Interpretable Research Replication Prediction via Variational Contextual Consistency Sentence Masking

Tianyi Luo[1], **Rui Meng**[2], **Xin Eric Wang**[1], **Yang Liu**[1]
[1]Computer Science and Engineering, University of California, Santa Cruz
[2]Lawrence Berkeley National Laboratory, University of California, Berkeley
{tluo6, xwang366, yangliu}@ucsc.edu
rmeng@lbl.gov

Abstract

Research Replication Prediction (RRP) is the task of predicting whether a published research result can be replicated or not. Building an interpretable neural text classifier for RRP promotes the understanding of why a research paper is predicted as replicable or non-replicable and therefore makes its real-world application more reliable and trustworthy. However, the prior works on model interpretation mainly focused on improving the model interpretability at the word/phrase level, which are insufficient especially for long research papers in RRP. Furthermore, the existing methods cannot utilize a large size of unlabeled dataset to further improve the model interpretability. To address these limitations, we aim to build an interpretable neural model which can provide sentence-level explanations and apply weakly supervised approach to further leverage the large corpus of unlabeled datasets to boost the interpretability in addition to improving prediction performance as existing works have done. In this work, we propose the Variational Contextual Consistency Sentence Masking (**VCCSM**) method to automatically extract key sentences based on the context in the classifier, using both labeled and unlabeled datasets. Results of our experiments on RRP along with European Convention of Human Rights (ECHR) datasets demonstrate that VCCSM is able to improve the model interpretability for the long document classification tasks using the area over the perturbation curve and post-hoc accuracy as evaluation metrics.

1 Introduction

Scientific research results that cannot be reproduced are unreliable and negatively impact the development of science. Therefore, it is important to know whether a published research result can be replicated or not. To this end, domain researchers have conducted several direct replication projects in contemporary published social science studies (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2018; Collaboration et al., 2015). Such direct replication, however, is very time-consuming and expensive. A much more efficient and cheaper alternative, Machine Learning (ML), is utilized for predicting research replication (Dreber et al., 2019; Yang, 2018; Altmejd et al., 2019; Luo et al., 2020). In this paper, we model the task of predicting research replication as a binary classification problem and name it Research Replication Prediction (RRP) task which is shown in Figure 1(a). Nonetheless, applying the neural network models in the context of RRP faces two challenges. The first challenge is that the existing neural network models used in RRP are characterized as a black box because their predictions are hardly understandable. Without intelligible explanations for the predictions, results of RRP may not be widely accepted as reliable and trustworthy. Despite the progress in interpretable machine learning (Hechtlinger, 2016; Smilkov et al., 2017; Singh et al., 2018; Serrano

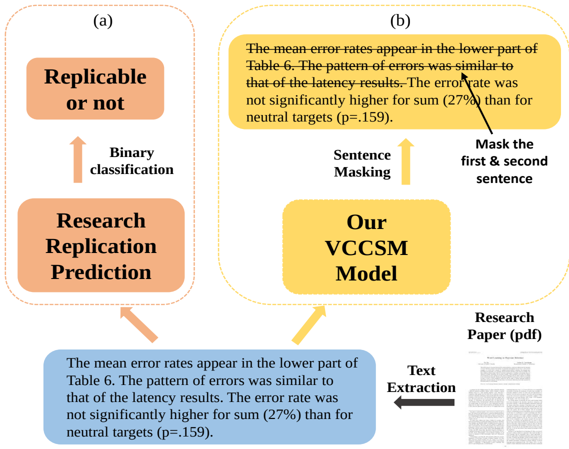


Figure 1: (a) Given the text information of a research paper, Research Replication Prediction (RRP) task predicts whether the paper can be reproduced or not. (b) Having the same input as (a), our VCCSM model can keep the important sentences (through masking unimportant ones) which are related to reproducibility.

Findings of the Association for Computational Linguistics: ACL 2022, pages 3864 - 3876

and Smith, 2019; Han et al., 2020; Chen and Ji, 2020; Chen et al., 2021) , the existing works mostly focus on improving the interpretability only at the word/phrase level which might work well for short documents. However, research papers in our RRP problem are usually lengthy (the average length of words is about 10,000). Building interpretable models for long documents is a challenging task due to the massive amount of textual information.

The second salient challenge is the small size of labeled dataset in RRP due to the high cost (e.g., funding requirement, human labor, etc.) of direct replications. Training an interpretable neural network also requires a large size of labeled dataset and weakly supervised learning can help utilize the unlabeled dataset. Although weakly supervised approaches have been utilized to make use of a large size of unlabeled dataset (Berthelot et al., 2019; Xie et al., 2019; Chen et al., 2020), they have mainly focused on improving the prediction performance but not the interpretability. We therefore aspire to build a weakly supervised interpretable neural text classifier for predicting research replication that can leverage the existence of the large corpus of unlabeled articles to boost up both the prediction performance and the interpretability.

To tackle the first challenge mentioned above, we built an interpretable neural network model which can automatically select key sentences instead of words/phrases by adding a *variational sentence masking* layer on the input layer which is a simple modification of network architecture but can effectively improve the model interpretability. By adding a *variational sentence masking*, we can adopt information bottleneck framework (Tishby et al., 2000; Alemi et al., 2016) to train the model and improve both the prediction performance and interpretability by identifying important sentences. In addition, we hypothesize that whether to mask a sentence or not should also depend on its context (whether other sentences in the same paper are masked) in the case for long research papers because the information provided by extracted key sentences should not be redundant. Therefore, we invoke a contextual sentence masking approach using the LSTM model (Hochreiter and Schmidhuber, 1997). The extracted key sentences after masking are considered as our interpratable outcomes for each research paper.

To resolve the second challenge, we developed a new weakly supervised method which makes use of unlabeled dataset to improve both the prediction performance and interpretability. Specifically, we adopted the consistency training methods (Laine and Aila, 2016; Tarvainen and Valpola, 2017; Xie et al., 2019) which regularize model predictions to be invariant to the small noises added to the input. Consistency training were used to improve the prediction performance with the help of unlabeld dataset (Xie et al., 2019). In this paper, to improve the interpretability along with the prediction performance, we propose a consistency training method with sentence masking through replacing the noises-added input of unlabeled dataset by masked sentences. Specifically, for each unlabeled research paper, we generate the first prediction by using only the extracted key sentences after the sentence masking. Then, we generate the second prediction using all the sentences in the research paper without masking. The consistency check is then imposed upon the two predictions by minimizing the difference between them. Through the consistency training, an extra large size of unlabeled dataset can be utilized to make model continually learn how to extract the key sentences of a research paper so that the model interpretability is further improved.

In sum, our main contribution is the proposal of a variational contextual consistency sentence masking (VCCSM) method as shown in Figure 1(b) that is able to (1) extract the key sentences based on the context of a research paper and (2) leverage the large number of unlabeled sets of papers using a consistency checking mechanism. We present experimental results to validate the usefulness of our proposed methods on two neural network models, LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2018) on the RRP along with ECHR datasets. In particular, we find VCCSM is able to improve both the replication prediction accuracy and the interpretability for long research papers and general long documents.

2 Related Work

Blackbox Research Replication Prediction Research Replication Prediction, knowing whether a published research result is replicable or not, is important. Recently, several large scale of direct replication projects have been conducted in social science studies to alleviate the replication crisis. But the cost of direct replication is too high to have a large size of annotated dataset. Therefore, an alternative ML method that is much cheaper and more efficient than direction replication is utilized in RRP. Luo et al. (2020) proposed a neural text classifier to achieve the best performance on RRP. But their model is a blackbox and cannot provide faithful explanations about why a research paper is predicted as replicable or non-replicable.

Interpreting Neural Networks Various approaches have been proposed to interpret neural network models from the post-hoc manner, such as gradient-based (Simonyan et al., 2014; Hechtlinger, 2016; Sundararajan et al., 2017), attention-based (Serrano and Smith, 2019), decomposition-based (Murdoch et al., 2018; Singh et al., 2018), example-based methods (Koh and Liang, 2017; Han et al., 2020), and word masking (Chen and Ji, 2020). However, these interpretation methods have their own limitations, including only work with specific neural network model, render doubts on faithfulness, and need additional work to provide the explanations based on trained models. In this paper, we focus on model-agnostic explanation methods. More specifically, we follow the research of masking methods which can improve both the prediction performance and interpretability by adopting information bottleneck framework (Tishby et al., 2000; Alemi et al., 2016) to identify important sentences.

Improving interpretability via word masking Chen and Ji (2020) proposed a word masking method which can automatically select important words in the training process and build interpretable neural text classifiers by formulating their problem in the framework of information bottleneck. The proposed solution mainly deals with the short text and the average length (words) in all the seven datasets they used are less than 300. Four of them are less than 25. In constrast, the average length (words) of research papers in our RRP task is about 10,000 which is much longer than the ones used in (Chen and Ji, 2020). Therefore, we view word masking as insufficient for our task. On the other hand, Chen and Ji (2020) learn independently on whether each word is masked or not. But context matters, especially for long documents. Different from prior work, we utilized the context information (whether other sentences in the same paper are masked or not) of each sentence by applying LSTM models to decide whether to mask this sentence or not. We hypothesize that context masking is better than independent masking, especially for long documents such as the research papers in RRP.

Consistency Training on Unlabeled Dataset The annotated data in RRP is collected using direction replication and its size is small. Therefore, weakly supervised learning methods need to be used to improve the model performance in RRP with the help of the unlabeled dataset. The existing weakly supervised methods applied in RRP focus mainly on improving the prediction performance, but less so about the model interpretability.

Consistency training can improve the robustness of models by regularizing model predictions to be invariant to small noise applied to input examples (Sajjadi et al., 2016; Clark et al., 2018). Xie et al. (2019) proposed to substitute the traditional noise injection methods in the consistency training with high quality data augmentations so that a new consistency training based weakly supervised method is proposed and the performance is improved with the help of unlabeled dataset. But they focused only on improving the prediction performance.

In this paper, we conduct the consistency training on the unlabeled dataset to improve both prediction performance and interpretability by substituting the traditional noise injection methods with sentence masking methods, which is the major contribution of our paper. More specifically, we first mask the unimportant sentences and keep the critical sentences. Then we make the predictions on the kept key sentences the same as the ones based on all the sentences in the research paper without masking. Finally, we conducted the consistency check by minimizing the difference between them.

3 Problem Statement

In this paper, our main goal is to improve the intepretibility of neural textual classifier for Research Replication Prediction (RRP). First we introduce the RRP task.

Research Replication Prediction (RRP) task In RRP, we hope to build a model f that takes each research `article` as input and predicts whether the made research claim is replicable or not $f(\texttt{article}) \in \{0 \text{ (non-replicable)}, 1 \text{ (replicable)}\}$. There are different definitions and criteria for claiming a research paper to be replicable. In this work, a research paper is replicable means that an independent replication can provide evidence of a statistically significant effect in the same direction as the original paper.

Interpretable Research Replication Prediction In this paper, we aim to build an interpretable neural textual classifier for RRP. Improving the model interpretability can help us understand why a research paper is predicted as replicable or non-replicable and make its application in the real world achieve more reliability and trustworthiness. Different from generating post-hoc explanations based on well-trained models, we adopt the information bottleneck framework (Tishby et al., 2000; Alemi et al., 2016) to train our model and build a more interpretable neural textual classifier for RRP.

Preliminaries and notations To perform the above task, we have an labeled training dataset $\mathcal{L} := \{(x_i, y_i)\}_{i=1}^{L}$, an unlabeled dataset $\mathcal{U} := \{x_i\}_{i=1}^{U}$, and a test dataset $\mathcal{T} := \{(x_i, y_i)\}_{i=1}^{T}$, where L, U, and T are the number of labeled training, unlabeled training, and testing datasets respectively. x_i contains a sequence of sentences $x_i = [x_{i1}, x_{i2}, ...x_{ij}..., x_{iS}]$ in the i_{th} research paper and S is the maximum number of sentences in a research paper in RRP task. For the j_{th} sentence in x_i, $x_{ij} = [x_{ij1}, x_{ij2}, ...x_{ijk}..., x_{ijK}]$, where n is the maximum number of words in a sentence and $x_{ijk} \in \mathbb{R}^d$ which indicates the word embeddings as the model input. All the sentences have the same length K by truncating. And y_i is x_i's binary classification label which is either '1' (replicable) or '0' (non-replicable). A neural textual classifier can be trained to output the replication labels given any new research paper x_i.

4 Method

The details of our proposed variational contextual consistency sentence masking (VCCSM) method are described in this section.

4.1 Model Overview

Our model contains two key modules: variational contextual sentence masking and consistency training. Variational contextual sentence masking module is applied in the training on both labeled and unlabeled datasets. Consistency training is only used in the training on the unlabeled dataset.

In the training on labeled dataset, variational contextual sentence masking module extracts the key sentences via contextual masking (LSTM model). Then the supervised loss is calculated and optimized to minimize the difference between prediction using the extracted sentences as the input and the ground truth label in the information bottleneck framework. The formula of supervised loss will be described later in this section and the architecture of the model on how to train the labeled dataset is shown in the left part of Figure 2.

In the training on unlabeled dataset, different from the prior works, we conduct the consistency training on the unlabeled dataset to improve both prediction performance and interpretability by substituting the traditional noise injection methods with sentence masking methods. Consistency training can improve the model robustness by regularizing model predictions to be invariant to small noise applied to input examples (Sajjadi et al., 2016; Miyato et al., 2018; Clark et al., 2018). Typical noise injection methods included additive Gaussian noise, dropout noise or adversarial noise. The existing consistency training based work e.g., (Xie et al., 2019) focuses only on improving the prediction performance instead of interpretability. The consistency training methods utilized in this paper are based on variational contextual sentence masking and can also improve the model interpretability. Our optimization goal is to minimize the difference between prediction using the extracted vital sentences and prediction made on all the sentences without masking in the information bottleneck framework. The formula of unsupervised loss will be described later in this section and the architecture on how to train the unlabeled dataset is shown in the right part of Figure 2.

4.2 Variational Contextual Sentence Masking

Inspired by Chen and Ji (2020), we want to add a mask layer M after the sentence embeddings layer to help the model select the key sentences, where $M = [M_1, M_2, ...M_j..., M_S]$ and S is the maximum number of sentences in a research paper. The embedding of each sentence is concatenated by word embeddings included in this sentence.

Each $M_j \in \{0, 1\}$ is a binary random variable to decide whether we mask this sentence or not. For each sentence in one research paper, M_j should be related to both the current sentence and the sentences around it (context). Therefore, we use LSTM model to generate the contextual sentence mask M_j for the j_{th} sentence in one research paper, where $M_j = \text{LSTM}(x_j)$, $j = 1, 2, ..., S$. x can be any given research paper. This contextual sentence mask layer M together with the sentence embed-

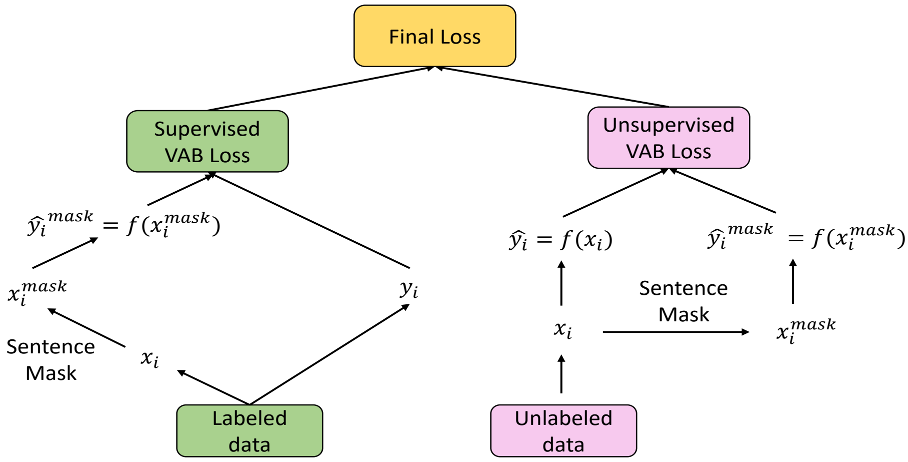


Figure 2: The architecture of variational contextual consistency sentence masking (VCCSM).

dings are considered as the input of neural network text classifiers in RRP, which is denoted as follows,

$$Z = X^{\text{mask}} = M \bigodot X, \tag{1}$$

where $\bigodot$ is an element-wise multiplication, X are all the examples in any given dataset, and X^{mask} denotes the internal representations of all the examples. Our goal is to optimize M so that the model can extract the key sentences for each research paper.

The information bottleneck theory aims to learn an encoding Z of the input X with maximal information on predicting the target Y while keeps X's the least redundant information (Tishby et al., 2000; Alemi et al., 2016). As proven effective and flexible in identifying important features (Chen and Ji, 2020), the information bottleneck framework is employed in our model. we want to make $Z = X^{\text{mask}}$ maximally expressive on predicting Y while being maximally compressive on X. Therefore, following the standard information bottleneck theory (Tishby et al., 2000), our objective function is denoted as follows:

$$\max_{Z} I(Z;Y) - \beta \cdot I(Z;X), \tag{2}$$

where the definitions of X and $Z = X^{\text{mask}}$ are given in Equation 1. Y is the target output, $I(\cdot;\cdot)$ denotes the mutual information, and $\beta \in \mathbb{R}_+$ is a coefficient that balances the two terms in the information bottleneck loss function. The formula of mutual information I should include the parameters θ which need to be optimized. For simplicity, we ignore θ in the following formulae.

However, computing the mutual information in Equation 2 is usually computationally challenging. Therefore, we adopted variational inference method to construct a lower bound for Equation 2. After constructing the lower bound and applying the reparameterization trick (Kingma and Welling, 2013), we can optimize the objective utilizing stochastic gradient descent. In this subsection, we simply listed the lower bound of Equation 2. The complete details on the derivation of lower bound for variational contextual sentence masking is shown in Appendix A.1.

Assuming that the true joint distribution is $P(X, Y, Z)$ and X, Y, Z are random variables which have the following conditional dependency: $Y \leftrightarrow X \leftrightarrow Z$. And x, y, z are instances of random variables. The lower bound of Equation 2 is as follows:

$$\sum_{x,y,z} P_X(x)P_{Y|X}(y|x)P_{Z|X}(z|x)\log Q_{Y|X}(y|z) - \beta \sum_{z,x} P_X(x)P_{Z|X}(z|x)\log \frac{P_{Z|X}(z|x)}{Q_Z(z)} \tag{3}$$

To compute Equation 3, we use the empirical data distribution including two Delta functions to approximate the $P_{X,Y}(x, y)$. Therefore we have the loss function of variational information bottleneck (VAB) as follows:

$$\ell_{vib} = -(\mathbb{E}_{P_{X,Y}(x,y)}[\mathbb{E}_{P_{Z|X}(z|x)}[\log(Q_{Y|Z}(y|z)] - \beta \cdot \text{KL}[P_{Z|X}(z|x)||Q_Z(z)]]]) \tag{4}$$

4.3 Consistency Training based on Variational Contextual Sentence Masking

In this work, we utilized a particular consistency training setting where the masked input x^{mask} is generated by applying variational contextual sentence masking mentioned above on each input x, which can be written as follows: $x^{\text{mask}} = M \cdot x$, to improve both the interpretability and prediction performance.

More specifically, inspired by Xie et al. (2019), we propose to substitute the traditional noise injection methods with our Contextual Sentence Masking module to generate the masked input x^{mask} given each input x in the unlabeled dataset which can be written as follows: $x^{\text{mask}} = M \cdot x$. We also use the information bottleneck framework in the consistency training. The only change is to replace the ground truth label y_i with the prediction $\hat{y}_i$ given the original research paper x_i as the input. To be noted, the sentence mask layer is not used when predicting $\hat{y}_i$.

4.4 Variational Information Bottleneck (VAB) Loss Function

As shown in Figure 2, our VAB loss functions contains two parts: a supervised VAB loss ℓ_{su} and an unsupervised VAB loss ℓ_{un}. The same model is optimized in both losses.

Supervised VAB Loss Since we have ground truth labels in the labeled dataset, the supervised VAB loss ℓ_{su} is the same as the VAB loss ℓ_{vlb} in Equation 4 and it is denoted as follows:

$$\ell_{su} = -(\mathbb{E}_{P_{X,Y}(x,y)}[\mathbb{E}_{P_{Z|X}(z|x)}[\log(Q_{Y|Z}(y|z)] - \beta \cdot \mathrm{KL}[P_{Z|X}(z|x)||Q_Z(z)]]]) \tag{5}$$

where $P_{X,Y}(x, y)$ refers to empirical distribution of complete observations.

Unsupervised VAB Loss As for the unsupervised VAB loss, the only difference from the supervised one is to replace the ground truth label y by the prediction $\hat{y} = f(x)$ given the original research paper x as the input and and it is denoted as follows:

$$\ell_{un} = -(\mathbb{E}_{P_X(x)}[\mathbb{E}_{P_{Z|X}(z|x)}[\log(Q_{Y|Z}(\hat{y}|z)] - \beta \cdot \mathrm{KL}[P_{Z|X}(z|x)||Q_Z(z)]]]) \tag{6}$$

where $P_X(x)$ refers to empirical distribution of incomplete observations.

Total Loss In summary, our full training objective ℓ can be written as follows:

$$\ell = \ell_{su} + \alpha \cdot \ell_{un} \tag{7}$$

where $\alpha > 0$ is a balancing hyper parameter about these two items of losses. Our goal is to minimize the full training objective ℓ.

5 Experimental Setup

The proposed VCCSM method is evaluated with two typical neural network models commonly used on text classification tasks, LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2018) on two datasets.

5.1 Datasets

RRP Dataset RRP dataset is proposed by Luo et al. (2020). RRP dataset contains 399 labeled and 2,170 unlabeled research articles in social science fields. In this paper, randomly selected 300 (150:1;150:0) labeled and 2,170 unlabeled samples are treated as the training dataset. The remaining 99 (51:1;48:0) labeled research articles are considered as the testing dataset. More details about the RRP dataset are shown in Appendix A.2. PDFMiner (Shinyama, 2014) is used to extract the text in the raw pdf files for both labeled and unlabeled datasets. Therefore, the text format of labeled and unlabeled datasets are the same.

ECHR Dataset European Convention of Human Rights (ECHR) (Chalkidis et al., 2019) is a publicly available English legal judgment prediction dataset which contains 11,478 cases. Each case has a list of paragraphs describing the facts. The task is to predict whether one given case is judged as violated or not. The ECHR dataset is split into training, development, and testing datasets with the number of cases of 7,100, 1,380 and 2,998. The average number of tokens for training, development, and testing datasets are 2,421, 1,931, and 2,588, respectively.

5.2 Implementation Details

The LSTM model we used has a bidirectional hidden layer, and it's initialized with 300-dimensional google's pre-trained word embeddings. We fix the embedding layer and update other parameters in LSTM to achieve the best performance. As for BERT model, a published BERT pre-trained model ("bert-base-uncased"[1]) is utilized as the embedding layer of LSTM model. We first use our corpus to pre-train the BERT model and then fine-tune it in the VCCSM classifier's training. In each epoch, the model is first trained on labeled data, followed by unlabeled data. The hidden state of the [CLS] token of the last layer is considered as the sentence representation.

[1] https://huggingface.co/bert-base-uncased

Because the average length (words) of all the documents in the labeled and unlabeled datasets is about 10,000, we set the the maximum length of words in our paper to 10,000. Since VCCSM method is sentence masking and we need to split the text of research paper into sentences. We use period, question mark, and semicolon to conduct the splitting. After some statistical analysis, the average length (words) of each sentence is around 25. For a fair comparison with word masking method, we set the maximum length of sentences in each document to 400. It means that we set the maximum length of words in each document to 10,000 in all models. In the experiments, for RRP dataset, the number of labeled and unlabeled datasets are 4,00 and 2,170 research papers respectively. As for ECHR dataset, 2,000 cases in the training dataset are considered as the labeled and the remaining 5,100 cases as the unlabeled.

5.3 Interpretability Metrics

5.3.1 AOPC

The first interpretability metric we used is area over the perturbation curve (AOPC) (Samek et al., 2016; Nguyen, 2018) which is obtained by computing the average change of prediction probability by deleting top n important words and it can evaluate the model interpretablity on faithness. Since our proposed VCCSM is sentence masking method, we calculate the average change of prediction probability by deleting top n key sentences in the explanations of the papers. Therefore, AOPC used in our paper is defined as follows:

$$\mathrm{AOPC}(f) = \frac{1}{T+1}\sum_{i=1}^{T}\left(f(x_i) - f(x_i \backslash \{s_1, ..., s_n\})\right),$$

where $f(x_i \backslash \{s_1, ..., s_n\})$ is the probability for the predicted class on the i_{th} document in RRP when the top n sentences on importance are removed. Higher AOPC score is better.

5.3.2 Post-hoc Accuracy

The second interpretability metric utilized in this paper is post-hoc accuracy metric (Chen et al., 2018) which is computed by counting how many testing examples' predictions are changed by utilizing only extracted top n words to classify. For our VCCSM models, we used top n key sentences. The formula to calculate the post-hoc accuracy in our paper is as follows:

$$\mathrm{Acc}_p(f, n) = \frac{1}{T}\sum_{i=1}^{T} 1[f(\{s_1, ..., s_n\}) = f(x_i)],$$

where T is the number of examples in the testing dataset, $\{s_1, ..., s_n\}$ are the top n sentences on importance in the i_{th} document. Higher post-hoc accuracy is better.

6 Results

We tested our proposed models on two text classification datasets (RRP along with ECHR), and the details about prediction accuracy and interpretability are described in this section.

6.1 Quantitative Evaluation

We evaluate the interpretability of VCCSM model against other types of models via the AOPC (Samek et al., 2016; Nguyen, 2018) and post-hoc accuracy (Chen et al., 2018) metrics. We also listed the performance with varying number of the unlabeled data in Appendix A.3 and it shows that the performance become higher with more unlabeled data.

Table 1 shows the results of VCCSM (LSTM & BERT) and other interpretable models on the RPP and ECHR datasets with top 500 words (word based methods) or 20 sentences (sentence based methods). Simialr results are obtained with varying number of sentences. For BERT's attention weights model, we extracted the words' attention weights of all heads in the last layer and average them. As for BERT's attention weights (sentences), we average the words' averaged weights in each sentence as its sentence representation. Extractive summarization models can also extract the key sentences for each document. In this section, we used the recent extractive summarization method (Cui and Hu, 2021) as the baseline. We conduct the training on arXiv + PubMed (Cohan et al., 2018) and our labeled + unlabeled datasets (the abstract are the summary). Training on arXiv + PubMed aims to generalize the model and make the model extract a more comprehensive of information instead of only abstract in the research paper. We can observe that our proposed models perform better than other methods in both interpretability and prediction performance on both RRP and ECHR datasets.

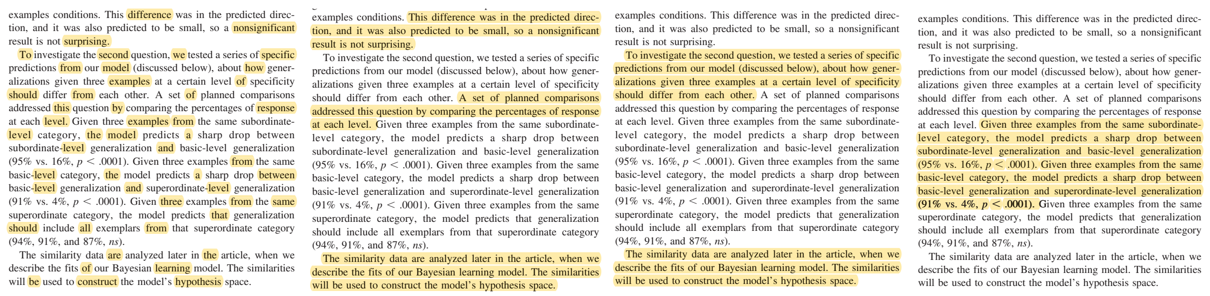
Ablation Study In order to validate different modules in our proposed VCCSM method, we conduct the ablation study on the RRP dataset as shown in Table 2. We observe the drop after removing contextual masking or consistency training (on the unlabeled data) which shows that each component benefit to the model. It is noting that we observe a

Methods	RRP			ECHR		
	Acc	AOPC	Post-hoc	Acc	AOPC	Post-hoc
LSTM Word Masking (Chen and Ji, 2020)	60.61%	11.16%	50.51%	84.86%	10.32%	65.84%
BERT's Attention Weights (words)	64.65%	11.70%	60.61%	84.26%	15.06%	73.75%
BERT Word Masking (Chen and Ji, 2020)	65.66%	12.05%	61.62%	85.06%	16.30%	76.38%
SOTA Extractive Summarization (Cui and Hu, 2021)	65.66%	12.86%	57.58%	85.39%	19.57%	75.52%
BERT's Attention Weights (sentences)	65.66%	13.62%	62.63%	85.39%	22.61%	81.49%
LSTM Sentence Masking + Contextual + Consistency	65.66%	22.19%	63.64%	86.06%	30.53%	84.22%
BERT Sentence Masking + Contextual + Consistency	**68.69%**	**24.02%**	**65.66%**	**87.66%**	**32.78%**	**86.59%**

Table 1: Comparison between VCCSM and other methods on testing accuracy, area over the perturbation curve (AOPC), and post-hoc accuracy on RRP and ECHR datasets.

Model	Methods	Accuracy	AOPC	Post-hoc
LSTM	Proposed LSTM VCCSM	65.66%	22.19%	63.64%
	w/o consistency training	62.63%	14.29%	60.61%
	w/o contextual masking	63.64%	19.10%	62.63%
BERT	Proposed BERT VCCSM	68.69%	24.02%	65.66%
	w/o consistency training	65.66%	16.38%	62.63%
	w/o contextual masking	66.67%	21.16%	64.65%

Table 2: Ablation study of proposed VCCSM (LSTM & BERT Sentence Masking + Contextual + Consistency) on testing accuracy, area over the perturbation curve (AOPC), and post-hoc accuracy on RRP dataset.



BERT Word Masking | Attention Weights (sentences) | Extractive Summarization | BERT VCCSM

Figure 3: Highlighted explanations (words or sentences) of BERT word masking, attention weights (sentences), SOTA extractive summarization, and BERT VCCSM methods for a paragraph in one replicable research paper "Word Learning as Bayesian Inference" in Psychological Review.

larger drop on both accuracy and two interpreatability metrics without the consistency training on the unlabeled data which demonstrates that consistency training contributes more to the model.

6.2 Qualitative Evaluation

In this section, we conduct the qualitative evaluations and compare the explanations of different models intuitively by highlighting the words or sentences. Specifically, we draw on the Open Science pratices (e.g., mentioning how to access the data) as indicators of high reproducibility, because these practices are proposed as solutions to the reproducibility crisis in the science community (Simonsohn et al., 2015; Foster and Deardorff, 2017; Brodeur et al., 2020; Dienlin et al., 2021; Markowitz et al., 2021). Some of those indicators which are easier to check are listed as below: (1) Publish materials, data, and code; (2) Preregister studies and submit the reports; (3) Conduct the replications by themselves; (4) Collaborate with others; (5) P-value[2] is close to 0.5.

We conduct the case studies on the testing dataset and find that our proposed methods can highlight more sentences which are related to the indicators mentioned above. A case study is shown in Figure 3. More specifically, Figure 3 shows highlighted explanations (words or sentences) of BERT word masking, attention weights (sentences), SOTA extractive summarization, and BERT VCCSM methods for a paragraph in one replicable research paper "Word Learning as Bayesian Inference" (Xu and Tenenbaum, 2007) in Psychological Review. In this case study, we extracted top 200 sentences or 5,000 words (only

[2] Probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

for BERT word masking method) but only show one paragraph highlighted results. Although all the methods provide the correct prediction, our VCCSM highlights the sentences which are related to the indicators described above. It is noting that the highlight words of BERT word masking is not so readable for the long research paper. Attention weights (sentences) and SOTA extractive summarization methods can provide informational sentences but the highlighted sentence are not related to the indicators described above. BERT VCSSM can highlight p-value sentences which are related to the indicators mentioned above.

6.3 Discussion on Plausibility of Predicting Research Replicability using Text

By looking into RRP's labeled dataset and conducting the cases studies carefully such as in Figure 3, we discuss on whether classifying results in a research paper as replicable using text is actually sufficient to replicate the results, which is the central premise this paper is based on. Non-replicability of scientific studies largely results from unscientific, unethical research practices (e.g., p-hacking, selective reporting, data manipulation). Such practices can be manifested in the texts of research papers such as the reports of p-values, experimental procedures, etc. Generally speaking, the more problematic practices a research paper involves, the less likely its findings are valid, and the less likely it will be reproduced. Hence, by modeling the replicability of research paper with regard to its textual components that are potentially linked with the problematic practices, we can classify whether a research paper can be replicated and identify the focal sentences relevant to the prediction.

7 Concluding Remarks

In this paper, we proposed VCCSM to improve both interpretability and prediction accuracy on RRP along with ECHR datasets, using largely unlabeled datasets. We tested VCCSM with two different neural text classifiers (LSTM and BERT) and evaluated both prediction accuracy and interpretability metrics. As future work, we plan to explore other advanced interpretable models and weakly supervised methods to further improve the prediction performance and interpretability of long document classification tasks.

Acknowledgments

This research is partially supported by National Science Foundation (NSF) under Grant No. IIS-2007951 and the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC Pacific) under Contract No. N66001-19-C-4014. The authors would like to thank Dr. Bingjie Liu for her valuable comments on the manuscript. We also thank all the anonymous reviewers for their insightful comments.

8 Broader Impacts

Our paper proposed VCCSM method to build an interpretable model for long document datasets such as RRP and ECHR. Our model can provide the explanations about why a research paper is predicted as replicable or non-replicalbe and why a case is judged as violated or not so that the prediction results obtained by neural text classifier are more reliable and trustworthy. However, sometimes, our proposed methods can be misused. For example, people may try to adversarially write the new text in a research paper to fool the research replication prediction tool when they can obtain the explanations by using our interpretable models. Therefore, the proposed methods in this paper should be used with careful consideration of its potential misusing when deployed in the real-world.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2019. Predicting the replicability of social science lab experiments. *PloS one*, 14(12).

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*.

Abel Brodeur, Nikolai Cook, and Anthony Heyes. 2020. Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60.

Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan

Chan, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Hanjie Chen, Song Feng, Jatin Ganhotra, Hui Wan, Chulaka Gunasekara, Sachindra Joshi, and Yangfeng Ji. 2021. Explaining neural network predictions on sentence pairs via learning word-group masks. *arXiv preprint arXiv:2104.04488*.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. *arXiv preprint arXiv:2010.00667*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Open Science Collaboration. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6):657–660.

Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Peng Cui and Le Hu. 2021. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5881–5891.

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

Tobias Dienlin, Niklas Johannes, Nicholas David Bowman, Philipp K Masur, Sven Engesser, Anna Sophie Kümpel, Josephine Lukito, Lindsey M Bier, Renwen Zhang, Benjamin K Johnson, et al. 2021. An agenda for open science in communication. *Journal of Communication*, 71(1):1–26.

A Dreber, T Pfeiffer, E Forsell, D Viganola, M Johannesson, Y Chen, B Wilson, BA Nosek, and J Almenberg. 2019. Predicting replication outcomes in the many labs 2 study. *Journal of Economic Psychology*.

Charles R Ebersole, Olivia E Atherton, Aimee L Belanger, Hayley M Skulborstad, Jill M Allen, Jonathan B Banks, Erica Baranski, Michael J Bernstein, Diane BV Bonfiglio, Leanne Boucher, et al. 2016. Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82.

Erin D Foster and Ariel Deardorff. 2017. Open science framework (osf). *Journal of the Medical Library Association: JMLA*, 105(2):203.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. *arXiv preprint arXiv:1611.07634*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. 2014. Investigating variation in replicability. *Social psychology*.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. 2018. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Tianyi Luo, Xingyu Li, Hainan Wang, and Yang Liu. 2020. Research replication prediction using weakly supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1464–1474.

David M Markowitz, Hyunjin Song, and Samuel Hardman Taylor. 2021. Tracing the adoption and effects of open science in communication research. *Journal of Communication*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078.

H Pashler, B Spellman, S Kang, and A Holcombe. 2019. Psychfiledrawer: archive of replication attempts in experimental psychology. *Online< http://psychfiledrawer. org/view_ article_list. php*.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Yusuke Shinyama. 2014. Pdfminer.

Daniel J Simons, Alex O Holcombe, and Barbara A Spellman. 2014. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5):552–555.

Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. 2015. Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to ulrich and miller (2015).

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

Chandan Singh, W James Murdoch, and Bin Yu. 2018. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.

Yang Yang. 2018. The replicability of scientific findings using human and machine intelligence. https://www.metascience2019.org/presentations/yang-yang/ Metascience 2019.

A Appendix

A.1 Detailed Derivation of Lower Bound for Variational Contextual Sentence Masking in Section 4.2

In this section, we provided the complete details on the derivation of lower bound for variational contextual sentence masking in Section 4.2.

Assuming that the true joint distribution is $P(X, Y, Z)$ and X, Y, Z are random variables which have the following conditional dependency: $Y \leftrightarrow X \leftrightarrow Z$. And x, y, z are instances of ramdom variables. We can have

$$\begin{aligned} P(X, Y, Z) &= P(Z|X, Y)P(Y|X)P(X) \\ &= P(Z|X)P(Y|X)P(X). \end{aligned} \tag{8}$$

According to the definition of $I(Z;Y)$, we have

$$I(Z;Y) = \sum_{z,y} P_{Z,Y}(z,y) \log \frac{P_{Z,Y}(z,y)}{P_Z(z)P_Y(y)} = \sum_{z,y} P_{Z,Y}(z,y) \log \frac{P_{Y|Z}(y|z)}{P_Y(y)}. \quad (9)$$

And we also have

$$\begin{aligned} P_{Y|Z}(y|z) &= \sum_x P_{X,Y|Z}(x,y|z) \\ &= \sum_x P_{Y|X}(y|x)P_{X|Z}(x|z) \\ &= \sum_x \frac{P_{Y|X}(y|x)P_{Z|X}(z|x)P_X(x)}{P_Z(z)}. \end{aligned} \quad (10)$$

Since $P(Y|Z)$ can be intractable, $Q(Y|Z)$ is considered as a variational approximation to $P(Y|Z)$. $Q(Y|Z)$ is our decoder and a neural network. Because the Kullback Leibler divergence is non-negative, we have

$$\begin{aligned} &\mathrm{KL}[P(Y|Z)||Q(Y|Z)] \geq 0 \\ \Rightarrow &\sum_y p(y|z)\log p(y|z) \geq \sum_y p(y|z) \log q(y|z). \end{aligned} \quad (11)$$

Therefore, we can obtain the lower bound of $I(Z;Y)$ as follows:

$$\begin{aligned} I(Z;Y) &\geq \sum_{z,y} P_{Z,Y}(z,y) \log \frac{Q_{Y,Z}(y|z)}{P_Y(y)} \\ &= \sum_{z,y} P_{Z,Y}(z,y) \log Q_{Y|Z}(y|z) + H(Y). \end{aligned} \quad (12)$$

where $H(Y) = -\sum_y P_Y(y) \log P_Y(y)$ is entropy. According to Equation 8, we have

$$\begin{aligned} P(Y|Z) &= \sum_x P_{X,Y,Z}(x,y,z) \\ &= \sum_x P_{X,Y,Z}(x,y,z) \\ &= \sum_x P_X(x)P_{Y|X}(y|x)P_{Z|X}(z|x). \end{aligned} \quad (13)$$

Hence, we obtain the lower bound of $I(Z,Y)$ as follows:

$$\sum_{x,y,z} P_X(x)P_{Y|X}(y|x)P_{Z|X}(z|x) \log Q_{Y|Z}(y|z).$$

As for $I(Z;X)$, similar to Equation 9 in the derivation of $I(Z;Y)$, we first obtain

$$\begin{aligned} I(Z;X) &= \sum_{z,x} P_{Z,X}(z,x) \log \frac{P_{Z|X}(z|x)}{P_Z(z)} \\ &= \sum_{z,x} P_{Z,X}(z,x) \log P_{Z|X}(z|x) \\ &\quad - \sum_z P_Z(z) \log P_Z(z). \end{aligned} \quad (14)$$

Because the marginal distribtuion of Z, $P(Z) = \sum_x P_{Z|X}(z|x)P_X(x)$ in which the computation might be difficult, we replace $P(Z)$ by a variational approximation of $Q(Z)$. Since $\mathrm{KL}[P(Z)||Q(Z)] \geq 0 \Rightarrow \sum_z P_Z(z) \log P_Z(z) \geq \sum_z P_Z(z) \log Q_Z(z)$, we can get the upper bound of $I(Z;X)$ as follows:

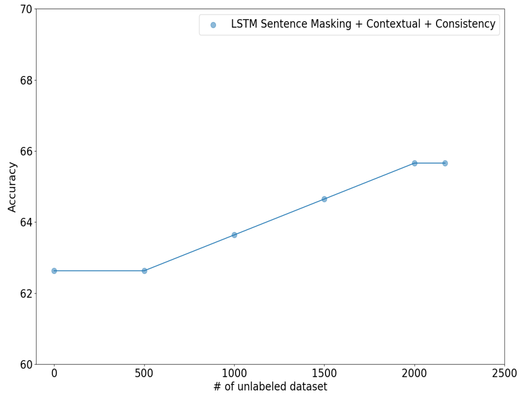
$$\begin{aligned} I(Z;X) &\leq \sum_{z,x} P_{Z,X}(z,x) \log P_{Z|X}(z|x) \\ &\quad - \sum_{z,x} P_{Z,X}(z,x) \log Q_Z(z) \\ &\leq \sum_{z,x} P_X(x)P_{Z|X}(z|x) \log \frac{P_{Z|X}(z|x)}{Q_Z(z)}. \end{aligned} \quad (15)$$

Combining Equation 12 and 15, we can get the lower bound of $I(Z;Y) - \beta I(Z;X)$ as follows:

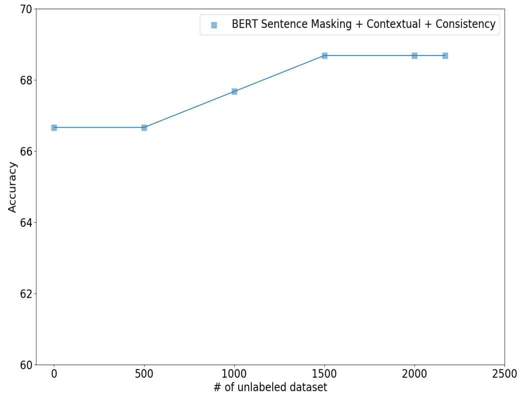
$$\begin{aligned} &\sum_{x,y,z} P_X(x)P_{Y|X}(y|x)P_{Z|X}(z|x) \log Q_{Y|X}(y|z) \\ &- \beta \sum_{z,x} P_X(x)P_{Z|X}(z|x) \log \frac{P_{Z|X}(z|x)}{Q_Z(z)}. \end{aligned}$$

A.2 Details of RRP Dataset

In the RRP dataset proposed by Luo et al. (2020), the labeled datset are collected from eight research replication projects which are the Registered Replication Report (RRR) (Simons et al., 2014), Many Labs 1 (Klein et al., 2014), Many Labs 2 (Klein et al., 2018), Many Labs 3 (Ebersole et al., 2016), Social Sciences Replication Project (SSRP) (Camerer et al., 2018), PsychFileDrawer (Pashler et al., 2019), Experimental Economics Replication Project (Camerer et al., 2016), and Reproducibility Project: Psychology (RPP) (Collaboration, 2012). Among 399 labeled data in the RRP dataset, 201 are labeled as ‘1’ (replicable) and the remain 198 are annotated as ‘0’ (non-replicable). We observe that the labeled data in the RRP dataset is balanced.



(a) LSTM Sentence Masking + Contextual + Consistency



(b) BERT Sentence Masking + Contextual + Consistency

Figure 4: Testing accuracy (%) on RRP dataset with varying number of unlabeled dataset for VCCSM applied on two neural text classifiers (LSTM and BERT)

In addition, RRP dataset also contains 2,170 research articles as the unlabeled dataset. Luo et al. (2020) observed that most papers in the labeled dataset in the RRP dataset are economical and psychology related. Among those papers, they are mainly from American Economic Review and Psychological Science journals. Therefore, a python crawler is written by Luo et al. (2020) to get 2,170 published research articles on the American Economic Review (Jan 2011 - Dec 2014) and Psychological Science websites (Jan 2006 - Dec 2012). The number of articles crawled from American Economic Review and Psychological Science websites are 981 and 1,189 respectively.

A.3 Performance with Varying Number of Unlabeled Data

We conducted the experiments to test our model's effectiveness by varying number of unlabeled data for VCCSM applied on two neural text classifiers (LSTM and BERT). From Figure 4, we can observe that, with more unlabeled data, the testing accuracy become higher on both LSTM Sentence Masking + Contextual + Consistency and BERT Sentence Masking + Contextual + Consistency models, which validates the effectiveness of using unlabeled data.