Meta-Tuning LLMs to Elicit Lexical Knowledge of Language Style

Ruohao Guo **Wei Xu** **Alan Ritter**
Georgia Institute of Technology
rguo48@gatech.edu; {wei.xu, alan.ritter}@cc.gatech.edu

Abstract

Language style is often used by writers to convey their intentions, identities, and mastery of languages. In this paper, we show that current large language models struggle to capture some of the language styles without fine-tuning. To address this challenge, we investigate whether LLMs can be meta-trained based on representative lexicons to recognize new language styles that they have not been fine-tuned on. Experiments on 13 established style classification tasks, as well as 63 novel tasks generated using LLMs, demonstrate that meta-training with style lexicons consistently improves zero-shot transfer across styles. Code and data to reproduce our experiments will be released upon publication.

1 Introduction

The style of a text refers to unique ways authors select words and grammar to express their message (Hovy, 1987). It can provide insights into social interactions and implicit communication. A notable example underscoring the importance of studying linguistic style used in communication is the analysis of body camera footage and transcripts (Voigt et al., 2017), where police officers have been found to use less respectful language towards black people than white people. Moreover, the open-ended and ever-evolving nature of language styles (Xu, 2017; Kang and Hovy, 2021) motivates the need for zero-shot classification, as it is costly to annotate data for every possible style in every language.

This leads to a natural question: *can recently developed instruction-tuned language models do well in identifying the style of texts without labeled data?* As we show in the paper (Table 2), this remains a challenge, even though these models have demonstrated impressive zero-shot performance on many other tasks (Chung et al., 2022; Ouyang et al., 2022). On the other hand, before the paradigm in NLP shifted to pre-trained language models, lexicons of words that are stylistically expressive were commonly used as important lexical knowldge (Verma and Srinivasan, 2019) in rule-based (Wilson et al., 2005; Taboada et al., 2011), feature-based (Mohammad et al., 2013; Eisenstein, 2017), and deep learning models (Teng et al., 2016; Maddela and Xu, 2018) for style identification. Many lexicons have been developed for varied styles, such as politeness (Danescu-Niculescu-Mizil et al., 2013), happiness (Dodds et al., 2015), emotions (Mohammad and Turney, 2010; Tausczik and Pennebaker, 2010), etc. This leads to another research question: *can we leverage lexicons during instruction fine-tuning of large language models (LLMs) to improve their understanding of language style?*

In this paper, we examine the effectiveness of fine-tuning LLMs to interpret lexicons that are provided as inputs to elicit latent knowledge (Kang et al., 2023) of language styles that were acquired during pre-training. We first compile a benchmark of 13 diverse writing styles with both annotated test sets and style-representative lexicons. Using this benchmark, we show that **meta-tuning with lexicons** enables different pre-trained LLMs to generalize better to new styles that have no labeled data. For example, meta-tuning LLaMA-2-7B (Touvron et al., 2023) on seven styles can improve the average F1 score on a separate set of six held-out styles by 12%, and by 8% over a general instruction-tuned model, LLaMA-2-Chat.

To further verify the capability of LLMs to generalize to novel styles using lexicons as the only source of supervision, we created a diverse set of 63 unique writing styles with examples using self-instruction (Wang et al., 2023). We demonstrate that using a small lexicon of just five words can effectively improve generalization to new styles. We found it helpful to replace class names with random identifiers when meta-training with lexicons, which prevents models from ignoring lexicons and simply memorizing source styles' class names.

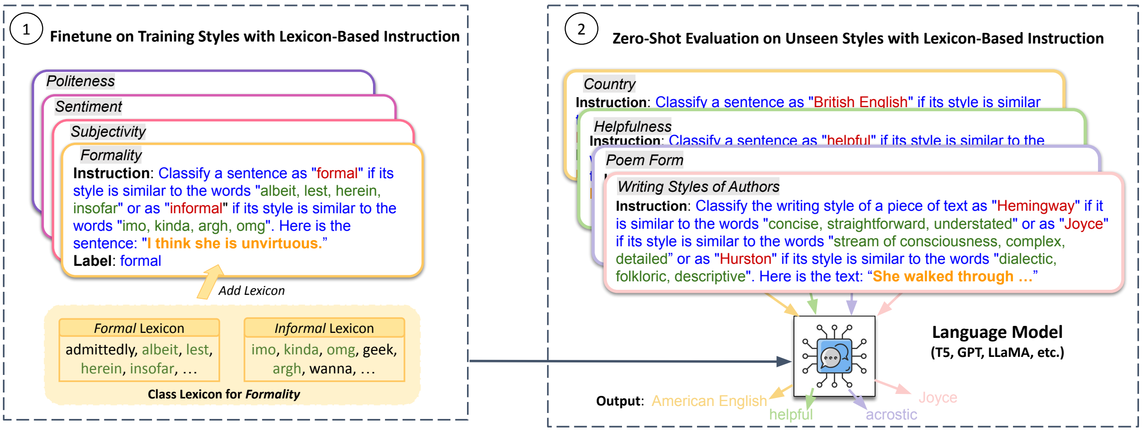

Figure 1: Overview of using lexicon-based instructions for cross-style zero-shot classification. It consists of two steps: (1) instruction tuning the model on training styles; (2) evaluating the learned model on unseen target styles zero-shot. A lexicon-based instruction is composed of instruction, class names, lexicons and an input.

Style Dataset	$\|C\|$	B?	Domain	#Tra, Val, Test	Lexicon Sources
Age (Kang and Hovy, 2021)	2	✗	caption	14k, 2k, 2k	ChatGPT, Dict
Country (Kang and Hovy, 2021)	2	✗	caption	33k, 4k, 4k	ChatGPT, Dict
Formality (Rao and Tetreault, 2018)	2	✓	web	209k,10k,5k	NLP (Wang et al., 2010), Dict
Hate/Offense (Davidson et al., 2017)	3	✗	Twitter	22k,1k,1k	NLP (Ahn, 2005), Dict
Humor (CrowdTruth, 2016)	2	✓	web	40k,2k,2k	ChatGPT, Dict
Politeness (Danescu-Niculescu-Mizil et al., 2013)	2	✓	web	10k,0.5k,0.6k	NLP (Danescu-Niculescu-Mizil et al., 2013), Dict
Politics (Kang and Hovy, 2021)	3	✗	caption	33k, 4k, 4k	NLP (Sim et al., 2013), Dict
Readability (Arase et al., 2022)	2	✗	web, Wiki	7k,1k,1k	NLP (Maddela and Xu, 2018), Dict
Romance (Kang and Hovy, 2021)	2	✓	web	2k,0.1k,0.1k	ChatGPT, Dict
Sarcasm (Khodak et al., 2018)	2	✓	Reddit	11k,3k,3k	ChatGPT, Dict
Sentiment (Socher et al., 2013)	2	✗	web	236k,1k,2k	NLP (Mohammad, 2021), Dict
Shakespeare (Xu et al., 2012)	2	✓	web	32k,2k,2k	NLP (Xu et al., 2012), Dict
Subjectivity (Pang and Lee, 2004)	2	✓	web	6k,1k,2k	NLP (Wilson et al., 2005), Dict

Table 1: Statistics of datasets. "$|C|$" denotes the number of classes in each style dataset. "B?" indicates whether or not the class distribution is balanced. "#Tra, Val, Test" lists the number of examples in train, validation and test sets.

Model	Meta-Tuned?	Instruction	Shakespeare	Romance	Humor	Country	Sarcasm	Age	Avg.
Flan-T5$_{base}$	✗	Standard	33.36	33.33	33.33	43.15	33.33	33.92	35.07
	✗	+ Lex	49.95	51.30	48.66	35.34	49.40	49.02	**47.28**
Style-T5$_{base}$	✓	Standard	33.31	43.57	36.43	19.86	33.37	35.75	33.72
	✓	+ Lex	55.10	78.98	60.56	49.09	49.25	50.80	**57.30**
Style-GPT-J	✓	Standard	58.16	87.82	33.33	53.11	44.10	35.25	51.96
	✓	+ Lex	56.76	83.99	55.86	44.97	48.84	47.47	**56.32**
LLaMA-2-Chat (7B)	✗	Standard	60.20	85.72	43.84	49.19	36.02	38.91	52.31
	✗	+ Lex	62.59	88.95	51.01	50.88	42.88	36.54	**55.47**
LLaMA-2-Chat (13B)	✗	Standard	61.99	97.00	47.42	17.96	43.26	48.16	52.63
	✗	+ Lex	63.49	95.00	55.15	24.41	44.66	53.88	**56.10**
LLaMA-2 (7B)	✗	Standard	42.13	64.41	37.38	48.27	48.84	37.13	46.36
	✗	+ Lex	50.21	77.86	45.44	49.86	47.72	47.63	**53.12**
Style-LLaMA (7B)	✓	Standard	40.91	41.65	48.88	48.92	49.02	49.80	46.53
	✓	+ Lex	59.03	88.97	57.64	51.52	50.83	50.53	**59.75**

Table 2: Macro-average F1 scores for zero-shot performance on unseen evaluation styles. We compare the models fine-tuned on general instruction tuning data (i.e., not meta-tuned) and the "Style-*" models that are instruction-tuned on our training styles (i.e., meta-tuned). For each model, we evaluate its zero-shot learning capabilities when the standard and lexicon-based instructions are used, respectively.

Ethics Statement

Style classification is widely studied in the NLP research community. We strictly limit to using only the existing and commonly used datasets that are related to demographic information in our experiments. As a proof of concept, this research study was only conducted on English data, where human annotations for multiple styles are available for use in the evaluation. We also acknowledge that linguistic styles are not limited to what are included in this paper, and can be much more diverse. Future efforts in the NLP community could further extend research on stylistics to more languages and styles.

References

Luis Von Ahn. 2005. Useful resources: Offensive/profane word list. https://www.cs.cmu.edu/~biglou/resources/.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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

Cristian Danescu-Niculescu-Mizil, 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 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394.

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

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

Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-style language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Junmo Kang, Hongyin Luo, Yada Zhu, James Glass, David Cox, Alan Ritter, Rogerio Feris, and Leonid Karlinsky. 2023. Self-specialization: Uncovering latent expertise within large language models. *arXiv preprint arXiv:2310.00160*.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

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*.

Saif M. Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement (Second Edition)*. Elsevier.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Annual Meeting of the Association for Computational Linguistics*.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*.

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*.

Zhiyang Teng, Duy-Tin Vo, and Yue Zhang. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1629–1638, Austin, Texas. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Gaurav Verma and Balaji Vasan Srinivasan. 2019. A lexical, syntactic, and semantic perspective for understanding style in text. *arXiv preprint arXiv:1909.08349*.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*.

Tong Wang, Julian Brooke, and Graeme Hirst. 2010. Inducing lexicons of formality from corpora.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Wei Xu. 2017. From shakespeare to Twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation*, pages 1–9, Copenhagen, Denmark. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING*, pages 2899–2914.