Enhancing the Model Robustness and Generalization in Sentence Sentiment Classification Based on Causal Representation Learning Techniques

Yanyi Peng[1], **Xiang Li**[2*]

[1] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[2] School of Information Science and Engineering, Yanshan University
yanyipeng@link.cuhk.edu.hk, lixiang_222@stumail.ysu.edu.cn

Abstract

Sentiment classification is an important task in natural language processing (NLP), aiming to perform sentiment analysis on sentences. One of widely used method based on the causal word detection first estimates the treatment effect between words and sentiment of sentences, and then removing words with low treatment effect in sentiment classification model training. However, the previous works regard whether the specific word appears in the sentence as the binary treatment, which limits the robustness of identify the treatment effect of word, especially for the low-frequency word. To bridge this gap, we propose a novel causal representation learning method that regarding word representation as treatments to ensure the generalization of the sentiment classifier. Specifically, the method begins by clustering words based on their representations obtained from a pre-trained language model. Subsequently, corresponding to the clusters, a multi-head word classifier is trained to estimate the treatment effect of each word to identify whether this word is causally or spurious correlated to the sentiment. To ensure covariate balancing between each treatment cluster, we utilize the integral probability metric (IPM) distance to learn the balanced representation of the context. Then, the balanced representation and estimated treatment effects are used to train a more robust and generalizable sentiment classification model. Extensive experiments on public datasets demonstrate the effectiveness of our method in identifying causal words and improving the performance of sentiment classification.

Introduction

Sentiment classification, which aims to determine the sentiment polarity (e.g., positive or negative) expressed in a given sentence, is a fundamental task in natural language processing (NLP) (Pang, Lee et al. 2008; Aggarwal and Zhai 2012). This task has found widespread applications across various domains, including customer service (Bagheri, Saraee, and De Jong 2013; Barik and Misra 2024), online content moderation (Hettiachchi and Goncalves 2019; Risch and Krestel 2020), and large language model pre-training (Sun et al. 2023; Miah et al. 2024).

A widely used method (Clark, Yatskar, and Zettlemoyer 2019; He, Zha, and Wang 2021; Wang, Shu, and Culotta 2021) in sentiment classification relies on determining the correlation between specific words to sentiment of a sentence. These approaches begin by identifying the top words that exhibit strong correlations with sentiment labels, then these top words are utilized to assist in classifying the sentiment of the entire sentence.

An important challenge in these method is the spurious correlations, for example, in IMDB dataset (Pang and Lee 2005), the term *Spielberg* is frequently associated with positively reviewed movies despite its inherently neutral semantic meaning, resulting in spurious correlations between word *Spielberg* and positive sentiment. Meanwhile, some words are causally related to the sentiment, such as the word *disappointed* to the negative sentiment or the word *happy* to the positive sentiment.

Various prior works have been proposed to improve the robustness of sentiment classifiers by addressing spurious correlations.Techniques such as masked reconstruction (Moon et al. 2021) and contrastive learning (Choi et al. 2022) have demonstrated effectiveness in out-of-distribution scenarios but face challenges in accurately identifying causal keywords. Feature selection approaches (Wang and Culotta 2020) and counterfactual data augmentation (Wang and Culotta 2021) effectively reduce spurious associations but exhibit limited generalization capabilities.

Although previous research has made significant progress in identifying the treatment effects of words, most existing methods simplify the task by modeling the presence of a specific word in a sentence as a binary treatment, i.e., 1 if the word in this sentence, and 0 otherwise. However, for low-frequency words in the dataset, almost treatments are labeled as 0, resulting in inaccurate treatment effect estimation, then leading to suboptimal performance in causal word classification and sentiment classification.

To build a more robust and generalizable sentiment classifier, we propose a method that regards word representations as treatments, enhancing the robustness and generalization of treatment effect estimation model, especially for the low-frequency words. Specifically, our approach consists of three main steps. First, we perform clustering based on the word representations extracted by a pre-trained language model and utilize the integral probability metric (IPM) distance to learn the balanced representation of the context between each cluster. Next, we train a multi-head treatment effects estima-

*Corresponding author.

tion model based on the word cluster to identify causal and spurious words, which can enhance the model robustness and generalization. Finally, we apply the balanced covariate and learned treatment effects to train a classification model.

In conclusion, our main contributions are:

- We adopt the the embedding of the words as treatment and propose a causal representation learning method to train a robust and generalizable sentiment classifier.
- We first cluster words by their representations and estimate their treatment effects based on clusters using a multi-head classifier. In addition, we use IPM distance to learn a balanced context representations and use the learned treatment effects and representation to train a sentiment classifier.
- Extensive experiments on two publicly available datasets demonstrate the effectiveness of our method.

Related Work

A lot of previous methods focused on enhancing sentence sentiment classifier robustness by addressing spurious correlations. Early studies (Paul 2017; Wood-Doughty, Shpitser, and Dredze 2018) introduced causal inference to identify robust features by learning the causal relationships between words and sentence categories. Wang and Culotta (2020) used feature selection to distinguish spurious from genuine correlations, but fine-grained token-level semantics were not captured. MASKER (Moon et al. 2021) improved out-of-distribution detection with masked reconstruction but suffered from errors in attention-based keyword retrieval. To mitigate spurious correlations, some studies (Wang and Culotta 2021; Du et al. 2021; Garg et al. 2019; Kaushik, Hovy, and Lipton 2020; Khashabi, Khot, and Sabharwal 2020) employ counterfactual causal inference by introducing specific perturbations to the original samples, assessing whether the altered positions serve as causal features through observed changes in outcomes. Contrastive learning frameworks like C2L (Choi et al. 2022) and spurious word masking (Wang et al. 2021) showed promise but remained limited by causal keyword inaccuracies. Chew et al. (2023) introduced a regularization method to reduce spurious clusters without auxiliary data but lacked interpretability. Song et al. (2025) adopt a bifurcated approach by separately considering spurious and causal features, mining them through two distinct, interpretable methodologies.

However, these methods often overlooked the limitation of inaccurate and non-robust estimation of treatment effects for words, particularly for low-frequency words. To fill this gap, we propose a novel method that clusters words and considers their representations as treatments, significantly enhancing the robustness in detecting spurious correlations and improving performance in sentence sentiment classification.

Preliminary

Sentence Sentiment Classification

This paper focuses on the task of sentence sentiment classification, formulating it as a binary classification problem. Specifically, the dataset is composed of a set of labeled sentence samples $D_s = \{(s_1, y_1^s), \ldots, (s_N, y_N^s)\}$, where N represents the total number of sentences in D_s. Label $y_i^s \in \{0, 1\}$, where $y_i^s = 1$ indicates positive sentiment, and $y_i^s = 0$ indicates negative sentiment.

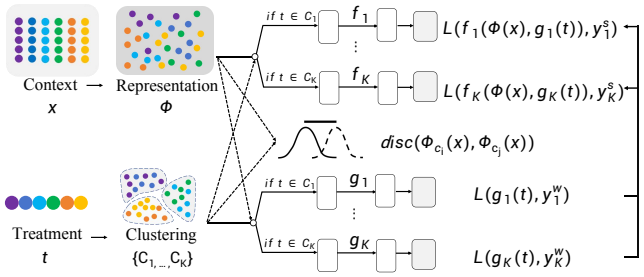

Figure 1: Architecture for sentiment classification: colored points represent treatments (words) and their contexts (sentences). Multi-head classifiers f and g handle sentiment and causal word classification, respectively, with L as the loss function. During training, only one head is updated per sample to focus on either sentiment or causal word classification.

Top Words Selection

Given a large number of words and some meaningless words such as "am" and "and", estimating the causal effect for every word is inefficient and unnecessary. Therefore, we first identify the words with strong correlations to the label, referred to as "top words". Specifically, following Wang and Culotta (2020), we use a bag-of-words model on each sentence s to obtain a word frequency vector x_s for the sentence. Then we define a logistic regression model to classify the sentiment of the sentence s, as follows: $h(x_s; \theta) = \frac{1}{1+e^{-\langle x_s, \theta \rangle}}$. Words with higher absolute coefficients in θ are considered more strongly correlated with the sentiment of the sentence. We are interested in the M words with the largest absolute coefficient, denoted as $\mathcal{W} = \{w_1, \ldots, w_M\}$. Following the setting in Wang and Culotta (2020). There is a labeled set $\{y_1^w, \ldots, y_M^w\}$. The label $y_i^w = 1$ means the word is a causal word, and $y_i^w = 0$ means a spurious word.

Causal Word Classification

For all top words, previous work (Wang and Culotta 2020) regards whether the words in a sentence as the binary treatment. However, due to the limited number of positive samples, the treatment effect estimation will not be accurate, especially for the low-frequency words, leading to struggles to identify spurious correlations and sub-optimal performance in sentence-level sentiment classification. In this paper, we treat the word embeddings as the treatment and the context embeddings as the covariate. Specifically, we initialize the representations using BERT (Kenton and Toutanova 2019). The sentence representation $E_s \in \mathbb{R}^{d_s}$ is obtained by concatenating the outputs from the last four layers of BERT, while the word representation $E_w \in \mathbb{R}^{d_w}$ is directly extracted from the corresponding layer, where d_s and d_w denote the dimensions of the context and word embeddings, respectively.

Methodology

To enhance the robustness and generalization of sentence sentiment classification, we propose a novel causal representation learning method, which is illustrated in Figure 1. First, top words are clustered based on their embedding. Then, we train a multi-head classifier to estimate the treatment effect of each word based on the word cluster on sentence sentiment to identify causal or spurious top words. Meanwhile, we learn a balanced context representation based on the IPM distance between each cluster. Then, the treatment effects derived from the word classifiers and the balanced representation are regarded as input to train a robust and generalizable sentiment classifier. Next, we explore each component of our method in detail.

Treatment Clustering & Context Representation Learning

To accurately estimate the treatment effect of each top word, especially for low-frequency words, we perform clustering based on the word embedding E_w of the top words. Using the representations of all top words, we group the words into K clusters using clustering algorithms such as K-means or Hierarchical Clustering (Cohen-Addad et al. 2019).

For K word clusters, we can obtain the corresponding K context clusters as well. To ensure balancing covariate, we learn the context representation $\Phi(\cdot)$ using the following IPM distance:

$$\text{IPM}_{\mathcal{Z}}(p_i, p_j) := \sup_{z \in \mathcal{Z}} \left| \int_S z(\Phi(E_s)) \big(p_i(\Phi(E_s)) - p_j(\Phi(E_s))\big)\, ds \right|, \tag{1}$$

where $p_i(\cdot)$ and $p_j(\cdot)$ represent the distribution of the context distribution in cluster i and j. IPM is always symmetric and obeys the triangle inequality, and trivially satisfies $\text{IPM}_{\mathcal{Z}}(p, p) = 0$. For rich enough function families $\mathcal{Z}$, we also have that $\text{IPM}_{\mathcal{Z}}(p, q) = 0 \implies p = q$, and then $\text{IPM}_{\mathcal{Z}}$ is a true metric over the corresponding set of probabilities.

Until now, we have obtained K clusters of top words and contexts with the balanced representation $\Phi(E_s)$.

Identify Causal/Spurious Relations of Top Words

To estimate the average treatment effect (ATE) of top words and identify whether a word is spurious, we define a multi-head word classifier, denoted as $\{g_1, \ldots, g_K\}$ for corresponding clusters, where $g_i \in \mathbb{R}^{d_w} \to \mathbb{R}^1$ is trained using manually labeled spurious label. Given a word w, we use its representation to determine the cluster C_k to which it belongs. By matching the word to its corresponding cluster, the classifier can effectively handle low-frequency words. Using the function g, we estimate the ATE τ_w as:

$$\hat{\tau}_w = E[\hat{Y}(E_w) - \hat{Y}(0)] = g(E_w) - g(0), \tag{2}$$

where $\hat{Y}(E_w)$ is the predicted sentiment outcome when the word embedding E_w is as treatment, $\hat{Y}(0)$ is the predicted outcome in the absence of treatment, i.e., replace the treatment words with the <blank> token, and $g(E_w) - g(0)$ is the word classifier output, which represents the estimated treatment effect.

Algorithm 1: Identifying spurious correlations for text sentiment classification

Input: training data $\mathcal{D}_s = \{(s_1, y_1^s), \ldots, (s_N, y_N^s)\}$

Extract words that are most strongly indicative of the sentence's sentiment;

Manually label the spurious words to create a dataset: $\mathcal{D}_w = \{(w_1, y_1^w), \ldots, (w_M, y_M^w)\}$;

Group the top words into multiple clusters;

for $w_m \in \{w_1, \ldots, w_M\}$ **do**

  Compute the cluster k of w_m;

  Update the word classifier: $\theta_g \leftarrow \theta_g - \eta \nabla_{\theta_g} L(g_k(E_{w_m}), y_m^w)$;

  Estimate the treatment effect: $\hat{\tau}_w = E[\hat{Y}(E_{w_m}) - \hat{Y}(0)] = g_k(E_{w_m}) - g(0)$;

end

for $s_n \in \{s_1, \ldots, s_N\}$ **do**

  Compute the cluster k of w_m in s_n;

  Update the sentence representation: $\theta_\Phi \leftarrow \theta_\Phi - \eta \nabla_{\theta_\Phi} L(f_k(\Phi(E_{s_n}), g_k(E_{w_m})), y_n^s)$;

  Update the sentence classifier: $\theta_f \leftarrow \theta_f - \eta \nabla_{\theta_f} L(f_k(\Phi(E_{s_n}), g_k(E_{w_m})), y_n^s)$;

end

Output: robust transferable word classifier $g(E_{w_m})$, and robust transferable text classifier $f(\Phi(E_{s_n}), g(E_{w_m}))$

Sentence Sentiment Classification

Given a sentence s, we first apply the representation function $\Phi : \mathcal{X} \to \mathcal{R} \in \mathbb{R}^{d_s}$ to map the sentence representation from the original space $\mathcal{X}$ to the balanced space $\mathcal{R} \in \mathbb{R}^{d_s}$.

Next, based on the top word w in s and the context corresponding to the top word, we determine the cluster C_k to which it belongs. Similar to word classification, we define a K-head classifier $\{f_1, \ldots, f_K\}$ for sentence classification. Using the treatment effect of the top word on the sentence, learned in the previous stage, we compute the final prediction $\hat{y}$ as follows:

$$\hat{y} = f_k(\Phi(E_s) \cdot g_k(E_w)) + g_k(E_w). \tag{3}$$

We add $g_k(E_w)$ to ensure the monotonicity, i.e., the larger the absolute value of $g_k(E_w)$, the greater the effect of the specific word on the sentiment of the entire sentence.

The outline of the algorithm is presented in Algorithm 1.

Experiments

Experimental Setup

Following previous work (Wang and Culotta 2020), we experiment with two datasets for sentiment classification and causal word classification.

IMDB Movie Reviews: This dataset is a sampled subset of the original IMDB dataset (Pang and Lee 2005), collected and published by Kaushik, Hovy, and Lipton (2020). Each document in this dataset is a long paragraph that has historically been used for sentiment classification tasks. We provide

Dataset	IMDB		Kindle	
	pos	*neg*	*pos*	*neg*
Sentences	4411	4471	14824	14606
Top words	366		270	
Spurious words	145		154	
Causal words	221		116	

Table 1: Statistics of the datasets.

a version of the dataset where paragraphs are split into multiple single sentences, each labeled with their overall sentiment polarity (positive or negative).

Amazon Kindle Reviews: This dataset contains book reviews from the Amazon Kindle Store, with ratings ranging from 1 to 5 (He and McAuley 2016). We label reviews with ratings of {4, 5} as positive and {1, 2} as negative, excluding those with a rating of 3. The dataset is then processed into single sentences using the same method as applied in the IMDB dataset.

Causal Words Classification: We utilize manually annotated causal word data from Wang and Culotta (2020). Specifically, a word was identified as a causal word if, all else being equal, it was deemed a determining factor for the sentiment polarity of a sentence. This dataset was annotated by two student annotators, who labeled all top words as either causal or spurious. While manual annotation inherently involves a degree of subjectivity, the annotation consistency for this task was generally high, with a raw agreement rate of 96%. It is worth noting that the annotated words are specific to the dataset's domain, and annotations are not typically transferable to new datasets.

The detailed statistics of two datasets and manually labeled ground truth are shown in the Table 1.

Baseline: Wang and Culotta (2020) estimates the ATE of a word based on the covariate matching algorithm. A word classifier is then trained to identify spurious words and remove spurious words during sentiment classification model training.

Experimental Scenarios

On the one hand, based on our frequency analysis of all top words, 50% of the low-frequency words account for 22.06% of occurrences in the sentence dataset, while 10% of the high-frequency words account for 38.04% of occurrences. These low-frequency words have only a small number of positive samples in the training set, making it difficult to perform accurate and robust treatment effect estimation. To evaluate whether the method can still perform effectively under low-frequency scenarios, we categorized the words based on their frequency in the sentences and provided experimental settings for high frequency, low frequency, and all samples.

- **High frequency**: Words with frequencies below the 50th percentile in the frequency distribution.
- **Low frequency**: Words with frequencies above the 50th percentile in the frequency distribution.
- **All samples**: Includes all top words.

IMDB			
	High freq.	Low freq.	All
baseline	0.7983	0.7188	0.7470
baseline (trans.)	0.7852	0.7204	0.7429
ours	0.7929	0.8482	0.7933
ours (trans.)	0.7847	0.8968	0.7545
Kindle			
	High freq.	Low freq.	All
baseline	0.7362	0.7471	0.7440
baseline (trans.)	0.8360	0.7562	0.7732
ours	0.7785	0.9129	0.7797
ours (trans.)	0.8191	0.8346	0.8198

Table 2: Performance of the **sentence sentiment classifier** (AUC score) under different frequency distributions in same-domain and transfer-domain scenarios.

On the other hand, considering that manually annotating causal words requires significant effort and that causal words labeled in one dataset cannot be directly transferred to another dataset, we aim to reduce the annotation burden. Specifically, we hope that a word classifier trained in one domain can also be effectively applied to another domain. To this end, we explored the performance of the word classifier in both same-domain and transfer-domain scenarios.

- **Same domain**: We used 5-fold cross-validation to evaluate the accuracy of the word classifier within the same domain.
- **Transfer domain**: We measured cross-domain accuracy by, for instance, training the word classifier on the IMDB dataset and subsequently evaluating its performance on the Kindle dataset to assess its generalization ability.

Sentence Sentiment Classification

Table 2 presents the AUC scores for causal word classification. Our method outperforms the baseline and exhibits consistent trends across different datasets. Specifically, using the IMDB dataset as an example: When the training domain and testing domain are the same, our method achieves a 6.20% improvement over the baseline on all samples and significantly outperforms the baseline on low-frequency words, with a remarkable increase of 13.53%. When the classifier trained on the source domain is transferred to the target domain for testing, our method shows a 1.56% improvement on all samples and a 22.19% improvement on low-frequency words compared to the baseline.

The baseline method models sentiment classification as a binary treatment problem, where the presence of a top word in a sentence is denoted as $Y(1)$, and its absence as $Y(0)$. However, this approach fails to generalize to words that appear infrequently or do not appear at all in the training set, making it incapable of accurately estimating treatment effects for low-frequency words. In contrast, our method uses the representation of words as the treatment. By assigning clusters to all top words, our approach enables precise treatment

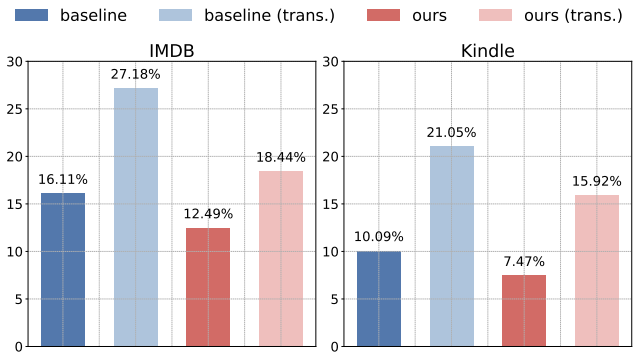

Figure 2: In both same domain and transfer domain scenarios, we measure the percentage decrease in the **causal word classifier**'s performance (AUC score) for low-frequency words relative to all samples. A lower percentage indicates better robustness.

effect estimation and classification, even for low-frequency words. By utilizing representations as treatments and modeling clusters for different words, our approach ensures superior performance and enhanced robustness, particularly in transfer-domain scenarios.

Causal Word Classification

As shown in Table 2, we measure the performance decrease of the causal word classifier on low-frequency words relative to the all-sample scenario. We found that on both datasets, our method exhibits significantly less performance decrease than the baseline does. This indicates that our proposed ITE estimation method can effectively address the issue of inaccurately estimating a word's treatment effect when the sample size is limited, thereby enhancing the robustness of causal word classification.

Conclusion and Future Work

This paper discovers that existing methods for sentiment classification tasks aimed at identifying word with spurious correlations fail to generalize to words that appear infrequently in the training set. To address this limitation, we propose a new spurious word identification and sentiment classification model. By clustering words and using word representations as treatments, we achieve more accurate and robust treatment effect estimation for low-frequency words, ultimately enhancing sentiment classification. We evaluate our method on two text datasets, and it outperforms the baseline in both low-frequency and all-sample scenarios.

In future work, we will further enhance the ability to identify causal words and explore how causal features influence the learning and representation of contexts in large language models (LLMs). Additionally, we aim to investigate methods based on causal inference to identify and mitigate the impact of spurious correlations on LLMs.

References

Aggarwal, C. C.; and Zhai, C. 2012. A survey of text classification algorithms. *Mining text data*, 163–222.

Bagheri, A.; Saraee, M.; and De Jong, F. 2013. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, 52: 201–213.

Barik, K.; and Misra, S. 2024. Analysis of customer reviews with an improved VADER lexicon classifier. *Journal of Big Data*, 11(1): 10.

Chew, O.; Lin, H.-T.; Chang, K.-W.; and Huang, K.-H. 2023. Understanding and Mitigating Spurious Correlations in Text Classification with Neighborhood Analysis. *arXiv preprint arXiv:2305.13654*.

Choi, S.; Jeong, M.; Han, H.; and Hwang, S.-w. 2022. C2l: Causally contrastive learning for robust text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10526–10534.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. 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)*, 4069–4082.

Cohen-Addad, V.; Kanade, V.; Mallmann-Trenn, F.; and Mathieu, C. 2019. Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM (JACM)*, 66(4): 1–42.

Du, M.; Manjunatha, V.; Jain, R.; Deshpande, R.; Dernoncourt, F.; Gu, J.; Sun, T.; and Hu, X. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. *arXiv preprint arXiv:2103.06922*.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 219–226.

He, H.; Zha, S.; and Wang, H. 2021. Unlearn dataset bias in natural language inference by fitting the residual. In *2nd Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing, DeepLo@ EMNLP-IJCNLP 2019*, 132–142. Association for Computational Linguistics (ACL).

He, R.; and McAuley, J. 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*, 507–517.

Hettiachchi, D.; and Goncalves, J. 2019. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian conference on human-computer-interaction*, 342–346.

Kaushik, D.; Hovy, E.; and Lipton, Z. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language

understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.

Khashabi, D.; Khot, T.; and Sabharwal, A. 2020. More Bang for Your Buck: Natural Perturbation for Robust Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 163–170.

Miah, M. S. U.; Kabir, M. M.; Sarwar, T. B.; Safran, M.; Alfarhood, S.; and Mridha, M. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, 14(1): 9603.

Moon, S. J.; Mo, S.; Lee, K.; Lee, J.; and Shin, J. 2021. Masker: Masked keyword regularization for reliable text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13578–13586.

Pang, B.; and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2): 1–135.

Paul, M. 2017. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 163–172.

Risch, J.; and Krestel, R. 2020. Toxic comment detection in online discussions. *Deep learning-based approaches for sentiment analysis*, 85–109.

Song, R.; Li, Y.; Tian, M.; Wang, H.; Giunchiglia, F.; and Xu, H. 2025. Causal keyword driven reliable text classification with large language model feedback. *Information Processing & Management*, 62(2): 103964.

Sun, X.; Li, X.; Zhang, S.; Wang, S.; Wu, F.; Li, J.; Zhang, T.; and Wang, G. 2023. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Wang, T.; Sridhar, R.; Yang, D.; and Wang, X. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Wang, Z.; and Culotta, A. 2020. Identifying spurious correlations for robust text classification. *arXiv preprint arXiv:2010.02458*.

Wang, Z.; and Culotta, A. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14024–14031.

Wang, Z.; Shu, K.; and Culotta, A. 2021. Enhancing Model Robustness and Fairness with Causality: A Regularization Approach. In *Proceedings of the First Workshop on Causal Inference and NLP*, 33–43.

Wood-Doughty, Z.; Shpitser, I.; and Dredze, M. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, 4586. NIH Public Access.