Text Embedding as Treatment: A Meta Causal Approach for Robust Sentiment Classification

Fengxiang Cheng[1,2], Chuan Zhou[3,2], Xiang Li[4], Haoxuan Li[4], Wenli Wang[5], Jinkun Chen[6], Mingming Gong[3,2,*], Kun Zhang[7,2,*]

[1]University of Amsterdam, [2]MBZUAI, [3]The University of Melbourne, [4]Peking University, [5]Renmin University of China, [6]Dalhousie University, [7]CMU

Correspondence: mingming.gong@unimelb.edu.au, kunz1@cmu.edu

Abstract

Sentiment classification is a crucial task in natural language processing (NLP). To mitigate the spurious correlation, the causal word identification method estimates the impact of treatment words on sentence sentiment and removes those with low treatment effects. However, previous works regard the presence or absence of a specific word in a sentence as a binary treatment. This approach limits the generalizability to novel words and the robustness of low-frequency words. To bridge this gap, we propose a meta-causal approach that achieves causal word identification for arbitrary words with a single training task. Specifically, we begin by clustering contexts based on their embeddings obtained from a pre-trained language model. Subsequently, for each cluster, a representation and multi-head prediction networks are trained to estimate the treatment effect of each word to distinguish causally related words from spuriously correlated ones. The trained word classifier is then used to give weights for different words 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.

1 Introduction

Sentiment classification, a fundamental NLP task aiming to determine sentiment polarity (Pang et al., 2008; Aggarwal and Zhai, 2012), is widely applied in customer service (Bagheri et al., 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). While many methods identify sentiment by correlating keywords with labels (Clark et al., 2019; He et al., 2021; Wang et al., 2021), they are susceptible to learning spurious, non-causal correlations (Kong et al., 2024; Cheng et al., 2025a,b; Wang et al., 2026a,b). For instance, in FineFood review dataset (McAuley and Leskovec, 2013), the word *coffee* might become strongly associated with positive sentiment due to its frequent appearance in favorable reviews.

To enhance the robustness of sentiment classification, an important class of methods involves identifying the causal relationships between words in sentences and sentiment classification (Wang et al., 2022; Choi et al., 2022), followed by removing spuriously correlated words before performing classification, as Step 1 and Step 2 in Figure 1 shows. Such methods aim to address the counterfactual question: For a given sentence, how would the probability of the sentiment being positive change when a keyword were *present* as compared to when it were *absent*? In practice, for each keyword selected as treatment, an independent binary treatment dataset is constructed: sentences containing the keyword serve as the treatment group, while those without it act as the control group. Datasets (and their corresponding counterfactual tasks) for different keywords are mutually independent.

Unfortunately, such methods have two crucial limitations: (i) They cannot generalize to novel treatments due to the lack of sample sentences for constructing binary treatment datasets. (ii) For existing treatments with low occurrence frequencies, the identification performance is poor due to the unbalanced sample distribution between the treatment and control groups.

To fill these gaps, we propose a meta-causal approach that focuses on a different counterfactual question: "If a keyword were intervened and replaced with any other word, how much would the probability of the sentiment being classified as positive change?". Methodologically, we treat text embeddings as the treatment variable. Our approach is meta-learning based, meaning that only one dataset

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 41309–41320
July 2-7, 2026

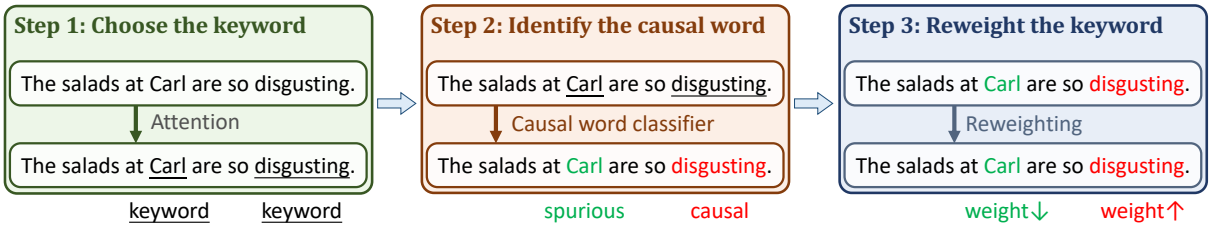


Figure 1: Framework Overview. A three-step process consists of selecting keywords, identifying the causal words, and reweighting the keywords in the sentence sentiment classifier.

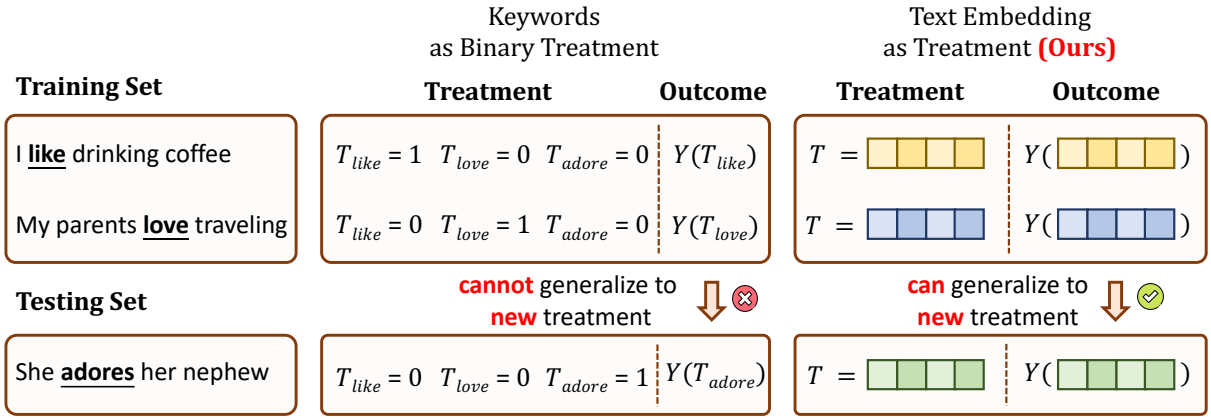


Figure 2: Illustration of the motivation for our proposed method.

(and correspondingly, one task) is required to generalize to the causal relationship identification tasks between novel keywords and sentiment classification, as illustrated in Figure 2.

In our meta-causal learning approach, a model is built for simultaneous sentiment classification and causal word identification. The specific methods include: (1) clustering the contexts after keyword removal using embeddings of a pre-trained language model (PLM); (2) computing the conditional average causal effect (CATE) as a causal word probability estimation; and (3) increasing the weights of causal word and decreasing the weights of spurious word in subsequent sentiment classifier training, instead of hard removal of the spurious words. The approach, therefore, enhances the accuracy and robustness of sentiment classification. The contributions are shown below:

- To the best of our knowledge, this is the first work that achieves meta learning in causally related words identification, thus the classifier is generalizable to novel treatment words.
- We propose a framework using the continuous embedding of words as treatment, and addressing the imbalance treatment problems via representation balancing.
- We conduct extensive experiments to validate the effectiveness of our method for both in-sample and out-of-sample generalization.

2 Related Work

Keyword-Based Sentiment Classification. Previous methods enhance the robustness of sentence sentiment classification by addressing spurious correlations. Earlier studies (Paul, 2017; Wood-Doughty et al., 2018) utilized causal inference to identify robust features. Wang et al. (2022) distinguish between spurious and genuine correlations through feature selection, but does not capture fine-grained semantics. MASKER (Moon et al., 2021) improves out-of-distribution detection through mask reconstruction, but suffers from errors in attention-based keyword retrieval. Causal inference methods (Garg et al., 2019; Kaushik et al., 2020; Khashabi et al., 2020; Wang and Culotta, 2021) evaluate feature causality through sample perturbation. Contrastive learning frameworks C2L (Choi et al., 2022) and masking methods (Wang et al., 2022) are limited by causal keyword accuracy. Chew et al. (2024) proposes regularization methods without auxiliary data but lacks interpretability. Song et al. (2025) uses a bifurcation framework to mine spurious and causal fea-

tures separately. However, these methods often overlooked the limitation of inaccurate and non-robust estimation of treatment effects for words, particularly for low-frequency words.

Causal Effect Estimation. Estimating causal effects in observational studies is challenging due to feature imbalance and confounding bias (Zhang et al., 2024; Zheng et al., 2025; Zhou et al., 2025). Previous methods include propensity-score-based approaches (Rosenbaum and Rubin, 1983) for matching (Dehejia and Wahba, 2002), reweighting (Hirano et al., 2003), and doubly robust estimation (Robins et al., 2000); confounder balancing techniques to align context distributions (Hainmueller, 2012; Athey et al., 2018); and deep representation learning models like TARNet (Shalit et al., 2017), BNN (Johansson et al., 2016), CFR (Shalit et al., 2017), Dragonnet (Shi et al., 2019), DRCFR (Hassanpour and Greiner, 2019), ES-CFR (Wang et al., 2024), and CFR-DF (Zheng et al., 2026). Besides, generative methods, such as CEVAE (Louizos et al., 2017), TEDVAE (Zhang et al., 2021), and GANITE (Yoon et al., 2018), use generative models to alleviate confounding bias by simulating the data generation process. Despite these methods mitigating confounding issue, their extension to causal effect estimation with continuous treatment in non-trivial (Zhu et al., 2024, 2025), and high-dimensional continuous treatment in NLP community remains rarely studied.

3 Preliminary

3.1 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 $\mathcal{D}_s = \{(s_1, y_1^s), \ldots, (s_N, y_N^s)\}$, where N represents the total number of sentences in $\mathcal{D}$. Label $y_i^s \in \{0, 1\}$, where $y_i^s = 1$ indicates positive sentiment, and $y_i^s = 0$ indicates negative sentiment. Our ultimate goal is a classifier that predicts the sentiment labels of input sentences.

3.2 Keywords Selection

Previous approaches propose to find causal words in a sentence based on the magnitude of causal effect of each word on the sentiment label. However, computing the causal effect of each word one by one is inefficient and meaningless, so keywords that are strongly correlated to the label are often sought first as a candidate set of causal words.

Specifically, most of existing work follows Wang et al. (2022), using a bag-of-words model on each sentence s to obtain a word frequency vector x_s for the sentence. Then 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. The L words with the largest absolute coefficient are the keywords, denoted as $\mathcal{T} = \{t_1, \ldots, t_L\}$. The keywords are then labeled as $\{y^{t_1}, \ldots, y^{t_L}\}$. The label $y^{t_i} = 1$ means the word t_i is a causal word, while $y^{t_i} = 0$ means t_i is a spurious word.

3.3 Causal Term Definition

Treatment: In causal inference, this denotes the intervention variable. Here, keywords serve as treatments ("Carl"/"disgusting").
Covariate: In causal inference, covariate refer to variables that may influence the outcome but are not directly treated as treatment variables. In this paper, covariate refers to other parts of the sentence apart from the keywords, such as "The salads at [MASK] are so [MASK]".
CATE: In causal inference, CATE refers to a statistical concept that estimates the average causal effect of a treatment on an outcome variable. For the causal word "disgusting", we observe significantly higher CATE than for "Carl".

3.4 Causal Word Classification

For all keywords, previous work (Wang et al., 2022) regards the presence of words in a sentence as a binary treatment. It estimates the average causal effect (ATE) by computing the difference in sentence sentiment labels with and without each word. Words whose absolute ATE exceeds a given threshold are classified as causal, while those below the threshold are considered spurious.

4 Methodology

In this paper, we propose a three-stage framework for robust sentiment classification, as illustrated in Figure 1. First we select candidate keywords using PLMs and label them via an LLM. Then we design a novel causal word identification model, which is trained with a twofold representation balancing strategy to accurately estimate the causal effects of words on sentiment. Finally, we use these estimated

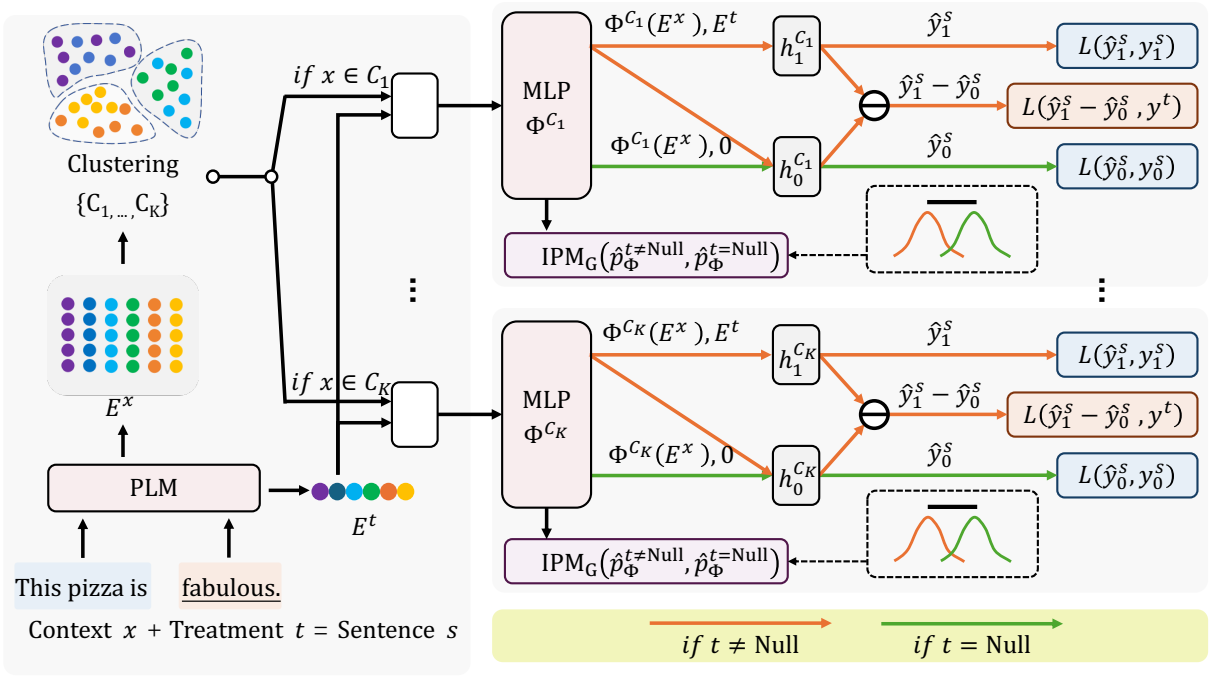


Figure 3: Model designation for Step 2 of our proposed framework, including clustering of contexts, representation and prediction networks of each cluster.

effects to dynamically reweight keywords, guiding the final classifier to focus on genuinely causal features over spurious ones.

Step 1 (Keywords Selection). Select keywords using the attention scores of transformer-based PLMs and label the causal words in the keyword set using large language models (LLMs) GPT-o1.

Step 2 (Causal Word Identification). Train a robust causal word classifier by balancing covariate representations twice and multi-task learning for sentence sentiment classification and word classification using a representation learning framework.

Step 3 (Keyword Reweighting). Training a robust sentence sentiment classifier by increasing the weight of causal words and decreasing the weight of causal words.

Notice that the framework improves the accuracy of rare treatment word prediction and the robustness of the model through the continuous representation of treatment words in Step 2 and the dynamic weight adjustment mechanism in Step 3.

4.1 Keywords Selection

First, we train a sentence sentiment classifier on the dataset $\mathcal{D}_s$ using the transformer-based pre-training language model BERT (Devlin et al., 2019). Then we calculate the attention of each word in all the sentences one by one and take the average value, and select the L words with the highest average attention as keywords $\{t_1, t_2, \dots, t_L\}$. After obtaining the keywords, we use the state-of-the-art large language model GPT o1 (Jaech et al., 2024) to label them as causal or spurious words, obtaining labels $\{y^{t_1}, \dots y^{t_L}\}$.

4.2 Causal Word Identification

We will use the data labeled by LLM to train a generalizable causal word classifier. We learn sentence sentiment classification at the same time as we train the word classifier, making sure that the causal words learned are particularly for sentiment classification. The model designation for the word classifier g in Step 2 is shown in Figure 3.

(1) Sentence Splitting. Firstly, if there is a keyword t in a sentence s, we treat it as the treatment word and replace it with a blank token [MASK] to get the context as the covariate x in the causal framework, i.e., $x = \text{mask}(s, t)$. Otherwise, we define the treatment variable as $Null$ and the covariate x as the sentence itself.

(2) Representation Extraction. We then used a pre-trained BERT to compute the embeddings for context x and keyword t as E^x and E^t respectively.

(3) Clustering-based Representation Balancing. Considering the semantic heterogeneity of different contexts x's, we cluster them based on embeddings E^x's, aiming to make contexts in each cluster C_k have similar semantics, such that the causal effect

of each treatment word on a specific cluster is homogeneous. Subsequently, we design separate representation network Φ^{C_k} and prediction networks for each cluster.

(4) IPM-based Representation Balancing. For contexts that pass through the same representation network Φ, we classify them into two groups based on whether their corresponding treatment variables are $Null$ or not. Inspired by representation learning, we minimize the integral probability metric (IPM) distance (Frogner et al., 2015) of the empirical distribution of representations between the two groups, which is defined below:

$$\mathcal{L}_{imb} = \text{IPM}_{\mathcal{G}}(\hat{p}_{\Phi}^{t=Null}, \hat{p}_{\Phi}^{t\neq Null}). \quad (1)$$

By introducing IPM distance, we aim to obtain a balanced representation that can generalize well across different groups, which is defined below:

$$\text{IPM}_G(p,q) := \sup_{g\in G} \left| \int_S g(s)(p(s)-q(s))\, ds \right|,$$

where p and q are two probability distributions, and G is a class of functions for which we seek to optimize the difference in expectations. In our scenario, p and q are the distributions on groups with/without top words.

(5) CATE Calculating: When $t = Null$ (**green path**), the representation $\Phi(E^x)$ is concatenated with a zero vector into the prediction header h_0. On the contrary, if $t \neq Null$ (**orange path**), the representation $\Phi(E^x)$ is concatenated with E^t into h_1, and with the zero vector into h_0, respectively. The outputs of h_1 and h_0, i.e., $\hat{y}_1^s$ and $\hat{y}_0^s$ are the estimates for sentiment labels of s with and without t respectively. Here we denote the sentiment classification loss as $\mathcal{L}_s$, corresponding to the **blue loss block** in Figure 3. Let y_0^s be the sentiment label for control group, and y_1^s be the sentiment label for treatment group, we define the classification loss:

$$\mathcal{L}_s = \mathbb{1}_{t=Null} L(\hat{y}_0^s, y_0^s) + \mathbb{1}_{t\neq Null} L(\hat{y}_1^s, y_1^s), \quad (2)$$

where $\mathbb{1}_{\{\cdot\}}$ is indicator function, and $L(\cdot,\cdot)$ is a loss function such as cross-entropy. We take the difference in the two estimates as the CATE estimate. Typically, the larger the CATE, the greater the probability that this treatment variable t is a causal word. Therefore, we denote the causal word loss as $\mathcal{L}_t$, corresponding to the **orange loss block** in Figure 3:

$$\mathcal{L}_t = \mathbb{1}_{t\neq Null} L(\sigma(\hat{y}_1^s - \hat{y}_0^s), y^t), \quad (3)$$

where y^t represents the label for whether a keyword is a causal word and $\sigma(x) = 1/(1+e^{-x})$ is sigmoid function. In summary, we train a robust word classifier g by minimizing the following loss, which is a combination of the (1) imbalance loss $\mathcal{L}_{imb}$, (2) sentiment classification loss $\mathcal{L}_s$, and (3) causal word loss $\mathcal{L}_t$:

$$\mathcal{L}(\theta_g) = \mathcal{L}_{imb} + r_s\mathcal{L}_s + r_t\mathcal{L}_t, \quad (4)$$

where $r_s, r_t > 0$ are hyper-parameters. It is worth noting that our word classifier g takes not just a keyword t as input, but both the context x and the keyword t at the same time, due to the fact that we consider CATE, i.e., determining whether the keyword has a causal effect on the sentiment of the sentence while concerned with the contextual information.

4.3 Keyword Reweighting

Previous methods always perform a hard classification after obtaining the word classifier g, dividing the keyword set $\{t_1, \ldots, t_L\}$ into causal and spurious words, and then mask out the spurious words from the sentences in the training data $\mathcal{D}_s$, and using the remaining sentences and sentiment labels to train sentence sentiment classifiers. Instead, we propose to use the probability of a keyword being a causal word given by the word classifier g to adjust the weights of the keywords.

Specifically, we use BERT as the backbone. For each sentence s_i in the dataset $\mathcal{D}_s$, if there are keywords $\mathcal{T}_i = \{t_i^1, t_i^2, \ldots, t_i^J\} \subset \mathcal{T}$, we compute the corresponding contexts as $\{x_i^j | x_i^j = \text{mask}(s_i, t_i^j), j = 1, \ldots J\}$. The probability of a keyword t_i^j being a causal word considering the contexts x_i^j is given by the word classifier g as:

$$\tau_i^j = \sigma(\hat{y}_1^s - \hat{y}_0^s) \in (0,1), \quad (5)$$

where τ denotes the CATE of the keyword. We obtained τ using the trained keyword classifier. In order to increase the weight of the causal word, we monotonically map this probability to $(0,2)$ as the new weight of the word $w(x_i^j) = 2\tau_i^j$. On the other hand, for sentences s without any keywords, we set the weight of all words to one.

As shown in Step 3 of Figure 1, by doing so, we have increased the weight of causal words and decreased the weight of spurious words.

Dataset	Food	IMDB	SST-2
Samples	17,273	35,000	67,349
Positive samples	13,618	17,540	37,569
Negative samples	3,656	17,461	29,781

Table 1: Summary statistics of datasets.

Occurrence	1–10	10–100	100+
IMDB	**73.3%**	6.7%	20.0%
SST-2	**25.0%**	48.4%	26.6%
Food	**60.0%**	31.4%	5.7%

Table 2: Causal keyword occurrences in datasets.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct the sentiment analysis experiments on three widely-used datasets: FineFood (Food) (McAuley and Leskovec, 2013), IMDB movie reviews (IMDB) (Maas et al., 2011), and Stanford Sentiment Treebank (SST-2) (Socher et al., 2013). To further demonstrate the significance and the challenge of handling low-frequency causal words, we provide the summary statistics of datasets in Table 1 and details about the distribution of keyword occurrences across them in Table 2, showing that low-frequency causal words (occurring <10 times) constitute a substantial portion across all datasets.

Base Models. We use standard PLMs including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020) as the base models for obtaining the embedding of sentences and top words. In addition, for our method, we use the MLP as the base model for learning balanced representation.

Baselines. We consider the following baselines in the experiments: IPS (Saha et al., 2019), DR (Sridhar and Getoor, 2019), Matching (Wang et al., 2022), TARNet (Shalit et al., 2017), Masker (Moon et al., 2021), Causally Contrastive Learning (C2L) (Choi et al., 2022), NFL (Chew et al., 2024), SCC (Zhou et al., 2024), and CCR (Zhou and Zhu, 2025).

Implementation Details. We utilize a setup of 8 NVIDIA 3090 GPUs for parallel computing, supported by 300GB of RAM.

Model	Dataset	Accuracy	F1	Sensitivity	Specificity
BERT	IMDB	97.00%	93.75%	95.74%	97.39%
	SST-2	96.00%	90.00%	92.31%	96.89%
	Food	97.00%	93.02%	93.02%	98.09%
RoBERTa	IMDB	98.00%	94.44%	97.14%	98.18%
	SST-2	97.50%	95.12%	96.08%	97.99%
	Food	98.00%	92.31%	96.00%	98.29%
ALBERT	IMDB	98.00%	92.86%	100.00%	97.70%
	SST-2	99.00%	98.04%	100.00%	98.67%
	Food	97.50%	92.75%	94.12%	98.19%
Average	–	**97.56%**	**93.59%**	**96.05%**	**97.93%**

Table 3: Evaluation of LLM annotated causal words against human label on sentiment classification datasets.

Domain	Task	Accuracy	F1
Finance	RE	97.00%	95.50%
Medical	NLI	96.00%	93.75%
Legal	QA	97.00%	94.15%

Table 4: Evaluation of LLM annotated causal words against human label on other domains and NLP tasks.

5.2 Causal Words Annotation

We now elaborate on **Causal words labeling** by providing a detailed explanation of the (1) label generation process, (2) toy example for illustration, and (3) verification of the label generation.

5.2.1 Label Generation Process

We first select the most correlated words as keywords (Step 1), in which we annotate the causal words with the help of an LLM (Step 2 and Step 3), finally perform manual verification (Step 4).

- **Step 1 (Keyword Selection):** We first use sentiment classification **PLMs**, such as BERT, to select the words with the most strong correlation with sentiment labels as **keywords**.
- **Step 2 (Editing Prompt):** Then we edit the following prompt for each keyword:

  *"In the sentiment classification task, please judge whether the word **[keyword]** can be used as a basis for sentiment classification without any additional explanation. Answer 'yes' or 'no'."*

- **Step 3 (Generating Causal Label):** Afterwards, we will input the edited prompt into LLM (GPT-o1) to generate the causal label.
- **Step 4 (Manual Verification):** The causal labels are manually verified and corrected.

	Method	Dataset					
		Food → IMDB	**Food → SST-2**	**IMDB → Food**	**IMDB → SST-2**	**SST-2 → Food**	**SST-2 → IMDB**
BERT	Vanilla	76.90 ± 0.08	77.32 ± 0.08	84.09 ± 0.08	83.23 ± 0.09	82.17 ± 0.08	84.23 ± 0.08
	IPS	75.05 ± 0.08	75.22 ± 0.10	84.59 ± 0.08	83.59 ± 0.09	79.67 ± 0.08	82.01 ± 0.09
	Macthing	76.91 ± 0.12	76.40 ± 0.06	86.29 ± 0.11	85.68 ± 0.09	82.75 ± 0.28	83.95 ± 0.11
	DR	77.55 ± 0.09	76.01 ± 0.07	86.19 ± 0.09	85.44 ± 0.09	83.89 ± 0.26	82.19 ± 0.08
	Tarnet	77.89 ± 0.08	77.23 ± 0.10	86.33 ± 0.09	85.80 ± 0.10	82.44 ± 0.26	83.40 ± 0.09
	Masker	75.23 ± 0.08	76.65 ± 0.08	84.35 ± 0.08	82.72 ± 0.08	78.69 ± 0.08	82.63 ± 0.07
	C2L	77.34 ± 0.08	76.11 ± 0.07	86.94 ± 0.09	84.01 ± 0.07	82.49 ± 0.09	84.78 ± 0.09
	SCC	76.87 ± 0.11	76.60 ± 0.11	85.35 ± 0.07	84.01 ± 0.10	83.28 ± 0.08	85.94 ± 0.08
	NFL	77.34 ± 0.08	77.76 ± 0.07	86.20 ± 0.10	84.00 ± 0.06	83.16 ± 0.10	85.22 ± 0.09
	CCR	77.28 ± 0.08	76.72 ± 0.09	86.23 ± 0.10	83.77 ± 0.10	84.24 ± 0.08	84.10 ± 0.07
	Ours	**78.60*** ± 0.08**	**78.86*** ± 0.08**	**86.98 ± 0.09**	**86.67*** ± 0.07**	**84.28*** ± 0.09**	**86.61*** ± 0.08**
RoBERTa	Vanilla	85.47 ± 0.09	82.22 ± 0.09	90.64 ± 0.09	84.91 ± 0.09	86.47 ± 0.08	86.65 ± 0.09
	IPS	85.43 ± 0.09	81.31 ± 0.09	90.33 ± 0.11	85.45 ± 0.07	86.02 ± 0.09	85.89 ± 0.08
	Macthing	85.29 ± 0.12	81.04 ± 0.10	90.76 ± 0.09	85.75 ± 0.09	85.59 ± 0.09	86.10 ± 0.09
	DR	85.34 ± 0.08	81.37 ± 0.09	90.83 ± 0.09	85.38 ± 0.09	86.29 ± 0.08	86.47 ± 0.08
	Tarnet	85.58 ± 0.10	81.88 ± 0.10	89.21 ± 0.10	86.28 ± 0.08	86.17 ± 0.09	87.21 ± 0.08
	Masker	82.10 ± 0.08	81.98 ± 0.10	91.57 ± 0.10	84.14 ± 0.07	89.32 ± 0.08	86.80 ± 0.10
	C2L	85.75 ± 0.10	82.79 ± 0.07	92.18 ± 0.09	85.51 ± 0.09	87.29 ± 0.10	87.35 ± 0.07
	SCC	85.05 ± 0.11	81.73 ± 0.10	92.40 ± 0.08	85.27 ± 0.10	86.83 ± 0.08	87.18 ± 0.09
	NFL	86.08 ± 0.08	82.33 ± 0.09	92.41 ± 0.10	85.91 ± 0.10	87.46 ± 0.09	87.40 ± 0.09
	CCR	86.17 ± 0.09	81.39 ± 0.10	92.32 ± 0.08	85.98 ± 0.09	88.49 ± 0.10	87.02 ± 0.09
	Ours	**87.50*** ± 0.10**	**83.65*** ± 0.08**	**93.24*** ± 0.10**	**87.15*** ± 0.08**	**89.46 ± 0.10**	**88.89*** ± 0.08**
ALBERT	Vanilla	81.96 ± 0.09	83.41 ± 0.10	87.09 ± 0.09	85.17 ± 0.07	81.17 ± 0.08	84.09 ± 0.09
	IPS	80.66 ± 0.09	81.56 ± 0.11	84.87 ± 0.08	87.21 ± 0.08	82.68 ± 0.08	84.32 ± 0.09
	Macthing	82.35 ± 0.10	81.85 ± 0.12	81.04 ± 0.08	86.54 ± 0.08	83.22 ± 0.08	84.48 ± 0.09
	DR	80.92 ± 0.10	80.47 ± 0.10	86.40 ± 0.08	86.87 ± 0.09	83.42 ± 0.09	84.01 ± 0.09
	Tarnet	83.08 ± 0.09	82.35 ± 0.12	85.64 ± 0.09	86.82 ± 0.08	83.48 ± 0.07	85.21 ± 0.09
	Masker	82.00 ± 0.08	84.30 ± 0.08	87.06 ± 0.08	85.31 ± 0.09	81.46 ± 0.09	83.44 ± 0.09
	C2L	82.53 ± 0.09	84.18 ± 0.08	87.68 ± 0.08	85.43 ± 0.08	82.66 ± 0.08	84.21 ± 0.09
	SCC	82.80 ± 0.09	84.14 ± 0.11	87.39 ± 0.08	85.78 ± 0.07	82.25 ± 0.11	84.62 ± 0.10
	NFL	82.96 ± 0.08	83.77 ± 0.09	87.74 ± 0.09	85.70 ± 0.07	82.13 ± 0.09	84.81 ± 0.07
	CCR	82.73 ± 0.08	84.11 ± 0.13	87.75 ± 0.08	85.59 ± 0.09	83.73 ± 0.10	84.11 ± 0.08
	Ours	**83.71*** ± 0.10**	**85.49*** ± 0.09**	**88.21*** ± 0.08**	**88.17*** ± 0.09**	**83.81 ± 0.07**	**85.52*** ± 0.09**

Table 5: Sentiment classification accuracy in cross-domain scenario. For example, Food → IMDB means training on the Food dataset and test on the IMDB dataset. * means statistical significance using t-test with p-value < 0.05.

Method	Food → IMDB	Food → SST-2	IMDB → Food	IMDB → SST-2
Vanilla	76.08	**76.66**	78.71	75.13
Masker	71.96	75.93	**84.12**	82.38
NFL	76.29	74.76	81.39	83.68
Ours	**77.99**	75.99	83.19	**84.73**

Table 6: Sentiment classification accuracy of DeBERTaV3 in cross-domain scenario.

Method	Food → Kindle	IMDB → Kindle	SST-2 → Kindle	Kindle → Kindle	Kindle → Food	Kindle → IMDB	Kindle → SST-2
Vanilla	76.33	81.17	85.45	89.34	86.79	82.90	82.70
Masker	75.89	81.87	85.31	88.81	86.25	82.34	82.68
NFL	77.58	82.23	86.79	91.00	87.00	82.97	84.32
Ours	**79.12**	**83.64**	**87.21**	**91.31**	**87.91**	**83.06**	**85.05**

Table 7: Sentiment classification accuracy of BERT in cross-domain scenario on the Kindle dataset.

5.2.2 Toy Example for Illustration

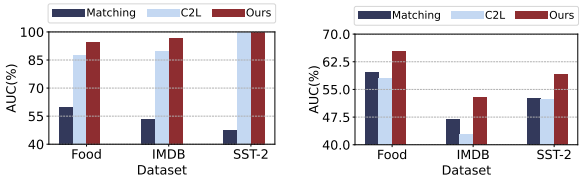
Taking the Food dataset as an example, the LLM labeled 200 keywords (which have high coefficient in the sentiment classification BERT model):

- **Keywords**: ['collects', 'calculated', 'ensued', 'carl', 'pension', 'merry', 'surpassed', ...],
- **Causal words labeled by LLM**: ['surpassed', 'snatched', 'repulsed', 'stunning', ...],
- **Manually corrected word**:

  'merry' **should be** causal word,
  'snatched' **should not be** causal word.

5.2.3 Verification of the Label Generation

We use three backbones (*BERT, RoBERTa, ALBERT*) to generate keywords on 3 datasets (*IMDB, SST-2, Food*) and then employ the large model GPT-o1 to annotate whether each keyword is a causal word. To further assess the annotation performance of the base model, we carried out a quality evaluation study. We used human-corrected results as



(a) Supervised learning (b) Unsupervised learning

Figure 4: AUC of causal word identification via estimated CATE on supervised and unsupervised learning.

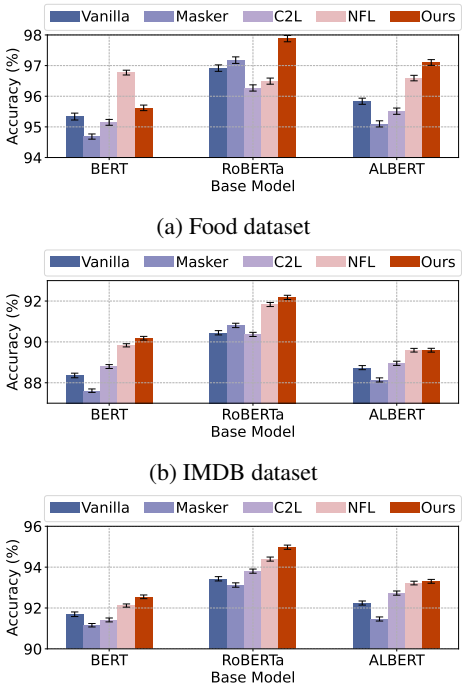
the golden label and evaluated the labels generated by the LLM (GPT-o1) using **Accuracy**, **F1**, **Sensitivity**, and **Specificity**, with results presented in Table 3, where Sensitivity represents the ability to identify causal words, and Specificity represents the ability to identify spurious words. These results demonstrate that the annotations show a high degree of agreement with human judgments.

Furthermore, to validate the quality of GPT-o1's annotations, we sampled 100 keywords each dataset from different domains and different tasks: (1) Relationship Extraction (RE) on Finance Domain (*FinRED*[1]); (2) Natural Language Inference (NLI) on Medical Domain (*MedNLI*[2]); (3) Question Answer (QA) on Legal Domain (*LexGLUE*[3]). We used manually annotated labels as ground truth to evaluate GPT-o1's performance. The experimental results are shown in Table 4.

5.3 Performance Comparison

In cross-domain scenario, we train the sentiment classification model in the source dataset and evaluate in the target dataset. Table 5 presents the AUC($\times$100) scores for this task. First, among the baselines, C2L and NFL show the best performance and outperform the vanilla base model, which shows the necessity of identifying spurious correlations in sentiment analysis. Second, the proposed method consistently outperforms baselines in all scenarios, achieving statistically significant improvements with narrow standard deviations, which demonstrates the effectiveness of our framework. Building upon the existing three pre-trained language models, we have added results for DeBERTaV3. As shown in Table 6 and Table 7, our method continues to significantly outperform the baselines in OOD testing across multiple datasets.

[1] https://github.com/soummyaah/FinRED
[2] https://physionet.org/content/mednli/1.0.0/
[3] https://github.com/coastalcph/lex-glue



(a) Food dataset

(b) IMDB dataset

(c) SST-2 dataset

Figure 5: Sentiment classification accuracy under the in-domain scenarios on three datasets.

5.4 Causal Word Identification Study

As shown in Figure 4, our method has the most superior AUC in estimating the CATEs in both supervised and unsupervised learning. In supervised settings, our method leverages estimated keyword CATE and causal word labels for supervised learning, refining causal effect estimation and outperforming prior methods, surpassing the best ablated variants by 0.4% (Food→IMDB) and 1.5% (SST-2→IMDB) with statistical significance. In unsupervised settings, it ranks keyword CATE values without causal labels and unlike methods such as C2L that struggle with causal estimation, effectively identifies causal words via reweighting and covariate balancing. These results demonstrate our method's robustness and versatility in both supervised and unsupervised scenarios.

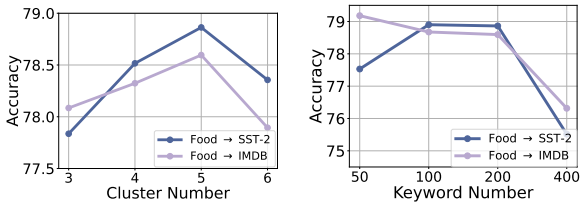
5.5 In-Domain Study

This subsection presents an analysis of our method's performance in in-domain scenarios. The in-domain scenario means training and testing the sentiment classification model in the same dataset.

As shown in Figure 5, our method also shows superior in-domain performance across various base models. While certain keywords may show false correlations, they often strongly associate with spe-

Method	Dataset					
	Food → IMDB	Food → SST-2	IMDB → Food	IMDB → SST-2	SST-2 → Food	SST-2 → IMDB
Vanilla	76.90 ± 0.08	77.32 ± 0.08	84.09 ± 0.08	83.23 ± 0.09	82.17 ± 0.08	84.23 ± 0.08
w/o reweighting	77.05 ± 0.08	76.72 ± 0.08	84.43 ± 0.09	83.11 ± 0.09	82.65 ± 0.07	84.11 ± 0.08
w/o causal loss	77.81 ± 0.07	76.81 ± 0.09	84.32 ± 0.08	82.88 ± 0.09	82.28 ± 0.08	84.00 ± 0.07
w/o sentiment loss	78.21 ± 0.07	78.18 ± 0.09	86.35 ± 0.08	84.48 ± 0.09	83.06 ± 0.08	85.29 ± 0.09
w/o clustering	77.07 ± 0.08	77.83 ± 0.07	86.07 ± 0.08	83.94 ± 0.08	83.94 ± 0.07	85.02 ± 0.09
All (Ours)	**78.60 ± 0.08**	**78.86 ± 0.08**	**86.70 ± 0.09**	**86.67 ± 0.07**	**84.28 ± 0.09**	**86.61 ± 0.08**

Table 8: Ablation studies on reweighting, causal loss, sentiment loss and clustering within our proposed approach.



(a) Cluster number K (b) Keyword number T

(c) Sentiment loss coef. r_s (d) Causal loss coef. r_t

Figure 6: Sensitivity study on sentiment classification.

cific labels in the dataset. Traditional methods that mask false keywords typically degrade performance. In contrast, our approach reweights keywords, reducing the influence of spurious ones while boosting causal keywords. This ensures the model can effectively extract sentiment to maintain strong performance.

5.6 Ablation Study

We conduct ablation studies to validate the effectiveness of each proposed module, as shown in Table 8. In all scenarios, the comprehensive method, incorporating all components, delivers the highest performance. Notably, removing the causal loss causes a significant performance drop in sentiment-sensitive tasks. The ability to estimate causal effects is impaired, which hinders accurate identification of causal keywords and their reweighting.

5.7 Parameter Analysis

The Figure 6 presents a parameter sensitivity analysis for four key parameters: Cluster number K, Keyword number T, Sentiment loss coefficient r_s, and Causal loss coefficient r_t. For K, performance improves as the cluster number increases up to a certain point, especially for the Food → SST-2 task, but plateaus for the Food → IMDB task, indicating that there is an optimal cluster number for each scenario. We determine its value via grid search on the validation set. For T, accuracy increases with the number of keywords until about 100, after which it sharply declines, highlighting the importance of selecting an optimal number of keywords to avoid overfitting or underfitting. The Sentiment loss coefficient r_s shows optimal performance when set to 2, with a decrease in accuracy as the coefficient increases, suggesting that too much emphasis on sentiment loss negatively impacts the model. Similarly, the Causal loss coefficient r_t performs best at 2, with performance dropping as the coefficient grows larger, indicating that overemphasizing causal loss could hurt the model's ability to learn other critical features. In conclusion, the analysis demonstrates that fine-tuning these parameters is essential for maximizing model performance, and an imbalance in any of these factors can lead to suboptimal results.

6 Conclusion

This paper introduces a novel meta-causal learning approach that includes a continuous treatment to identify causal words for robust sentiment classification, overcoming limitations of binary word presence treatments. Observing that text embedding as treatment further incorporates semantic information, and in practice the same CATE estimation model exhibits significant performance variations across treatments with different semantic meanings, we propose to cluster words using pre-trained embeddings and estimate treatment effects via a multi-head classifier. Experimental results confirm the superior performance of our method, particularly in challenging settings with sparse keywords in both in-domain and cross-domain scenarios.

Limitations

Some limitations are that our method relies on the pre-defined clustering thresholds, and our approach does not assess situations where multiple top words exist in a single sentence. When multiple top words are present, the average treatment effect estimation may not accurately reflect the significance of individual words in sentiment classification tasks. Moreover, our current work focuses on word-level analysis as a first step toward understanding how causal and spurious relationships affect sentence-level sentiment classification. In future work, we plan to extend our causal framework to incorporate higher-level linguistic structures, which provides additional signals not captured at the word level.

Acknowledgments

KZ would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program, and the Al Deira Causal Education project. MG was supported by ARC grant DP240102088, as well as WIS-MBZUAI grant 142571.

References

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

Susan Athey, Guido W Imbens, and Stefan Wager. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623.

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

Kousik Barik and Sanjay Misra. 2024. Analysis of customer reviews with an improved vader lexicon classifier. *Journal of Big Data*, 11(1):10.

Fengxiang Cheng, Haoxuan Li, Alina Leidinger, and Robert Van Rooij. 2025a. Revealing the limitations of exploiting causal effects to resolve linguistic spurious correlations. In *AAAI 2025 Workshop on Artificial Intelligence with Causal Techniques*.

Fengxiang Cheng, Chuan Zhou, Xiang Li, Alina Leidinger, Haoxuan Li, Mingming Gong, Fenrong Liu, and Robert Van Rooij. 2025b. Mitigating spurious correlations via counterfactual contrastive learning. *Findings of the Association for Computational Linguistics: EMNLP*.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1013–1025.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10526–10534.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 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)*, pages 4069–4082.

Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015. Learning with a wasserstein loss. *Advances in Neural Information Processing Systems*, 28.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46.

Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

He He, Sheng Zha, and Haohan Wang. 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*, pages 132–142. Association for Computational Linguistics (ACL).

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

Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029. PMLR.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 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)*, pages 163–170.

Lingjing Kong, Guangyi Chen, Petar Stojanov, Haoxuan Li, Eric Xing, and Kun Zhang. 2024. Towards understanding extrapolation: a causal lens. *Advances in Neural Information Processing Systems*, 37:123534–123562.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *The Eighth International Conference on Learning Representations*.

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*.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Andrew 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.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Seung Jun Moon, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. 2021. Masker: Masked keyword regularization for reliable text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13578–13586.

Bo Pang, Lillian Lee, and 1 others. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

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

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

JM Robins, MA Hernán, and B Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5):550–560.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451.

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y 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.

Rui Song, Yingji Li, Mingjie Tian, Hanwen Wang, Fausto Giunchiglia, and Hao Xu. 2025. Causal keyword driven reliable text classification with large language model feedback. *Information Processing & Management*, 62(2):103964.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1872–1878.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. 2024. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 36.

Haotian Wang, Hao Zou, Haoxuan Li, Haoang Chi, Yang Shi, Yuanxing Zhang, Wenjing Yang, Xinwang Liu, and Zhouchen Lin. 2026a. Transformers with endogenous in-context learning: Bias characterization and mitigation. In *The Fourteenth International Conference on Learning Representations*.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Findings of the association for computational linguistics: NAACL 2022*, pages 1719–1729.

Yingrong Wang, Haoxuan Li, Minqin Zhu, Anpeng Wu, Baohong Li, Keting Yin, Ruoxuan Xiong, Fei Wu, and Kun Kuang. 2026b. Causal inference with complex treatments: A survey. *ACM Computing Surveys*, 58(9):1–36.

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

Zhao Wang, Kai Shu, and Aron Culotta. 2021. Enhancing model robustness and fairness with causality: A regularization approach. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 33–43.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 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, page 4586. NIH Public Access.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *The Sixth International Conference on Learning Representations*.

Min Zhang, Haoxuan Li, Fei Wu, and Kun Kuang. 2024. Metacoco: A new few-shot classification benchmark with spurious correlation. In *The Twelfth International Conference on Learning Representations*.

Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10923–10930.

Chunyuan Zheng, Anpeng Wu, Chuan Zhou, Taojun Hu, Qingying Chen, Hongyi Liu, Chenxi Li, Huiyou Jiang, Haoxuan Li, and Zhouchen Lin. 2026. Uplift modeling with delayed feedback: Identifiability and algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 16468–16476.

Chunyuan Zheng, Haocheng Yang, Haoxuan Li, and Mengyue Yang. 2025. Unveiling extraneous sampling bias with data missing-not-at-random. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chuan Zhou, Yaxuan Li, Chunyuan Zheng, Haiteng Zhang, Min Zhang, Haoxuan Li, and Mingming Gong. 2025. A two-stage pretraining-finetuning framework for treatment effect estimation with unmeasured confounding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2113–2123.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492.

Yuqing Zhou and Ziwei Zhu. 2025. Fighting spurious correlations in text classification via a causal learning perspective. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4264–4274.

Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Fei Wu, and Kun Kuang. 2025. Learning double balancing representation for heterogeneous dose–response curve estimation. *Neural Networks*, 189:107600.

Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Xiaoqing Yang, Xuan Qin, Peng Zhen, Jiecheng Guo, Fei Wu, and 1 others. 2024. Contrastive balancing representation learning for heterogeneous dose-response curves estimation. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 38, pages 17175–17183.