COUNTERFACTUAL CONTRASTIVE LEARNING FOR ROBUST TEXT CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text classification has recently been promoted by large pre-trained language models (PLMs). However, derivative models of PLMs still suffer from sensitive performance on different datasets, the reasons are multiple such as cross-domain and label imbalance problems, from which most models may learn the spurious correlation between texts and labels. Existing research requires people to manually add counterfactual samples to the dataset or automatically match so-called counterfactual pairs that are already in the dataset for augmentation. In this paper, we propose a novel LDA-based counterfactual contrastive learning framework and three data augmentation methods, to capture the causal information in texts, which can promote the robustness of text classification. To confirm the effectiveness of our proposed model and methods, we design and conduct several couples of experiments. Experimental results demonstrate that our model works well on five popular text classification datasets on distinct tasks, we find that training with proposed data augmentation outperforms other augmentation methods on many superior models by 1% or above. Plus, robustness tests on different datasets also show a competitive performance, which proves the effectiveness of our model and data.

1 INTRODUCTION

Text classification is a fundamental task in natural language processing, deep learning (DL) models have achieved impressive success in the text classification task in recent years, especially the PLMs (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Lan et al., 2019; Clark et al., 2020; Sun et al., 2020) which are trained on a large number of unlabeled datasets. However, these models may have learned spurious associations between some irrelevant elements and the final label of given texts. In contrast to some predictive or descriptive models based on deep neural networks, causal inference aims to find the causal variables by understanding how intervening on one variable influences another, which is quite a supplement for DL-based or PLM-based text classification models. We think that introducing the causal inference to the training process of PLMs could improve the robustness of PLM derivative models. But two problems lie ahead: first, datasets that can combine causal inference and text classification are rare; second, we need a suitable model to learn the difference between the real-world data and the counterfactual data, i.e., the causal information. In this paper, we propose some novel causal data augmentation methods and an effective model to learn causal information and pay less attention to spurious associations, rely more on robust features, predict data labels and generalize better to cross-domain data.

Counterfactual is an irreplaceable component of causal inference. According to (Pearl & Mackenzie, 2018), there are three levels of human intelligence: association by observing, intervention by intervening, and counterfactual by imaging. Counterfactual is at the highest level, "Would the patient have lower blood pressure had she received another medicine?", "Would Kennedy be alive if Oswald had not killed him?". These are typical counterfactual questions (Rubin, 1974; Lewis, 1974; Pearl, 2009) that ask something never happened, they are constructed from real-world data, which is essentially a data augmentation. Although there exists work on counterfactual data augmentation (Kaushik et al., 2019) to find causal features in natural language processing tasks, they look for human annotations, collecting the minimally dissimilar yet differently labeled examples. By pinpointing the different parts between original and counterfactual texts, it can help models (Choi et al., 2022) identify the causal correlations of the given task. Whereas these methods yield well results,

with complicated details, the costs are still expensive, as there is always a preference for cheap and effective data enhancement methods. Considering the above, we put forward several novel automatically counterfactual data augmentation methods based on current datasets to introduce causal inference into text classification tasks, referring to the counterfactual strategy.

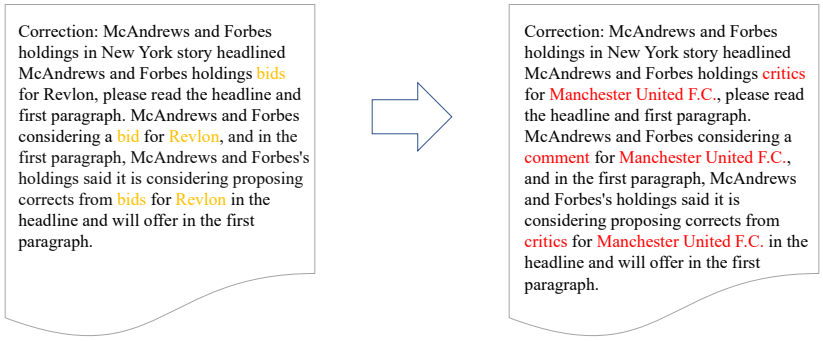


Figure 1: Counterfactual data augmentation on a document. The topic-related terms are replaced by words from other irrelevant topics in the left sub-sample figure's article from Reuters, which has its LDA-counterfactual augmentation on the right. In these two documents, we highlight the topic words in two different colors.

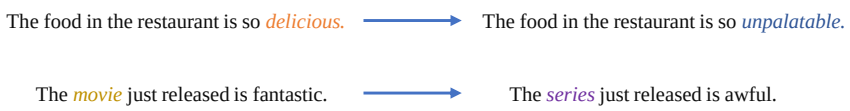


Figure 2: Counterfactual data augmentations on sentences. The counterfactual augmentation of the first line, sampling from Fine Food, consists of replacing the term that expresses emotion with an antonym on its right side. For the factual augmentation on the right, we merely change the word that expresses emotion to something unimportant. In these two texts, we use two distinct colors to highlight the terms.

Word substitution can provide a solution to counterfactual augmentation. During the research process on which element exactly affects human judgment on the classification of texts, we observe that topic words generally carry the central meaning of a document or an essay, and intuitively some specialized adjectives and adverbs used to modify specific subject words may also affect the meaning of texts. We can see that various text classification tasks concentrate on different semantic components, such as sentence-level sentiment analysis depends on adjectives and adverbs, while document classification is determined by specific topic words. As a result, we use distinct counterfactual augmentation skills as the positive and negative pairs of contrastive learning frameworks. Contrastive learning aims to learn effective representation by pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006), so from the perspective of model architecture the learning strategy of contrastive learning is in line with our initiative by which we can identify the differences between real-world and counterfactual data.

To realize the topic word substitution, we attempt to utilize the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model which can distinguish the topic words of a specific text from other words in the text. LDA is one of the most influential topic models, and it is often used for text classification and finding topic words of given texts (Chen et al., 2015; Xie et al., 2017; Li et al., 2018b). It can determine the topic of each document in a document set as a probability distribution so that by analyzing some documents to extract their topic distribution, thus it is possible to perform topic word generation or text classification based on the topic distribution. Considering the reliability and effectiveness of LDA, we adopt it to detect those topic words as a necessary procedure in topic word substitution.

After the extract-and-substitute topic words operation based on LDA, the left essay in Figure 1 which originally stated the financial theme of business activities between McAndrews, Forbes, and Revlon, is transformed into a sports story on the right side that corrected the previous critical coverage of Manchester United. Our counterfactual data augmentation operations consist of topic word substitution, adjective substitution, and adverb substitution, we will also show how much each augmentation operation can affect the model performance on different text classification tasks in the experiments section. When we try to replace all the topic words in an essay, we find it is hard to identify them as their original category, for example, "movie" and "fan" are more likely to appear in entertainment articles, while "vote" and "employment" are more likely in political news. Other linguistic components, such as adjectives and adverbs, may also influence judgments about the category of an article, for example, "bearish" and "inflationary" are more likely to appear in financial reports, and "dunk" and "shoot" tend to occur in sports stories more frequent.

To depict the relationships among these linguistic components and the final labels of texts more clearly, we draw several causal graphs to display the causal structure among the elements, then analyze them separately by the control-observation method. In the following sections, we will comprehensively evaluate counterfactual data augmentation on five benchmark classification tasks, showing that it provides substantial improvements on all five tasks and is particularly helpful for smaller datasets.

2 Background and Related Work

2.1 Robust Text Classification

Conventional text classification aims to give the text that has been provided labels. Despite the recent developments in natural language understanding (Devlin et al., 2018; Liu et al., 2019), large pre-training language models are still challenged by spurious correlations, associating "free" with negative sentiment (Wang & Culotta, 2020), "gay" with detriment (Wulczyn et al., 2017), and "not" with contradiction (Gururangan et al., 2018). Against spurious correlations, recent work pursued additional human annotations, such as human rationales (Jain & Wallace, 2019) and counterfactually-augmented datasets (Kaushik et al., 2019), for supervising neural attention (Zou et al., 2018; Choi et al., 2020), or model gradients (Liu & Avci, 2019; Teney et al., 2020). (Ng et al., 2020; Wang & Culotta, 2020) generate the counterfactual sentences, (Garg & Ramakrishnan, 2020) estimate token importance via counterfactual inference, and (Wang & Culotta, 2020; Klein & Nabi, 2020) find a similar counterpart in the given dataset. However, the automatically annotated methods of counterfactual datasets have been studied rarely.

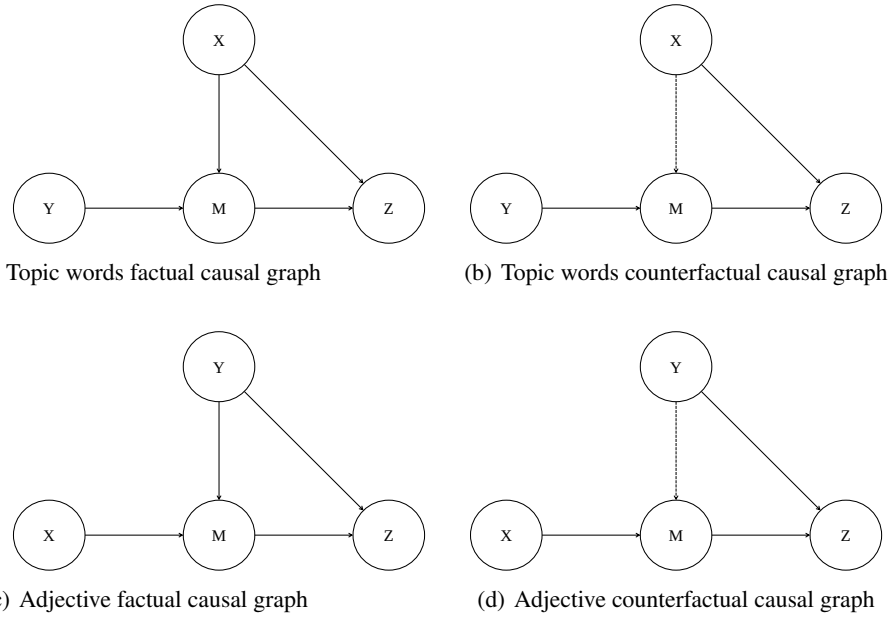
2.2 Text Augmentation

(Yu et al., 2018a) generate new data by translating English datasets into French and back into English, (Xie et al., 2017) noise data as data smoothing, and (Kobayashi, 2018) use predictive language models to replace the synonyms. Whereas these methods perform well, all have complicated details, and the costs are still expensive. As there is a preference for simple and effective data enhancement methods, random insertion, swap, and deletion skills proposed by (Wei & Zou, 2019) is a more widely used data augmentation method in practice. Regretfully, the above methods merely enlarge the scale of training data, none of them shows advantages in improving the model robustness over other methods.

2.3 Counterfactual Text

Counterfactual thinking is an exclusive ability of human beings, it thus has been considered by many researchers, to act as the highest level of causation on the ladder of causal reasoning (Pearl & Mackenzie, 2018). To the best of our knowledge, even the most advanced artificial intelligence system nowadays may still be far from achieving human-like counterfactual reasoning. The counterfactual text is defined based on the real text, which carries a distinctly different semantic meaning from the original text. This is done by replacing the topic-related entity words in the original text with entity words of other topics while keeping the non-topic-related semantics unchanged as much as possible. Our idea of a counterfactual text stems from counterfactual statements (Yang et al.,

2020), in which the authors depict events that did not happen or cannot happen, and the possible consequences had those events happened, e.g., "if kangaroos had no tails, they would topple over" (Lewis, 2013). By developing a connection between the antecedent (e.g., "kangaroos had no tails") and consequent (e.g., "they would topple over"), based on the imagination of possible worlds, humans can naturally form some causal judgments.



(a) Topic words factual causal graph

(b) Topic words counterfactual causal graph

(c) Adjective factual causal graph

(d) Adjective counterfactual causal graph

Figure 3: Causal graphs of linguistic components and text meaning. In all sub-figures, X denotes topic-related words, Y denotes adjectives and adverbs, M represents the actual meaning of texts, and Z is on behalf of the human-annotated text labels. The two upper sub-figures (a) and (b) depict the causal structure changes of documents before and after the LDA-based topic words substitution, while the bottom sub-figures (c) and (d) describe the similar changes when we conduct the StanfordNLP-based part-of-speech recognition and replacement.

2.4 Latent Dirichlet Allocation (LDA)

Probabilistic topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are a class of Bayesian latent variable models that have been adapted to model a diverse range of document genres. As it learns distributions over words, they have become a potent new tool for discovering valuable structures in an unstructured collection. It is predicated on the idea that each document in a collection is composed of several latent subjects, each of which is expressed using a variety of words, such as LDA has a long history of successful applications to news articles (Li et al., 2016) and academic abstracts (Li et al., 2018a; Kim & Gil, 2019). The high probability words in each distribution give us a way of understanding the contents of the corpus at a very high level. LDA is one of the most effective topic models of corpora of documents which seeks to represent the underlying thematic structure of the document collection (Mehrotra et al., 2013).

2.5 Contrastive Learning

Contrastive learning algorithms (Oord et al., 2018; Wu et al., 2018; Chen et al., 2020; He et al., 2020; Khosla et al., 2020) learn similar representations for positive data pairs and dissimilar representations for negative data pairs. Contrastive learning is a type of representation learning that aims to learn an embedding space where the vector representations of similar data are mapped close together, and vice versa (Lowe, 1995; Mika et al., 1999; Xing et al., 2002), they have achieved impressive success in representation learning via self-supervised (Chen et al., 2020; He et al., 2020; Gao et al., 2021) and supervised settings (Khosla et al., 2020; Chen et al., 2022). For instance,

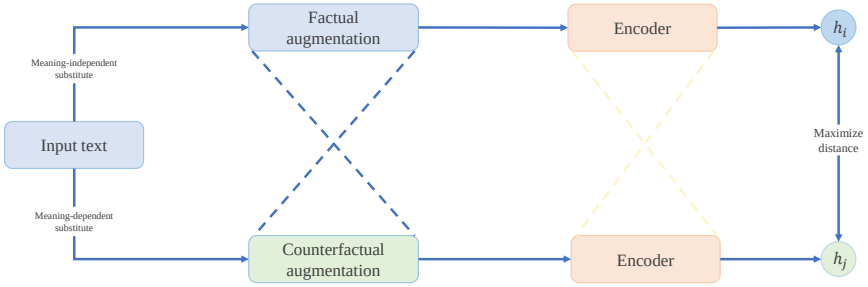


Figure 4: Overview of the counterfactual contrastive learning framework. For each document or sentence in a mini-batch, we augment in two directions: factual and counterfactual. But there is a difference between operations on document-level text and sentence-level text, we thus use LDA to identify the topic words in a document and substitute them with words from other topics; On the other hand, we use StandfordNLP to extract the adjectives and adverbs and then replace them with their antonym. After obtaining the factual and counterfactual augmentation text, we send them into the pre-trained language model to get their hidden representations and maximize the distance of vector space between them.

self-supervised visual contrastive learning defines two views of one image (applying different image augmentations to each view) as positive pair and different images as a negative pair. Supervised contrastive learning (Khosla et al., 2020) defines data with the same labels as a positive pair and data with different labels as a negative pair. We see that distinct contrastive approaches consider different positive and negative pairs constructions according to their learning goals. Early techniques would train using *triplet loss* (Weinberger & Saul, 2009; Chechik et al., 2010) to distinguish two similar objects from a third different object. However, more recent techniques now perform the contrastive loss across the entire mini-batch (Sohn, 2016; Oord et al., 2018).

3 Counterfactual Contrastive Learning Framework

We show the whole proposed counterfactual data augmentation framework in this section. First, we discuss how to identify keywords, including topic words, adjectives, and adverbs; Second, we display the causal graphs that illustrate the causal relationships that exist between these words and the recognition of people respectively; Finally, we discuss what proportion we choose and why.

3.1 Data Augmentation Methods

Causal Words Words of various lexical natures are the basic elements of texts, and words of different lexical natures assume different roles in the semantic representation of text. People tend to determine the category of one document with some keywords which we call topic words, however, when they face a sentence, keywords are always adjectives or adverbs expressing emotions. As mentioned before, we attempt to use the LDA model to identify the topic words and then replace them with words from other topics. For adverbs and adjectives, we plan to employ an automatic tool to locate them, because most current tagging algorithms are capable of the Part-Of-Speech (POS) tagging job with per-token accuracy of slightly over 97% (Manning, 2011; Heid et al., 2020), we choose the StanfordNLP (Qi et al., 2019) as the tool to employ the adjective recognition, for the adjective substitution, we use it to replace current words with their antonym.

Causal Graphs In this part, we make some premises about the causal relationships which may exist in the final label of texts and the above keywords. We here give a set of causal graphs to make the process of our counterfactual data augmentation clear. Aiming to conduct a proper causal intervention, we first formulate the causal graph (Pearl, 2022; Pearl & Mackenzie, 2018; Tang et al., 2020) for the text classification models, which sheds light on how the document contents and dataset biases affect the prediction. Formally, a causal graph is a directed acyclic graph $G = (N, E)$, indicating how a set of variables N causally interact with each other through the causal links E. In Figure 3, X denotes the topic words of a document, Y denotes the adjective and adverb in texts, M

represents the text meaning, and Z is the text category. Our initiative comes from the that it is topic words, adjectives, adverbs, and other semantic components which determine the meaning of each text; however, one may judge the text category correctly only after a glance, during which he/she can scroll some keywords. By contrast, models trained by a large corpus would inevitably capture unintended confounders existing in training data and its corresponding labeled category.

LDA for Topic Words Identification LDA conceives of a document as a mixture of a small number of topics, and topics as a distribution over word types (Blei et al., 2003), these priors are remarkably effective at producing useful results. In LDA, each document of the corpus is assumed to have a distribution over K topics, where the discrete topic distributions are drawn from a symmetric Dirichlet distribution. A topic is a discrete distribution over a fixed vocabulary of word types. As it follows from the definition in algotithm 1, a topic is a discrete distribution over a fixed vocabulary of word types. In the above-mentioned process, the parameters α and β are vectors of hyper-parameters that determine the Dirichlet prior on θ as a set of topic distributions for all documents and ϕ as a set of word distributions in all topics. Typically, symmetric Dirichlet priors are used, where $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$, which defines how probability distribution is concentrated into a single point.

Algorithm 1 LDA algorithm

Input: Corpus D
Parameter: Dirichlet distribution $Dir(\eta)$, topic union $\beta(k)$
Output: Your algorithm's output

Let $k = 1$
while $k \leq K$ **do**
  choose topic $\beta_{(k)} \sim Dir(\eta)$
end while
while each document d in D **do**
  choose a topic distribution $\theta_d \sim Dir(\alpha)$
  let word index $n = 1$
  while $n \leq N_d$ **do**
    choose a topic $z_n \sim Categorical(\theta_d)$
    choose a word $w_n \sim Categorical(\beta_{z_n})$
  end while
end while

Replacement Rule For document-level text classification tasks (e.g., news classification), after topic word identification by the LDA model, we prepare to replace all of the topic words in a document with words of another topic from a document annotated with another label. Such operations can convert the original text into a counterfactually augmented text which expresses a different meaning as we did in Figure 1. For sentence-level text classification (e.g., review classification), we simply exploit the StanfordNLP (Qi et al., 2019) to replace current adjectives and adverbs with their antonym as we did in Figure 2. Through these operations, we yield a counterfactual text to the original sample, we can further use them as positive or negative samples to instruct the contrastive learning progress.

3.2 Contrastive Learning Model

The counterfactual contrastive learning model aims to learn the causal information of the textual meaning itself based on the data augmentation methods. The core issue of counterfactual contrastive learning is constructing positive and negative pairs. First, we consider that one document or sentence D_i in which we have replaced all topic words with words of another topic as a negative sample D_i^- while replacing adjectives, or adverbs will be considered as a positive sample D_i^+, for such an operation alter the meaning of the text little. We here take advantage of the most classic learning strategy in contrastive learning models, and the loss function is as follows:

$$\mathscr{L}_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{h_i^+})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{h_j^+})/\tau}} \tag{1}$$

where h_i and h_i^+ denotes hidden representation of D_i and D_i^+, $\mathrm{sim}(\mathrm{h_i}, \mathrm{h_i^+})$ is the cosine similarity, and τ is a temperature hyper-parameter. In this work, we encode input texts using a text encoder: $h = f_\theta(x)$, and then fine-tune all the parameters using the contrastive learning objective in Eq 1.

3.2.1 CONTRASTIVE TEXT REPRESENTATION

Text Encoder The text encoder is to map the raw text onto a vector space where the metrics (or measurements) between texts can be computed. The large pre-trained language models, such as BERT (Devlin et al., 2018) or RoBERTa Liu et al. (2019), has recently been employed as efficient text encoders to obtain text representations and achieve promising results. Specifically, BERT takes a text x composed of a list of tokens as input, and outputs a hidden-state vector for each of the tokens; we take the hidden-state vector corresponding to the *CLS* token as the text representation of x. For later use, we denote the BERT text representation module as $f(\cdot)$ and denote all of its parameters as θ.

Supervised Contrastive Text Learning Our supervised counterfactual contrastive learning framework is a metric-based approach like former works (Wohlhart & Lepetit, 2015; Wen et al., 2016; Tao et al., 2016; Yu et al., 2018b), but different from Prototypical Networks that align query texts with prototypes, we optimize the measurement free of prototypes, by learning to align two text representations using supervised contrastive learning. It pulls closer the text representations belonging to the same class and pushes away text representations belonging to different classes among texts from both query and support sets. The model design of our supervised contrastive learning is based on the "batch contrastive learning" framework (Chen et al., 2020) and the supervised contrastive learning strategy (Khosla et al., 2020).

4 EXPERIMENTS

We conduct experiments on some quality datasets with our proposed contrastive learning framework that are discussed in Section 3: Section 4.1 introduces the datasets and experimental metrics, Section 4.2 for the specific implementation, Section 4.3 for the comparison methods, and Section 4.4 for the results and discussion.

4.1 DATASETS AND METRICS

We evaluate our counterfactual data augmentation method on 6 text classification datasets, including 2 news classification datasets: **20NewsGroup** (Lang, 1995), **Reuters**[1] in which we choose Reuters-52 (R-52) with 52 categories in total, and 4 review classification datasets: **Fine Foods** and **Movies** from **Amazon** (He & McAuley, 2016), **IMDB** (Maas et al., 2011), and **SST-2** (Socher et al., 2013). To demonstrate the robustness of our proposed model, we conduct natural language inference experiments on **MNLI** (Williams et al., 2017) dataset. It is noteworthy that our counterfactual augmentation method adopts the self-supervision signals without requiring any additional human efforts.

We use the official train and test splits if exist, or we randomly divide the dataset with a 9:1 ratio, using them for train and test, respectively. To ensure the training process is going to be convergence, we use 10% of the train set for validation purposes. The hyper-parameters are chosen by the best performance on the validation set. All the reported results are averaged over 5 trials.

4.2 IMPLEMENTATION DETAILS

We implement our model with PyTorch (Paszke et al., 2017), and the NT-Xent loss function with PyTorch Metric Library (Musgrave et al., 2020). For Transformer architectural pre-trained language models, we use Transformers library(Wolf et al., 2020). For the BERT classifier, we train bert-base-uncased with a batch size of 64 for SST-2, IMDb, Fine Food, and 8 for 20NewsGroup and R-52 over 10 epochs, ensuring convergence. We used AdamW (Loshchilov & Hutter, 2017) with a learning

[1] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

Model	20News	R-52	FineFoods	AmazonMovies	IMDB
One-hot-based Method					
TextGCN	85.9 ± 2.2	92.5 ± 1.4	-	-	-
Bag-of-words-based Method					
TextCNN	76.8 ± 0.3	85.3 ± 0.5	65.66 ± 0.16	76.9 ± 0.7	78.3 ± 0.6
TextRNN	75.4 ± 0.6	90.6 ± 0.8	65.66 ± 0.16	78.4 ± 0.9	80.1 ± 0.8
Unsupervised Method					
$\text{BERT}_{\text{base}}$	85.3 ± 1.2	96.2 ± 0.6	91.2 ± 1.3	95.4 ± 1.6	93.8 ± 1.1
$\text{RoBERTa}_{\text{base}}$	83.8 ± 0.7	96.1 ± 0.4	90.7 ± 0.6	93.2 ± 1.1	93.5 ± 0.7
DeCLUTR	85.6 ± 0.8	96.4 ± 0.5	91.0 ± 0.6	95.3 ± 1.1	94.2 ± 1.0
SimCSE	86.1 ± 0.5	96.8 ± 0.8	91.3 ± 0.7	93.1 ± 0.6	92.6 ± 0.4
C^2L	86.8 ± 0.4	96.6 ± 0.7	91.8 ± 1.3	93.2 ± 1.6	91.0 ± 1.2
LDACCL(ours)	$\mathbf{87.9 \pm 0.3}$	$\mathbf{97.6 \pm 0.4}$	$\mathbf{93.1 \pm 0.9}$	$\mathbf{95.9 \pm 0.5}$	$\mathbf{95.1 \pm 0.7}$

Table 1: Accuracy (%) on the counterfactually augmented news classification datasets and review classification datasets. As we mentioned in the former sections, the counterfactual contrastive learning model is trained on the given training sets and their counterfactually augmented data, and evaluated on the original test sets, respectively.

rate of 5e-5 and the linear scheduler with 50 warm-up steps. All models are trained on up to eight NVIDIA RTX3090 24GB GPUs.

4.3 Baselines

4.3.1 One-hot Method

TextGCN The most successful point of TextGCN (Yao et al., 2019) is that they build a single text graph for a corpus based on word co-occurrence and document word relations, then learn a TextGCN for the corpus, it then jointly learns the embedding for both words and documents, as supervised by the known class labels for documents.

4.3.2 Bag-of-words Method

TextCNN TextCNN (Yoon, 2014) we use here has a convolution layer, the kernel sizes of which are 2, 3, and 4, respectively, and each has 50 kernels. Then we apply global max-pooling and a 2-layer fully forward neural network with ReLU activation. The dropout rate is 0.5 and $L2$ regularization coefficient is $3e^{-4}$.

TextRNN TextRNN (Fu et al., 2014) uses a bidirectional GRU the same as the sentence encoder and max-pooling across all GRU hidden states to obtain the sentence embedding vector, and the output layer is a 2-layer FFN. The dropout rate and L_2 regularization coefficient are the same as TextCNN.

4.3.3 Unsupervised Method

DeCLUTR Deep contrastive learning for unsupervised textual representations (DeCLUTR) (Giorgi et al., 2020) is a self-supervised model for obtaining universal sentence embedding that does not require labeled training data, it samples sentences within one document as positive samples of origin and those from other documents as negative samples.

SimCSE SimCSE is an unsupervised approach, which takes an input sentence and predicts itself in a contrastive objective, with only standard dropout used as noise. The most critical of SimCSE (Gao et al., 2021) is that it makes dropout acts as minimal positive data augmentation, and removing it leads to a representation collapse.

SSMBA As a masking-based generative baseline, we implement SSMBA (Ng et al., 2020), a corrupt-and-reconstruct text augmentation method, which masks an arbitrary number of word positions and unmasks them using BERT. They augment 5 samples for each sample with RoBERTa (Liu et al., 2019) and train the BERT-Base classifier on the augmented dataset with soft-label.

MASKER MASKER (Moon et al., 2021) alters the fine-tuning process by enforcing BERT to make a prediction based solely on the surrounding contexts by masking out keywords, and it achieves

a robust text classification when some samples from the test set are out-of-distribution or under cross-domain scenarios.

C^{2}L Different from former methods, C^{2}L (Choi et al., 2022) aims to add robustness to the causal classification model by contrastive learning the sentence embedding with a mask the irrelevant words and keywords which are positive and negative samples, however, they also use the same augmentation and training method in SSMBA.

Model	Sentiment						MNLI		
	I→F	I→S	F→I	F→S	S→I	S→F	T→L	T→F	L→F
BERT$_{base}$	88.1±2.0	87.0±0.5	82.7±1.6	74.1±1.8	88.3±1.1	80.9±1.0	81.1±0.8	79.4±0.3	80.4±0.6
RoBERTa$_{base}$	87.6±1.7	87.2±0.6	82.3±1.3	74.8±1.6	88.0±0.7	81.1±1.2	80.9±0.6	80.1±0.5	80.7±0.7
DeCLUTR	88.5±1.3	87.5±0.8	82.6±1.1	74.5±1.5	88.6±0.8	81.2±1.4	81.3±0.4	81.2±0.9	80.6±0.6
SimCSE	88.2±1.1	87.9±0.5	83.4±1.0	75.3±1.3	88.7±0.8	81.8±1.2	81.5±0.5	81.6±0.4	81.4±0.3
SSMBA	88.9±0.3	87.2±0.9	83.5±1.2	74.8±1.0	87.8±0.3	80.6±0.4	80.4±0.5	79.8±0.4	80.2±0.5
MASKER	86.8±0.0	85.8±0.0	78.3±0.0	75.1±0.0	84.0±0.0	81.0±0.0	80.4±0.0	78.5±0.0	79.6±0.0
C^{2}L	89.0±0.6	87.7±0.6	84.7±1.1	77.5±0.3	89.7±0.6	83.8±1.2	82.1±0.8	80.3±0.5	81.5±0.6
LDACCL(ours)	**89.6±0.4**	**88.4±0.4**	**85.1±0.8**	**78.2±0.4**	**89.6±0.4**	**84.4±0.8**	**82.9±0.9**	**80.8±0.4**	**82.3±0.5**

Table 2: Cross-Domain Accuracy: accuracy (%) on the three sentiment analysis and MNLI datasets. $*$ indicates that the results are reproduced by the original implementation. We denote each sentiment dataset as follows: IMDB (I), FineFood (F), and SST-2 (S). For MNLI, each domain is denoted as follows: Telephone (T), Letters (L), and FaceToFace (F).

4.4 Results and Discussion

The text classification experimental results of the aforementioned models are shown in Table 1. We take the results of baseline models on 2 news classification datasets and 3 review classification datasets. The current state-of-the-art (SOTA) contrastive learning model on 3 review classification datasets is SSMBA, and SOTA counterfactual model on the 2 news classification datasets is C^{2}L. We can observe the proposed model LDA-based counterfactual contrastive learning (LDACCL) outperforms all the baselines after learning from factual semantics and counterfactuals. Specifically, LDACCL improves the accuracy for 1.3% on the Fine Foods, 0.5% on Amazon Movies, 0.9% on IMDB, and 1.1% on 20NewsGroup, 0.8% on R-52 from the BERT$_{BASE}$. Among the results of multiple experiments with different models, the LDACCL is the with the smallest fluctuation, which can demonstrate that counterfactual semantics makes the network more robust.

The performance of neural networks can deteriorate under a domain shift between training and test data. Previous literature (Moon et al., 2021) has shown that over-relying on the domain-specific keywords limits the generalization ability of networks, as the same keywords normally do not appear in another domain, for which we aim to remove such spurious features. Table 2 presents the classification accuracy for the cross-domain scenario, where each model is trained only on the source domain and evaluated on the target domain without further training. In Table 1 and Table 2, we can find LDACCL is more robust, outperforming all the baselines in cross-domain settings. Apart from that, against domain shifts. It demonstrates that the model becomes more robust against spurious correlations when the network learns to capture the causal components by comparing the original text and its counterfactual augmented text.

5 Conclusion

In this paper, we propose a novel LDA-based counterfactual contrastive learning model for robust text classification and present some efficient counterfactual data augmentation methods. Unlike existing efforts using causal features for contrastive learning, to the best of our knowledge, our work is the first to study the debias of task models, without increasing annotation overheads on the human side. Our model does not require a large amount of labeled training data and applies to any text encoder. We demonstrated the performance and robustness of our model by evaluating some datasets of top quality and showing effectiveness counterfactual data augmentation methods. When used to extend the pre-training of a transformer-based language model, our self-supervised objective closes the performance gap with existing methods that require human-labeled training data. Our experiments suggest that the performance can be further improved by increasing the model and training set size. We hope future research to explore generalization to other tasks and not be limited to natural language processing.

REFERENCES

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv preprint arXiv:2201.08702*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. Dataless text classification with descriptive lda. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seung-won Hwang. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6695–6704, 2020.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. C2l: Causally contrastive learning for robust text classification. 2022.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1199–1209, 2014.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*, 2020.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*, 2020.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Ruining He and Julian McAuley. 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*, pp. 507–517, 2016.

Stefan Heid, Marcel Wever, and Eyke Hüllermeier. Reliable part-of-speech tagging of historical corpora through set-valued prediction. *arXiv preprint arXiv:2008.01377*, 2020.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Sang-Woon Kim and Joon-Min Gil. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9(1):1–21, 2019.

Tassilo Klein and Moin Nabi. Contrastive self-supervised learning for commonsense reasoning. *arXiv preprint arXiv:2005.00669*, 2020.

Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*, 2018.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

David Lewis. *Counterfactuals*. John Wiley & Sons, 2013.

Changzhou Li, Yao Lu, Junfeng Wu, Yongrui Zhang, Zhongzhou Xia, Tianchen Wang, Dantian Yu, Xurui Chen, Peidong Liu, and Junyu Guo. Lda meets word2vec: a novel model for academic abstract clustering. In *Companion proceedings of the the web conference 2018*, pp. 1699–1706, 2018a.

Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. Dataless text classification: A topic modeling approach with document manifold. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 973–982, 2018b.

Zhenzhong Li, Wenqian Shang, and Menghan Yan. News text classification model based on topic model. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–5. IEEE, 2016.

Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

David G Lowe. Similarity metric learning for a variable-kernel classifier. *Neural computation*, 7(1): 72–85, 1995.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Christopher D Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pp. 171–189. Springer, 2011.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 889–892, 2013.

Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pp. 41–48. Ieee, 1999.

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

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Pytorch metric learning. *arXiv preprint arXiv:2008.09164*, 2020.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 373–392. 2022.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*, 2019.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8968–8975, 2020.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3716–3725, 2020.

Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1420–1429, 2016.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, pp. 580–599. Springer, 2020.

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

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3109–3118, 2015.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399, 2017.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*, 2017.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15, 2002.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. Semeval-2020 task 5: Counterfactual recognition. *arXiv preprint arXiv:2008.00563*, 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7370–7377, 2019.

K Yoon. Convolutional neural networks for sentence classification [ol]. *arXiv Preprint*, 2014.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018a.

Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 188–204, 2018b.

Yicheng Zou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics*, pp. 868–877, 2018.