some recent works have provided empirical evidence of style inter-dependencies but in a very limited range: Warriner et al. (2013) analyzed emotional norms and their correlation in lexical features of text. Chhaya et al. (2018) studied a correlation of formality, frustration, and politeness but on small samples (i.e., 960 emails). Nguyen et al. (2014) focused on correlation across demographic information (e.g., gender, age) and with some other factors such as emotions (Preoţiuc-Pietro and Ungar, 2018). Dankers et al. (2019); Mohammad et al. (2016) studied the interplay of metaphor and emotion in text. Liu et al. (2010) studied sarcasm detection using sentiment as a sub-problem. Brooke and Hirst (2013) conducted a topical analysis of six styles: literary, abstract, objective, colloquial, concrete, and subjective, on different genres of text. Passonneau et al. (2014) conducted a detailed analysis of Biber's genres and relationship between genres.

3 XSLUE: A Benchmark for Cross-Style Language Understanding and Evaluation

3.1 Style selection and groupings

In order to conduct a comprehensive style research, one needs to collect a collection of different style datasets. We survey recent papers related to style research published in ACL venues and choose 15 widely-used styles that have publicly available annotated resources and feasible size of training dataset (Table 1). We plan to gradually increase the coverage of style kinds and make the benchmark more comprehensive in the future.

	Style & dataset	#S	Split	#L	Label (distribution)	B	Domain	Public	Task
INTERPERS.	**Formality**								
	GYAFC (Rao and Tetreault, 2018)	224k	given	2	formal (50%), informal (50%)	Y	web	N	clsf.
	Politeness								
	StanfPolite (Danescu et al., 2013)	10k	given	2	polite (49.6%), impolite (50.3%)	Y	web	Y	clsf.
FIGURATIVE	**Humor**								
	ShortHumor (CrowdTruth, 2016)	44k	random	2	humor (50%), non-humor (50%)	Y	web	Y	clsf.
	ShortJoke (Moudgil, 2017)	463k	random	2	humor (50%), non-humor (50%)	Y	web	Y	clsf.
	Sarcasm								
	SarcGhosh (Ghosh and Veale, 2016)	43k	given	2	sarcastic (45%), non-sarcastic (55%)	Y	tweet	Y	clsf.
	SARC (Khodak et al., 2017)	321k	given	2	sarcastic (50%), non-sarcastic (50%)	Y	reddit	Y	clsf.
	SARC_pol (Khodak et al., 2017)	17k	given	2	sarcastic (50%), non-sarcastic (50%)	Y	reddit	Y	clsf.
	Metaphor								
	VUA (Steen, 2010)	23k	given	2	metaphor (28.3%), non-metaphor (71.6%)	N	misc.	Y	clsf.
	TroFi (Birke and Sarkar, 2006)	3k	random	2	metaphor (43.5%), non-metaphor (54.5%)	N	news	Y	clsf.
AFFECTIVE	**Emotion**								
	EmoBank$_{valence}$ (Buechel and Hahn, 2017)	10k	random	1	negative, positive	-	misc.	Y	rgrs.
	EmoBank$_{arousal}$ (Buechel and Hahn, 2017)	10k	random	1	calm, excited	-	misc.	Y	rgrs.
	EmoBank$_{dominance}$ (Buechel and Hahn, 2017)	10k	random	1	being_controlled, being_in_control	-	misc.	Y	rgrs.
	DailyDialog (Li et al., 2017)	102k	given	7	noemotion(83%), happy(12%)..	N	dialogue	Y	clsf.
	Offense								
	HateOffensive (Davidson et al., 2017)	24k	given	3	hate(6.8%), offensive(76.3%)..	N	tweet	Y	clsf.
	Romance								
	ShortRomance	2k	random	2	romantic (50%), non-romantic (50%)	Y	web	Y	clsf.
	Sentiment								
	SentiBank (Socher et al., 2013)	239k	given	2	positive (54.6%), negative (45.4%)	Y	web	Y	clsf.
PERSONAL	**Gender** PASTEL (Kang et al., 2019)	41k	given	3	Female (61.2%), Male (38.0%)..	N	caption	Y	clsf.
	Age PASTEL (Kang et al., 2019)	41k	given	8	35-44 (15.3%), 25-34 (42.1%)..	N	caption	Y	clsf.
	Country PASTEL (Kang et al., 2019)	41k	given	2	USA (97.9%), UK (2.1%)	N	caption	Y	clsf.
	Politics PASTEL (Kang et al., 2019)	41k	given	3	LeftWing (42.7%), Centerist(41.7%)..	N	caption	Y	clsf.
	Education PASTEL (Kang et al., 2019)	41k	given	10	Bachelor(30.6%), Master(18.4%)..	N	caption	Y	clsf.
	Ethnicity PASTEL (Kang et al., 2019)	41k	given	10	Caucasian(75.6%), African(5.5%)..	N	caption	Y	clsf.

Table 2: Style datasets in XSLUE. #S and #L mean the number of total samples and labels, respectively. **B** means whether the labels are balanced (Y) or not (N). Every label is normalized, rangin g in $[0, 1]$. **Public** means whether dataset is publicly available or not. clsf. and rgrs. in Task denotes classification and regression, respectively.

We follow the theoretical style grouping criteria based on their social goals in Kang (2020) that categorizes styles into four groups (Table 1): PERSONAL, INTERPERSONAL, FIGURATIVE, and AFFECTIVE group, where each group has its own social goals in communication. This grouping will be used in our case studies as a basic framework to detect their dependencies.

3.2 Individual style dataset

For each style in the group, we pre-process existing style datasets or collect our own if there is no publicly available one (i.e., ShortRomance). We do not include datasets with small samples (e.g., $\leq$ 1K) due to its infeasibility of training a large model. We also limit our dataset to classify a single sentence, although there exists other types of datasets (e.g., document-level style classifications, classifying a sentence with respect to context given) which are out of scope of this work.

If a dataset has its own data split, we follow that. Otherwise, we randomly split it by 0.9/0.05/0.05 ratios for the train, valid, and test set, respectively. If a dataset has only positive samples (ShortHumor, ShortJoke, ShortRomance), we do negative sampling from literal text as in Khodak et al. (2017). We include the detailed pre-processing steps in Appendix §A.

3.3 Cross-style diagnostic set

The individual datasets, however, have variations in domains (e.g., web, dialogue, tweets), label distributions, and data sizes (See domain, label, and #S columns in Table 2). Evaluating a system with these individual datasets' test set is not an appropriate way to validate how multiple styles are used together in a mixed way, because samples from individual datasets are annotated only when a single style is considered.

To help researchers evaluate their systems in the cross-style setting, we collect an additional diagnostic set, called *cross-set* by annotating labels of 15 styles together on the same text from crowd workers. We collect total 500 sample texts from

Sentiment	0.81	Sarcasm	0.38
Politeness	0.75	Country	0.38
Formality	0.48	Humor	0.37
Gender	0.47	Education level	0.36
Emotion: Valence	0.43	Age	0.35
Emotion	0.42	Political view	0.32
Romance	0.42	Metaphor	0.29
Offense	0.41	Emotion: Arousal	0.26
Ethnicity	0.41	Emotion: Dominance	0.24

Table 3: Annotator's agreement (Krippendorff's alpha). The degree of gray shading shows good, moderate, and fair agreements.

two different sources: the first half is randomly chosen from test sets among the 15 style datasets in balance, and the second half is chosen from random tweets that have high variations across style prediction scores using our pre-trained style classifiers. Each sample text is annotated by five annotators, and the final label for each style is decided via majority voting over the five annotations. In case they are tied or all different from each other for multiple labels, we don't include them. We also include Don't Know option for personal styles and Neutral option for two opposing binary styles (e.g., sentiment, formality). The detailed annotation schemes are in Appendix §B.

Table 3 shows annotator's agreement on the cross-set. We find that annotator's agreement varies a lot depending on style: sentiment and politeness with good agreement, and formality, emotion, and romance with moderate agreement. However, personal styles (e.g., age, education level, and political view), metaphor, and emotions (e.g., arousal and dominance), show fair agreements, indicating how difficult and subjective styles they are.

3.4 Contribution

Most datasets in XSLUE except for Romance are collected from others' work. Following the data statement (Bender and Friedman, 2018), we cite and introduce individual datasets with their data statistics in Table 2. Our main contribution is to make every dataset to have the same pre-processed format, and distribute them with accompanying code for better reproducibility and accessibility. Besides this engineering effort, XSLUE's main goal is to invite NLP researchers to the field of cross-style understanding and provide them a valid set of evaluation for further exploration. As the first step, using XSLUE, we study cross-style language variation in various applications such as classification (§4), correlation (§5), and generation (§6).

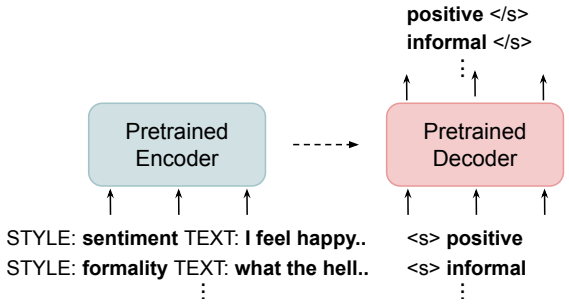

Figure 1: Our proposed cross-style classification model. The encoder and decoder are fine-tuned on the combined training datasets in XSLUE.

4 Case #1: Cross-Style Classification

We study how modeling multiple styles together, instead of modeling them individually, can be effective in style classification task. Particularly, the annotated cross-set in XSLUE will be used as a part of evaluation for cross-style classification.

Models. We compare two types of models: single and cross model. The single model is trained on individual style dataset separately, whereas the cross model is trained on shuffled set of every dataset together. For single model, we use various baseline models, such as majority classifier by choosing the majority label in training data, Bidirectional LSTM (biLSTM) (Hochreiter and Schmidhuber, 1997) with GloVe embeddings (Pennington et al., 2014), and variants of fine-tuned transformers; Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), robustly optimized BERT (RoBERTa) (Liu et al., 2019), and text-to-text transformer (T5) (Raffel et al., 2019).[2]

For cross model, we propose an encoder-decoder based model that learns cross-style patterns with the shared internal representation across styles (Figure 1). It encodes different styles of input as text (e.g., "STYLE: formality TEXT: would you please..") and decodes output label as text (e.g., "formal"). We use the pretrained encoder-decoder model from T5 (Raffel et al., 2019), and finetune it using the combined, shuffled datasets in XSLUE. Due to the nature of encoder-decoder model, we can take any training instances for classification tasks into the same text-to-text format. We also trained the single model (e.g., RoBERTa) on the combined datasets via a multi-task setup (i.e., 15 different heads), but showing less significant result.

[2]For a fair comparison, we restrict size of the pre-trained transformer models to 'base' model only, although additional improvement from the larger models is possible.

		Evaluation set →	*Individual-set evaluation*						*Cross-set evaluation* (§3.3)		
		Models →	single					cross	single		cross
	Style ↓	Dataset ↓	**Majority**	**biLSTM**	**BERT**	**RoBERTa**	**T5**	**Ours**	**BERT**	**T5**	**Ours**
INTER.	Formality	GYAFC	30.2	76.4	89.4	89.3	89.4	**89.9**	**37.3**	33.8	35.0
	Politeness	SPolite	36.2	61.8	68.9	70.4	**71.6**	71.2	60.0	62.1	**64.4**
FIGURATIVE	Humor	ShortHumor	33.3	88.6	97.3	97.5	97.4	**98.9**	-	-	-
		ShortJoke	33.3	89.1	98.4	98.2	98.5	**98.6**	**50.5**	47.2	47.9
	Sarcasm	SARC	33.3	63.0	71.5	**73.1**	72.4	72.8	**41.4**	37.7	37.4
		SARC_pol	33.3	61.3	73.1	**74.5**	73.7	**74.4**	-	-	-
	Metaphor	VUA	41.1	68.9	78.6	**81.4**	78.9	78.0	**49.8**	49.0	49.1
		TroFi	36.4	73.9	**77.1**	74.8	76.7	76.2	-	-	-
AFFECTIVE	Emotion	EmoBank$_{Valence}$	32.4	78.5	81.2	**82.8**	80.8	82.5	-	-	-
		EmoBank$_{Arousal}$	34.2	49.4	58.7	**62.3**	58.2	61.5	-	-	-
		EmoBank$_{Domin.}$	31.3	39.5	43.6	**48.3**	42.9	46.4	-	-	-
		DailyDialog	12.8	27.6	48.7	46.9	**49.2**	49.0	22.4	26.9	**33.3**
	Offense	HateOffens	28.5	68.2	91.9	92.4	91.7	**93.4**	34.4	36.9	**45.9**
	Romance	ShortRomance	33.3	90.6	99.0	**100.0**	98.0	99.0	**53.9**	55.2	48.2
	Sentiment	SentiBank	33.3	82.8	96.9	**97.4**	97.0	96.6	80.4	79.7	**84.6**
PERSONAL	Gender	PASTEL	25.7	45.5	47.7	47.9	47.3	**50.5**	29.2	32.4	**42.3**
	Age	PASTEL	7.3	15.2	23.0	21.7	21.3	**23.3**	**36.1**	27.0	28.1
	Country	PASTEL	49.2	49.3	54.5	49.3	51.8	**58.4**	**49.4**	46.7	48.7
	Political view	PASTEL	20.0	33.5	46.1	44.6	44.3	**46.7**	**27.7**	20.6	21.3
	Education	PASTEL	4.7	15.0	24.6	22.4	21.4	**27.3**	10.3	11.4	**15.7**
	Ethnicity	PASTEL	8.5	17.6	**24.4**	22.5	22.4	23.8	**10.8**	8.8	9.1
		Avearge	26.8	56.9	64.8	64.9	64.2	**65.9**	39.6	38.4	**40.7**

Table 4: Individual style (left) and cross style (right) classification in XSLUE. Every score is averaged over ten runs of experiments with different random seeds. For cross-style classification, we choose a single dataset per style, which has larger training data than the others. Otherwise, we leave it as a blank (-).

The detailed hyper-parameters used in our model training are in Appendix §C.

Tasks. Our evaluation has two tasks: *individual-set evaluation* for evaluating a classifier on individual dataset's test set (left columns in Table 4) and *cross-set evaluation* for evaluating a classifier on the annotated cross-set collected in §3.3 (right columns in Table 4).

Due to the label imbalance of datasets, we measure f-score (F1) for classification tasks and Pearson-Spearman correlation for regression tasks (i.e., EmoBank). For multi-labels, all scores are macro-averaged on each label.

Results. In the individual-set evaluation, compared to the biLSTM classifier, the fine-tuned transformers show significant improvements (+8% points F1) on average, although the different transformer models have similar F1 scores. Our proposed cross model, significantly outperforms the single model, by +1.7 percentage points overall F1 score, showing the benefit of learning multiple styles together. Particularly, the cross model significantly improves F1 scores on personal styles such as gender, age, and education level, possibly because the personal styles may be beneficial from detecting other styles. Among the styles, all personal styles, figurative styles (e.g., sarcasm and metaphor), and emotions are the most difficult styles to predict, which is similarly observed in the annotator's agreement in Table 3.

In cross-set evaluation, the overall performance significantly drops against the individual set evaluation, like from 65.9% to 40.7%, showing why it is important to have these annotated diagnostic set for valid evaluation of cross-style variation. Again, the cross-style model achieves +1.2% gain than the single models.

Figure 2 shows F1 improvement by the cross model against the single model BERT. Most styles obtain performance gain from the cross-style modeling, whereas not in the two metaphor style datasets (VUA, TroFi) and ethnicity style. This is possibly because metaphor tasks prepend the target metaphor verb to the input text, which is different from other task setups. Thus, learning them

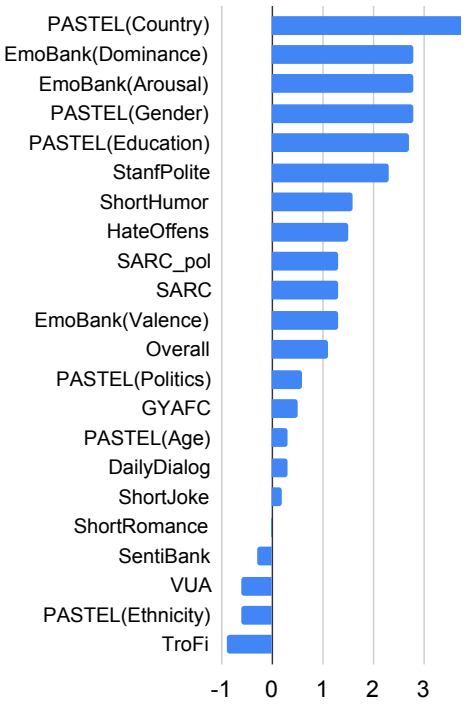

Figure 2: F1 improvement by our cross model over BERT in individual style classification task.

Target Style	Correlated styles	H
Humorous	Excitement emotion	5.0
	Negative sentiment	3.5
Polite	Positive valence emotion	4.5
	Happy emotion	4.0
	No offense	5.0
Positive sentiment	Happy emotion	4.5
	No offense	4.7
	No hate	4.7
Dominance emotion	Happy emotion	3.7
	Positive sentiment	3.7
Anger emotion	Disgust emotion	4.0
	Offense	5.0
Happy emotion	Romance	4.7
	Positive sentiment	4.7
Formal	Master education	4.0
Informal	High-school education	4.0
Non-humorous	Age 55<	3.7
	Doctorate education	4.0
High-school educ.	Excitement emotion	2.7
	Offense	3.0
Master education	Doctorate education	4.2
Caucasian	No Hispanic	4.2

Table 5: Some example pairs of positively (or negatively for "No") correlated styles with human judgement score (H).

together may harm the performance, although it is not significant.

5 Case #2: Style Dependencies

In addition to the theoretical style grouping in §3.1, we empirically find how two styles are correlated in human-written text using silver predictions from the classifiers.

Setup. We sample 1,000,000 tweets crawled using Twitter's Gardenhose API. We choose tweets as the target domain, because of their stylistic diversity compared to other domains, such as news articles. Using the fine-tuned cross-style classifier in §4, we predict probability of 53 style attributes[3] over the 1M tweets. We split a tweet into sentences and then average their prediction scores. We then produce a correlation matrix across the style attributes using Pearson correlation coefficients with Euclidean distance and finally output a 53×53 correlation matrix. We only show correlations that are statistical significant with p-value < 0.05 and cross out the rest.

Reliability. One may doubt about the classifier's low performance on some styles, leading to unreliable interpretation of our analysis. Although we only show correlation on the predicted style values, we also performed the same analysis on the human-annotated cross-set, showing similar correlation tendencies to the predicted ones. However, due to the small number of annotations, its statistical significance is not high enough. Instead, we decide to show the prediction-based correlation, possibly including noisy correlations but with statistical significance.

Results. Figure 3 shows the full correlation matrix we found. From the matrix, we summarize some of the highly correlated style pairs in Table 5 For each pair of correlation, two annotators evaluate its validity of stylistic dependency using a Likert scale. Our prediction-based correlation shows 4.18 agreement on average, showing reasonable accuracy of correlations.

We also provide an empirical grouping of styles using Ward hierarchical clustering (Ward Jr, 1963) on the correlation matrix. Figure 4 shows some interpretable style clusters detected from text, like Asian ethnicities (SouthAsian, EastAsian), middle ages (35-44, 45-54, 55-74), positiveness (happiness, dominance, positive, polite), and bad emotions (anger, disgust, sadness, fear).

6 Case #3: Cross-Style Generation

We study the effect of combination of some styles in the context of generation. We first describe our

[3] Attribute means labels of each style: *positive* and *negative* labels for sentiment style.

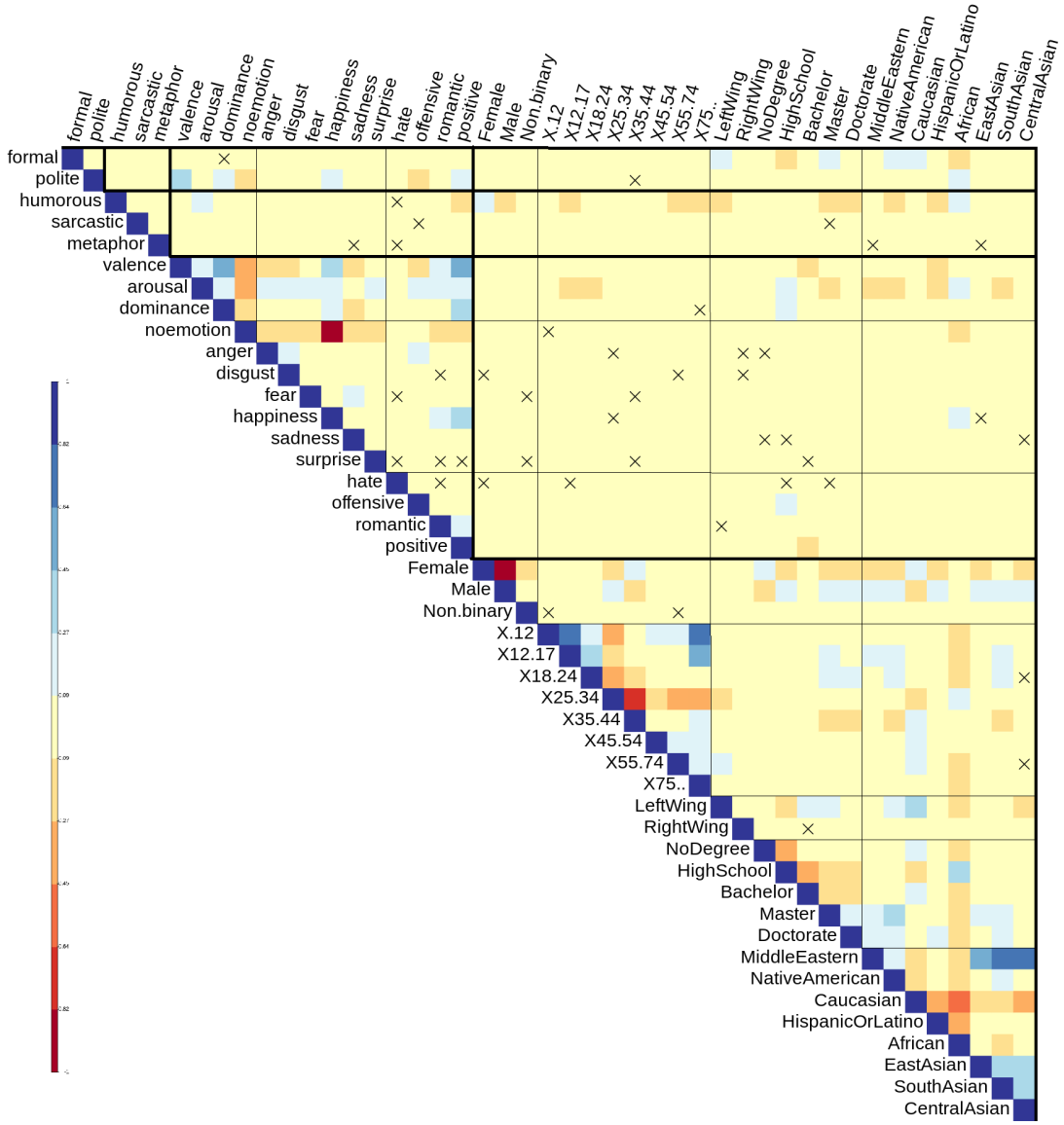

Figure 3: Cross-style correlation. Correlations with $p < 0.05$ (confidence interval: 0.95) are only considered as statistically significant. The degree of correlation gradually increases from red (negative) to blue (positive), where the color intensity is proportional to the correlation coefficients. We partition the correlation matrix into three pieces: across interpersonal, figurative and affective styles (upper left), between persona and a group of interpersonal, figurative, and affective styles (upper right), and across persona styles (lower right). IMPORTANT NOTE: *please be VERY CAREFUL not to make any unethical or misleading interpretations from these model-predicted artificial correlations.* Best viewed in color.

style-conditioned generators that combine the style classifiers in §4 with pre-trained generators (§6.1), and then validate two hypothetical questions using the generators: does better identification of styles help better stylistic generation (§6.2)? and which combination of styles are more natural or contradictive in generation (§6.3)?

6.1 Style-conditioned Generation

Let x an input text and s a target style. Since we already have the fine-tuned style classifiers $\mathrm{P}(s|x)$ from §4, we can combine them with a generator $\mathrm{P}(x)$, like a pre-trained language model, and then generate text conditioned on the target style $\mathrm{P}(x|s)$. We extend the plug-and-play language model (PPLM) (Dathathri et al., 2019) to combine our style classifiers trained on XSLUE with the pre-trained generator; GPT2 (Radford et al., 2019) without extra fine-tuning: $\mathrm{P}(x|s) \propto \mathrm{P}(x) \cdot \mathrm{P}(s|x)$. Table 6 shows example outputs from our style-conditioned generators given a prompt 'Every natural text is'.

We evaluate quality of output text: given 20 frequent prompts randomly extracted from our

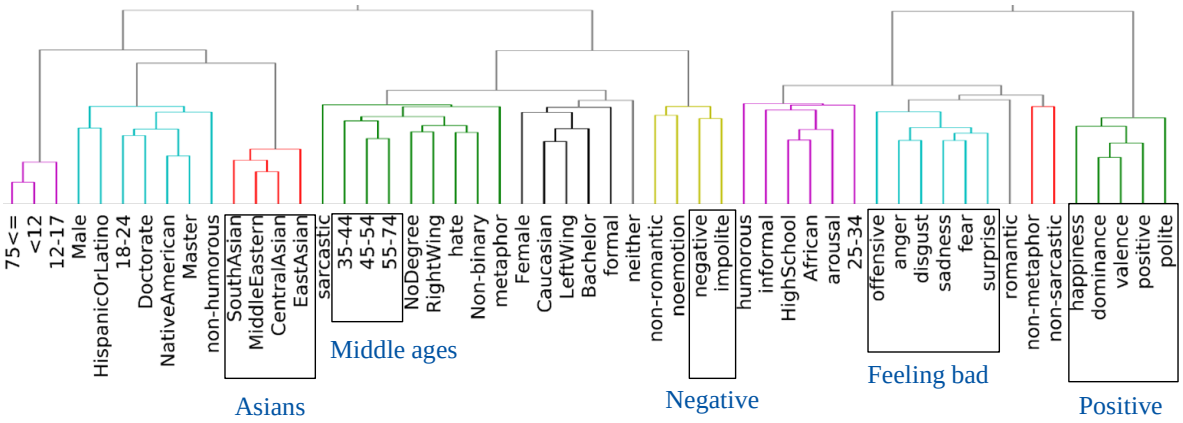

Figure 4: Empirical grouping of styles. Best viewed in color.

Output without style condition:

'Every natural text is' a series of images. The images, as they are known within the text, are the primary means by which a text is read, and therefore are ..

Output conditioned on **Formality** (F1 = 89.9%)
: Formal (left) and Informal (right)

'Every natural text is' different. You may find that the word you wrote does not appear on the website of the author. If you have any queries, you can contact us..	'Every natural text is' a bit of a hack. I don't think of it as a hack, because this hack is the hack.. and if you don't believe me then please don't read this, I don't care..

Output conditioned on **Offense** (F1 = 93.4%)
: Non-offensive (left) and Offensive (right)

'Every natural text is' a natural language, and every natural language is a language that we can speak. It is the language of our thoughts and of our lives..	'Every natural text is' worth reading...I'm really going to miss the music of David Byrne, and that was so much fun to watch live. The guy is a *ucking *ick. ..

Table 6: Given a prompt 'Every natural text is', output text predicted by our stylistic generator. The blue and red phrases are manually-labeled as reasonable features for each label. Offensive words are replaced with *.

training data,[4] we generate 10 continuation text for each prompt for each binary label of four styles (sentiment, politeness, offense, and formality)[5] using the conditional style generator; total $20 * 10 * 2 * 4$=1600 continuations.

We evaluate using both automatic and human measures: In automatic evaluation, we calculate F1 score of generated text using the fine-tuned classifiers, to check whether the output text reflects stylistic factor of the target style given. In human evaluation, scores (1-5 Likert scale) annotated by three crowd-workers are averaged on three metrics: *stylistic appropriateness*[6], *consistency with prompt*, and *overall coherence*.

	Sentiment	Politeness	Formality	Offense
XSLUE (F1)	96.5	71.2	89.8	93.3
Auto (F1)	73.7	70.1	60.0	63.7
Human(1^{st})	3.4/3.5/2.8	3.6/3.6/3.3	3.4/3.7/3.1	4.0/3.9/3.3
Human(2^{nd})	2.4/3.2/2.3	2.8/3.4/2.7	2.9/2.8/2.0	2.9/3.3/2.5

Table 7: Automatic and human evaluation on generated text. 1^{st} and 2^{nd} labels correspond to positive and negative for sentiment, polite and impolite for politeness, formal and informal for formality, and non-offensive and offensive for offense. Three numbers in human evaluation means stylistic appropriateness, consistency with prompt, and overall coherence in order.

In Table 7, compared to F1 scores on individual test set in XSLUE, automatic scores on output from the generator are less by 20.5% on average, showing sub-optimality of the conditional style generator between classification and generation. Interestingly, in human evaluation, negative labels (2^{nd} label for each style) for each style, like negative sentiment, impoliteness, informality, and offensiveness, show less stylistic appropriateness than positive or literal labels.

6.2 Better classification, better generation

To further investigate the relationship between classifier's performance and generation quality, we conduct a study by decreasing the training completion ratio (i.e., a fraction of epochs until completion; $C\%$) of the classifiers; $P_{C\%}(s|x)$ over the four styles and again evaluate the output continuation; $P_{C\%}(x|s) \propto P(x) \cdot P_{C\%}(s|x)$ using the same

[4]Some example prompts: "Meaning of life is", "I am", "I am looking for", "Humans are", "The virus is", etc

[5]We choose them by the two highest F1 scored styles each from inter-personal and affective groups, although we conduct experiments on other styles such as romance and emotions.

[6]Stylistically appropriateness means the output text includes appropriate amount of target style given.

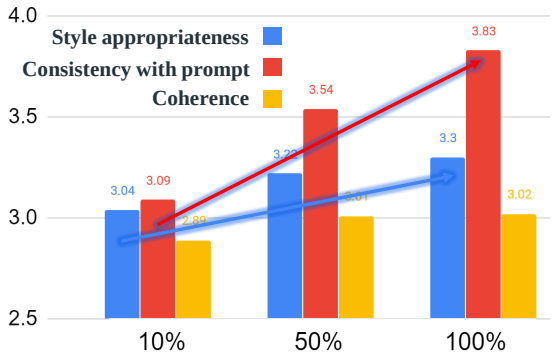

Figure 5: As the training completion ratio (x-axis, %) of classifiers increases, stylistic appropriateness (blue, y-axis) and consistency (red, y-axis) increase.

	Polite	Impolite
Pos	3.11	2.45
Neg	2.52	2.89

	Polite	Impolite
Pos	0.58	0.21
Neg	0.17	0.63

Table 8: Stylistic appropriateness scores (human judgement) on model-generated text with Likert scale (left) and style correlation scores from the correlation matrix (right) between politeness and sentiment.

human metrics. Figure 5 shows that the better style understanding (higher F1 scores in classification) yields the better stylistic generation (higher stylistic appropriateness and consistency scores).

6.3 Contradictive styles in generation

We have generated text conditioned on single styles. We now generate text conditioned on combination of multiple styles; $\mathrm{P}(x|s_1..s_k)\& \propto \mathrm{P}(x) \cdot \mathrm{P}(s_1|x) \cdots \mathrm{P}(s_k|x)$ where k is the number of styles. In our experiment, we set k=2 for sentiment and politeness styles, and generate text conditioned on all possible combinations between the labels of the two styles (e.g., positive and polite label, negative and impolite label). We again conduct human evaluation on the output text for measuring whether the generator produces stylistically appropriate text given the combination.

Table 8 shows averaged human-measured stylistic appropriate scores over the four label combinations (left) and the correlation scores observed in the style correlation matrix on written text in Figure 3 (right). Some combinations, like positive and impolite or like negative and polite, show less stylistic appropriateness scores, because they are naturally contradictive in their stylistic variation. Moreover, the stylistic appropriateness scores look similar to the correlation score observed from written text, showing that there exists some natural or unnatural combination of styles in both classification on human-written text and output generated by the model.

7 Conclusion and Discussion

We introduce a benchmark XSLUE of mostly existing datasets for studying cross-style language understanding and evaluation. Using XSLUE, we found interesting cross-style observations in classification, correlation, and generation case studies. We believe XSLUE helps other researchers develop more solid methods on various cross-style applications. We summarize other concerns we found from our case studies:

Style drift. The biggest challenge in collecting style datasets is to diversify the style of text but preserve the meaning, to avoid *semantic drift*. In the cross-style setting, we also faced a new challenge; *style drift*, where different styles are coupled so changing one style might affect the others.

Ethical consideration. Some styles, particularly on styles related to personal traits, are ethically sensitive, so require more careful interpretation of the results not to make any misleading points. Any follow-up research needs to consider such ethical issues as well as provides potential weaknesses of their proposed methods.

From correlation to causality. Our analysis is based on correlation, not causality. In order to find causal relation between styles, more sophisticated causal analyses, such as propensity score (Austin, 2011), need to be considered for controlling the confounding variables. By doing so, we may resolve the biases driven from the specific domain of training data. For example, generated text with the politeness classifier (Danescu et al., 2013) contains many technical terms (e.g., 3D, OpenCV, bugs) because its training data is collected from StackExchange.

Acknowledgements

This work would not have been possible without the efforts of the authors who kindly share the style language datasets publicly. We thank Edvisees members at CMU, Hearst lab members at UC Berkeley, and anonymous reviewers for their helpful comments.

References

Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Niko Besnier. 1990. Language and affect. *Annual review of anthropology*, 19(1):419–451.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*.

Julian Brooke and Graeme Hirst. 2013. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 673–679, Atlanta, Georgia. Association for Computational Linguistics.

Roger Brown, Albert Gilman, et al. 1960. The pronouns of power and solidarity.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Paul Chilton. 1990. Politeness, politics and diplomacy. *Discourse & Society*, 1(2):201–224.

Herbert H Clark and Dale H Schunk. 1980. Polite responses to polite requests. *Cognition*, 8(2):111–143.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

CrowdFlower. 2016. text Emotion. http://www.crowdflower.com/wp-content/uploads/2016/07/text_emotion.csv. [Online; accessed 1-Oct-2019].

CrowdTruth. 2016. Short Text Corpus For Humor Detection. http://github.com/CrowdTruth/Short-Text-Corpus-For-Humor-Detection. [Online; accessed 1-Oct-2019].

Niculescu-Mizil Cristian Danescu, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *IJCNLP 2019*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Penelope Eckert. 2000. *Language variation as social practice: The linguistic construction of identity in Belten High*. Wiley-Blackwell.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, 95(4):751–776.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169.

Michael Alexander Kirkwood Halliday. 2006. *Linguistic studies of text and discourse*, volume 2. A&C Black.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Suzana Ilić, Edison Marrese-Taylor, Jorge A Balazs, and Yutaka Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. *arXiv preprint arXiv:1809.09795*.

James B Jacobs, Kimberly Potter, et al. 1998. *Hate crimes: Criminal law & identity politics*. Oxford University Press on Demand.

Alexandra Jaffe et al. 2009. *Stance: sociolinguistic perspectives*. OUP USA.

Barbara Johnstone. 2010. Locating language in identity. *Language and identities*, 31:29–36.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong.

Dongyeop Kang. 2020. *Linguistically Informed Language Generation: A Multifaceted Approach*. PhD dissertation, Carnegie Mellon University.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, ph.d) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 89–94. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

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*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc T Tomlinson. 2016. Introducing the lcc metaphor datasets. In *LREC*.

Abhinav Moudgil. 2017. short jokes dataset. `https://github.com/amoudgl/short-jokes-dataset`. [Online; accessed 1-Oct-2019].

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Elinor Ochs. 1990. Indexicality and socialization. *Cultural psychology: Essays on comparative human development*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Rebecca J. Passonneau, Nancy Ide, Songqiao Su, and Jesse Stuart. 2014. Biber redux: Reconsidering dimensions of variation in American English. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 565–576, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Alfredo Láinez Rodrigo and Luke de Oliveira. Sequential convolutional architectures for multi-sentence text classification cs224n-final project report.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

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.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Samuel Walker. 1994. *Hate speech: The history of an American controversy*. U of Nebraska Press.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.