Data Augmentation via Knowledge Graph Representation as a Way of Mitigation of Spurious Correlations

Alina Hancharova
University of Tübingen
alina.hancharova@student.uni-tuebingen.de

Ali Zhunis
University of Tübingen
ali.zhunis@student.uni-tuebingen.de

Abstract

Robustness is essential for the successful performance of machine learning models, particularly in addressing the instability of predictions caused by spurious correlations between labels and features. This instability complicates the unbiased judgments of large language models (LLMs). Recent research has focused on reducing these spurious correlations in data before fine-tuning. Our experiment builds on the work of the University of Maryland team titled "Explore Spurious Correlations at the Concept Level in Language Models for Text Classification."(Zhou et al., 2023). We propose a novel approach to mitigate biased correlations by augmenting the data with graph representations before fine-tuning DistilBERT. Our findings indicate that this method achieves a more balanced reduction of bias, as evidenced by a decrease in the Bias@C metric while maintaining high accuracy.

1 Introduction

The LLM's robustness is crucial to its success. In an ideal world, the model shouldn't biased toward any concept and give as accurate an answer as possible. Nowadays, pre-trained LLM reaches impressive results in various fields(Brown, 2020; Wei et al., 2022; Koroteev, 2021), but, their performances might be worsened by spurious correlations(Tang et al., 2023). These correlations appear during the fine-tuning phase, creating a false-learned bound between label and feature due to their high frequency in data (Wang and Culotta, 2020). Unfortunately, are not easily eliminated in a model instance.

The current research on spurious correlations in LLMs shows that there is no certain answer to what level of information leads to their creation. The spurious correlations are detected on the lexical level (Joshi et al., 2022), syntactic level (McCoy, 2019), and contextual circumstances(Bian et al., 2023). It should be also mentioned that false association might occur out of linguistic features like social aspects of human lives(race, gender, age, etc.) (Cheng et al., 2022; Czarnowska et al., 2021; Hutchinson et al., 2020).

Our experiment is a continuation of the survey, which was held by the University of Maryland (Zhou et al., 2023) and described in the paper "Explore Spurious Correlations at the Concept Level in Language Models for Text Classification".We agree with the authors' suggestion:

> *LMs' tendency to learn concept-level shortcuts might stem from forming similar embeddings for expressions related to the same concept during fine-tuning or pre-training, driven by their semantic similarities.*

motivated us to explore ways to mitigate the embeddings' strong bends.

In Section 2, we will briefly describe the Maryland experiment (Zhou et al., 2023). The section will contain information on the method of concept selection for a dataset, the structure of the experiment, and central ideas on the mitigation of spurious correlations. Then, we will state our dataset-obtaining method. We give the reasoning for a graph representation of data as a key for enhancing the model awareness of relationships within a sentence. To receive sentences' graphs, we employ the Mistral model in the Ollama setting via the promoting technique (Chang et al., 2024).In section 4, we illustrate the results of the fine-tuning DistillBERT model on a dataset with graph representations. Also, we will replicate the Maryland experiments with the original dataset and the biased dataset to see the effectiveness of our approach.

During our research, we have found out that:

- While the knowledge graph approach showed a minor rise in accuracy, it consistently achieved a better balance between bias reduction and model performance compared to other mitigation approaches, making it a robust strategy for mitigating biased correlations.

2 Previous work

In the paper "Explore Spurious Correlations at the Concept Level in Language Models for Text Classification," the authors run several experiments to check if LLMs tend to shortcut on the concept level under fine-tuning and in-context-learning circumstances. Firstly, the authors assign concepts to the dataset using ChatGPT. The process of obtaining a concept involves an annotation prompt P_α that contains the annotation instruction (See B) and five demonstrations, a text input x, LLM M_α, and a candidate concept set $C = \{C_1, C_2, \cdots, C_k\}$. The annotation process is formalized as:

$$\alpha(x) = M_\alpha(P_\alpha \parallel C \parallel x),$$

where $\alpha(x)$, the set of concept labels for text x, may contain zero or several concepts selected from the pre-defined concept set ($\alpha(x) \subset C$), and $\parallel$ denotes the concatenation operation. After concepts were created, they analyzed the 3 top concepts of each of 4 datasets: Amazon Shoe(He and McAuley, 2016), IMDB(Maas et al., 2011), Yelp (Zhang et al., 2015), and CeBab(Abraham et al., 2022). (See Table 1). To highlight the impact of imbalanced concept-label distribution, a biased dataset is constructed, keeping only the majority class (positive or negative) for each concept. For example, the negative class is kept for "size" in Amazon Shoe and "service" in Yelp, while the positive class is kept for other concepts in the sentiment datasets. In the final, authors measure spurious correlations for both biased and original dataset versions to confirm that, indeed, LLms tend to shortcut the predictions to certain labels.

The discrepancy measurement is defined as:

$$\begin{aligned}\Delta c_i = E_x[p_M(\hat{y} = i \mid x, c, y = i)] \\ - E_{x'}[p_M(\hat{y} = i \mid x', \neg c, y = i)] = 0 \quad (1)\end{aligned}$$

where $\neg c$ denotes that concept c is not in the input x. They hypothesize that if there exists a spurious correlation between concept c and label i, the following conditions would hold:

$$\mathrm{E}_x[p_M(\hat{y} = i \mid x, c, y = i)] > E_{x'}[p_M(\hat{y} = i \mid x', \neg c, y = i)]$$
$$\mathrm{E}_x[p_M(\hat{y} = j \mid x, c, y = j)] < E_{x'}[p_M(\hat{y} = j \mid x', \neg c, y = j)]$$

Then we have $\Delta c_i > 0 > \Delta c_j$. Otherwise, if the spurious correlation is between c and j, then $\Delta c_j > 0 > \Delta c_i$.

To quantify the spurious correlation, we propose measuring the average discrepancy in the accuracy difference across all label combinations:

$$\text{Bias@C} = \frac{1}{n^2} \sum_{i,j \in Y} (\Delta c_i - \Delta c_j) \quad \text{for } i > j$$

For the binary classification task, this is simplified to:

$$\text{Bias@C} = \Delta c_1 - \Delta c_0$$

A Bias@C approaching 0 indicates minimal reliance on concept shortcuts. Conversely, a positive Bias@C value suggests that the model is more likely to predict larger labels when the input includes concept c, and the opposite for a negative value. Additionally, authors report test accuracy for utility performance. To address potential label imbalance with or without concept c, they rebalance the test set by downsampling (Chew et al., 2024). Then report inference accuracy on the balanced subset for examples with concept c (Acc@C) and without concept c (Acc@NoC). [1]

Dataset	# Training	# Test	# Labels	Concept
AS	70,117	8,000	5	size, color, style
IMDB	14,956	4,000	2	acting, comedy, music
Yelp	34,184	4,000	2	food, price, service
CeBaB	7,350	2,000	5	service, food, ambiance

Table 1: Dataset statistics with the labels and concepts for each.

3 Dataset and Method

Recent findings in Knowledge Graph (KG) as a beneficiary for LLMs reasoning in the task of question answering and recommendation systems(Liang et al., 2023, 2024) motivates us to implement KG data augmentation as a way of spurious correlations mitigation (all data and code notebooks can be found in our GitHub repository.). We claim that the graph representation of sentences makes information more LLM-friendly, which increases the certainty of sentimental bound estimation within words. To obtain a graph version of 4 datasets, we

[1] Check out the code on our GitHub Repository.

employ the Mistral-7B model via Ollama with a certain prompt (see Appendix A). After a generation of graphs, they are combined with original sentences and labels and sent to the BERT model for fine-tuning DistillBert.

4 Results

We fine-tune DistillBERT from the Maryland experiment on our dataset and compare them by Bias@C, Acc@NoC, and Acc@C. We divide our results section into two parts to check:

- if graph representations outperform original dataset
- if graph representations outperform mitigation approaches

4.1 Spurious correlations in comparison to original data

The spurious correlations across three dataset versions such as original, biased, and knowledge graphs were evaluated in order to identify how each data construction approach impacts Bert's reliance on concepts (See Table 2). In the original datasets, significant spurious correlations were observed, with certain concepts showing a strong association with output labels.

During the evaluation of the biased dataset, we noticed that it amplifies the spurious correlations.

When it comes to the IMDB dataset, the bias value for the "music" concept significantly rose from 14.57 in the original dataset to 17.8 in the biased one. These changes reveal that BERT's tendency to base predictions on certain concepts became more pronounced with the biased dataset approach.

In comparison, the knowledge graph dataset demonstrated a significant decrease in spurious correlations. In the Amazon Shoes dataset, the graph-based approach significantly reduced bias across all concepts in comparison to the original and biased datasets, with the lowest values for "size" (3.3), "color" (7.6), and "style" (12.1). Although the original dataset achieved the highest accuracy for "style" (61.83), the graph dataset offered a better balance of bias reduction and accuracy, particularly for "color" where it outperformed with an accuracy of 57.18.

This alignment suggests that the addition of knowledge graph representations can help in the reduction of BERT's dependence on specific concepts. Similar positive effects were observed in the IMDB, Yelp, and CeBaB datasets, where graph-based fine-tuning lowered bias values and balanced prediction accuracy between examples with and without concepts.

4.2 Spurious correlations mitigations in comparison to data augumintation

The authors of "Explore Spurious Correlations at the Concept Level in Language Models for Text Classification" confirm the hypothesis of concept-based dataset sensitivity and give ideas on how to manage spurious correlations. In our experiment, we mention only 2 of them: downsampling (cut sentences to balance data) and upsampling (after detection of a weak concept add an extra sentence to an example related to a weak topic to strengthen the difference of semantic meanings). We drop the Mask approach (masking emotionally colored words to decrease semantical impact) because the authors find this approach less effective in comparison to the two given above. The results of mitigations are given in Table 3.

For the Amazon Shoes dataset, the knowledge graph-based approach produced the lowest bias values for "size" (3.3) and "style" (12.1) concepts, outperforming both downsampling and upsampling techniques. Although downsampling slightly outperformed in terms of accuracy for "style" (63.11), the knowledge graph-based dataset achieved a perfect balance between bias reduction and accuracy, especially with accuracy values closely aligned across all methods.

In the IMDB dataset, the downsampling approach achieved the lowest bias value for "acting" (1.52) and the upsampling approach achieved the lowest value for the "comedy" (-0.18) concept. Also, the upsampling method achieved high accuracy with values of 91.86 and 91.63 for comedy and music concepts. The graph knowledge and upsampling methods performed similarly in terms of bias reduction for "music," with bias values of 9.47 and 10.07, respectively, but the graph dataset maintained slightly higher accuracy overall, particularly in accuracy with a concept for "music" at 92.33.

For the Yelp dataset, the knowledge graph-based approach produced the lowest bias values for the "food" (1.33) and "service" (-0.45) concepts, outperforming both downsampling and upsampling techniques. Moreover, the knowledge graph dataset reached the highest accuracy for "food" and "price" concepts achieving 97.59 and 95.28, respectively.

Amazon shoe	Size(pos<neg)			Color(pos>neg) S			Style(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Original dataset	6.6	**59.72**	53.33	9.5	**57.32**	54.38	16.5	57.02	**61.83**
Biased dataset	-5.9	59.50	49.05	21.2	57.30	50.90	16.8	56.23	60.98
Graph dataset	**3.3**	59.07	**53.79**	**7.6**	57.31	**57.18**	**12.1**	**57.13**	60.74
IMDB	Acting(pos>neg)			Comedy(pos>neg)			Music(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Original dataset	**0.32**	**91.09**	**93.36**	5.76	89.87	**91.93**	14.57	91.03	91.81
Biased dataset	17.8	88.40	89.30	-3.96	**90.45**	91.82	10.4	**91.66**	92.27
Graph dataset	5.23	90.60	92.95	**-0.24**	90.39	91.45	**9.47**	91.47	**92.33**
Yelp	Food(pos>neg)			Service(pos<neg)			Price(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Original dataset	4.32	93.81	**97.77**	-0.62	**93.00**	**97.43**	4.15	95.15	**95.57**
Biased dataset	1.94	**94.75**	97.24	-4.12	92.85	95.92	8.82	95.12	93.94
Graph dataset	**1.33**	94.65	97.59	**-0.45**	92.66	96.72	**3.00**	**95.22**	95.27
CeBaB	Service(pos>neg)			Food(pos>neg)			Ambiance(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Original dataset	**3.99**	69.52	**74.69**	4.73	**73.40**	**73.60**	4.79	67.78	**74.54**
Biased dataset	21.78	67.78	60.23	18.74	62.34	54.31	9.24	70.69	66.62
Graph dataset	6.30	**70.16**	74.47	**3.64**	71.40	71.99	**1.93**	**71.29**	73.18

Table 2: Results of DistillBERT fine-tuning in percentage representations

Amazon shoe	Size(pos<neg)			Color(pos>neg) S			Style(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Graph dataset	**3.3**	59.07	53.79	7.6	57.31	57.18	**12.1**	**57.13**	60.74
Downsampling	5.82	**59.42**	53.99	12.23	**57.59**	55.63	15.84	56.99	**63.11**
Upsampling	6.88	59.32	**54.36**	**7.03**	57.03	**57.30**	13.75	57.08	60.13
IMDB	Acting(pos>neg)			Comedy(pos>neg)			Music(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Graph dataset	5.23	90.60	92.95	-0.24	90.39	91.45	9.47	91.47	**92.33**
Downsampling	**1.52**	**90.86**	**93.32**	3.19	90.67	91.82	10.07	91.44	91.81
Upsampling	5.64	90.08	92.10	**-0.18**	**91.86**	**92.95**	**2.57**	**91.63**	91.81
Yelp	Food(pos>neg)			Service(pos<neg)			Price(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Graph dataset	**1.33**	**94.65**	**97.59**	**-0.45**	92.66	96.72	3.00	**95.22**	**95.28**
Downsampling	4.22	93.76	96.71	-1.43	93.40	**97.33**	3.52	94.52	95.27
Upsampling	2.1	94.26	**97.59**	1.87	**93.75**	96.78	**-2.93**	94.70	95.27
CeBaB	Service(pos>neg)			Food(pos>neg)			Ambiance(pos>neg)		
	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C	Bias@C	Acc@NoC	Acc@C
Graph dataset	6.30	70.16	74.47	**3.64**	71.40	71.99	1.93	**71.29**	73.18
Downsampling	**4.37**	**70.55**	**74.78**	3.89	**71.70**	**74.25**	4.97	70.92	74.96
Upsampling	5.83	70.44	74.69	6.81	70.85	72.55	**0.99**	71.00	**75.11**

Table 3: Results of fine-tuning on graphs representation in comparison to downsampling and upsampling techniques

It suggests that graph representations not only mitigated spurious correlations more effectively but also preserved high accuracy in the Yelp dataset.

In the CeBaB dataset, the knowledge graph-based approach yielded moderate bias values for "service" (6.30) and "ambiance" (1.93). Conversely, upsampling reduced the bias for "ambiance" even further to 0.99 and achieved the highest accuracy with the concept included, reaching 75.11. For the "food" concept, the graph dataset slightly outperformed others in bias reduction (3.64). These findings demonstrate that graph representations consistently balance reducing bias while maintaining performance across different concepts.

5 Conclusion

After running the experiment, we see a potential in the use of knowledge graphs as a mitigator of spurious correlations. In the comparison of a graph-based dataset to original and biased variants, the graphs-included datasets show a stable decrease over the different concepts in both accuracy and bias metrics, which leads to the conclusion that graph representations have a positive impact on lowering biased correlations.

In the case of mitigations, data augmentation with knowledge graph also reaches a significant decrease in the Bias@C metric as opposed to down-sampling and upsampling techniques, even though losing tenths of a percent of winning solutions. For instance, Yelp dataset numbers show that 2 concepts out of 3 knowledge graphs perform the best from the point of Bias@C and 2 out 3 in case of accuracy (see Table 3).

We suggest that an extension of information in the form of a graph representation makes a more machine-friendly sample for training. Future research could explore optimizing graph construction techniques, combining this approach with other augmentation methods, or applying it across different domains to further improve performance and bias mitigation.

6 Future Work

The field of graph representations can be extended with the actual creation of knowledge graphs from a given dataset like in KAG systems (Liang et al., 2024). These might lead to better control of our relationships within sentences and, of course, a chance to detect spurious correlations and edit them by fine-tuning edges between nodes.

The Maryland experiment also tests spurious correlations in context learning under Llama2. This will give more evidence for graph usage as a mitigation tool for spurious correlations, but currently, our experiment has only results in the case of Encoder-model architecture.

7 Limitations

The main obstacle to the research was a lack of computational power. Generating graph representation using an advanced model instead of Mitral-7B might be more proficient. Moreover, we needed to create a new dataset by preprocessing over 100,000 sentences to conduct this experiment, which took us over 150 GPU hours (we use T4 GPU). Also, we chose a promting technique to obtain graphs, which leads to the risk of losing accuracy in answers.

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. ArXiv:2205.14140.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

Lu Cheng, Suyu Ge, and Huan Liu. 2022. Toward understanding bias correlations for mitigation in nlp. *arXiv preprint arXiv:2205.12391*.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide*

Web, WWW '16. International World Wide Web Conferences Steering Committee.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. *arXiv preprint arXiv:2210.14011*.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2023. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multimodal.

Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, Huaidong Xiong, Lin Yuan, Jun Xu, Zaoyang Wang, Zhiqiang Zhang, Wen Zhang, Huajun Chen, Wenguang Chen, and Jun Zhou. 2024. Kag: Boosting llms in professional domains via knowledge augmented generation.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

RT McCoy. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. *arXiv preprint arXiv:2305.17256*.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. *arXiv preprint arXiv:2010.02458*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2023. Explore spurious correlations at the concept level in language models for text classification. *arXiv preprint arXiv:2311.08648*.

A Graph Promting

The current prompt was used in the Mistrael model to create graph representations of datasets:

You are a network graph maker who extracts terms and their relations from a given context. You are provided with a context chunk (delimited by "'). Your task is to extract the ontology of terms mentioned in the given context. These terms should represent the key concepts as per the context. Thought 1: While traversing through each sentence, think about the key terms mentioned in it. Terms may include object, entity, location, organization, person, acronym, documents, service, concept, etc. Terms should be as atomistic as possible. Thought 2: Think about how these terms can have one-on-one relations with other terms. Terms that are mentioned in the same sentence or the same paragraph are typically related to each other. Terms can be related to many other terms. Thought 3: Find out the relation between each such related pair of terms. Format your output as a list of JSON. Each element of the list contains a pair of terms and the relation between them, like the following:

```
[{
"node1":
"A concept from extracted
ontology",

"node2":
"A related concept from
extracted ontology",

"edge":
"relationship between the
two concepts,node1 and node2"
 }]
```

B Promt of previous experiment

The current prompt was used to create a training materials for DistillBERT fine-tuning

I will provide you with 5 reviews in {datasetname} dataset. Please find the concepts explicitly mentioned in this review only from the set with three concepts: {candidateconcepts}. Do not include other concepts. If you cannot find any of these concepts in the concept set, please annotate this review with "none". Wrap your answer for a review in a word sequence separated by commas and for each answer, start with a new line with an index. Here are a few examples: {demonstrations} The output is: {outputconcepts} Here is the review list of 5 OpenTable reviews: {textlists} The output is: