CDA: A Contrastive Data Augmentation Method for Alzheimer's Disease Detection

Junwen Duan[1], Fangyuan Wei[1], Hongdong Li[1], Tianming Liu[2], Jianxin Wang[1], Jin Liu[1]*

[1]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University

[2]School of Computing, The University of Georgia

{jwduan,weify9,hongdong,liujin06}@csu.edu.cn, jxwang@mail.csu.edu.cn

tliu@cs.uga.edu

Abstract

Alzheimer's Disease (AD) is a neurodegenerative disorder that significantly impacts a patient's ability to communicate and organize language. Traditional methods for detecting AD, such as physical screening or neurological testing, can be challenging and time-consuming. Recent research has explored the use of deep learning techniques to distinguish AD patients from non-AD patients by analysing the spontaneous speech. These models, however, are limited by the availability of data. To address this, we propose a novel contrastive data augmentation method, which simulates the cognitive impairment of a patient by randomly deleting a proportion of text from the transcript to create negative samples. The corrupted samples are expected to be in worse conditions than the original by a margin. Experimental results on the benchmark ADReSS Challenge dataset demonstrate that our model achieves the best performance among language-based models[1].

1 Introduction

Alzheimer's Disease (AD) is a debilitating neurodegenerative disorder characterized by a progressive cognitive decline that is currently incurable. It accounts for up to 70% of all cases of dementia (Association, 2020). With an aging population, the prevalence of AD is on the rise. As symptoms of Alzheimer's disease can be mistaken for a variety of other cognitive disorders, traditional diagnostic methods, such as physical screening or neurological testing, can be challenging and time-consuming. Furthermore, they require a certain degree of clinician expertise (Prabhakaran et al., 2018).

Consequently, the development of automatic detection methods for Alzheimer's disease is essential to the advancement of current medical treatment. The use of machine learning methods to detect AD or other diseases automatically has gained increasing attention in recent years (Luz et al., 2018; Martinc and Pollak, 2020; Liu et al., 2021; Yu et al., 2023). Nevertheless, these approaches have limitations due to a lack of data and the generalizability of the models. Some studies have attempted to address this problem by model ensembling (Syed et al., 2021; Rohanian et al., 2021), multi-task learning (Li et al., 2022; Duan et al., 2022) or data augmentation (Woszczyk et al., 2022), but the improvement in performance is not always substantial.

Inspired by previous research that AD patients often have language disorders, such as difficulties in word finding and comprehension (Rohanian et al., 2021), we propose a novel Contrastive Data Augmentation (CDA) approach for automatic AD detection. In our study, we simulated cognitive decline associated with Alzheimer's disease by randomly deleting words from the speech transcript to create negative samples. It is expected that the corrupted samples are in worse condition than the original due to the degradation of coherence and semantic integrity. Compared to traditional data augmentation methods, the CDA method expands the dataset scale and utilizes augmented data more effectively. We have demonstrated in our experiments on the ADReSS Challenge dataset that our approach uses linguistic features alone, is more generalizable to unseen data, and achieves superior results compared to strong baselines.

2 Data and Preprocessing

We use the data from the ADReSS Challenge (Alzheimer's Dementia Recognition through Spontaneous Speech) (Luz et al., 2020), a subset of the DementiaBank's English Pitt Corpus (Becker et al., 1994). It consists of recordings and transcripts of spoken picture descriptions from the Boston Diagnostic Aphasia Examination. During the examination, the subject is shown a picture and is asked to describe its content in their own language.

*Corresponding author

[1]Our code is publicly available at https://github.com/CSU-NLP-Group/CDA-AD.

Findings of the Association for Computational Linguistics: ACL 2023, pages 1819–1826

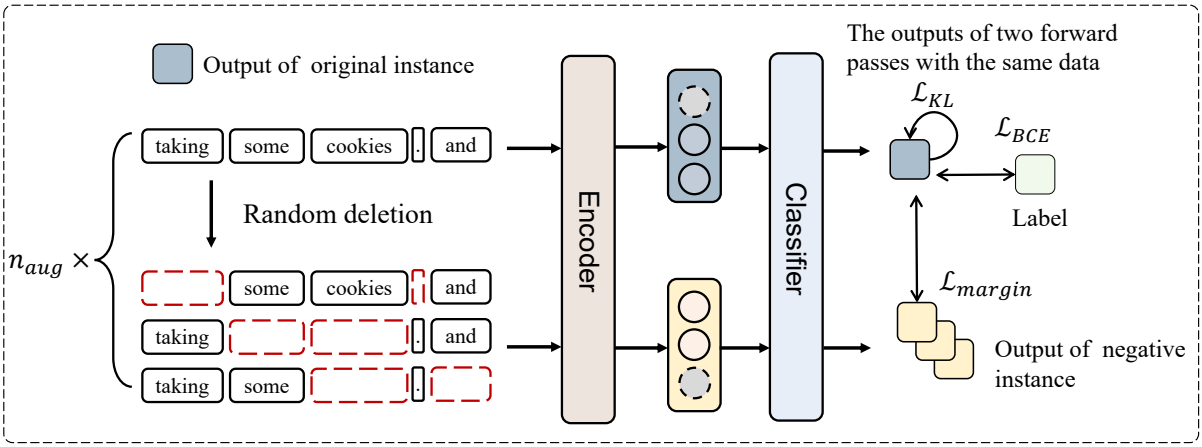

Figure 1: The overview of our proposed method.

A total of 156 speech audio recordings and transcripts were obtained from English-speaking participants in the ADReSS dataset, with an equal number of participants (N=78) diagnosed with and not suffering from Alzheimer's disease, as shown in Table 1. Annotated transcripts in the dataset are in CHAT format (MacWhinney, 2014). Participants' ages and genders are also balanced to minimize the risk of bias in prediction. As some of the tokens in CHAT format are highly specific and are unlikely to be included in BERT tokenizers, we converted them into actual repetitions of words. We remain with only words, punctuation, and pauses for input into the BERT model. Our method uses only the transcripts from the dataset.

	AD		non-AD	
	M	F	M	F
Train	24	30	24	30
Test	11	13	11	13
Total	35	43	35	43

Table 1: Statistics of ADReSS Dataset

3 Methods

Figure 1 illustrates the framework of the proposed model. Firstly, for each transcript, we generate a number of augmented instances, which are then input to Text Encoder along with the original transcripts to obtain their corresponding representations. Then the classifier uses feature vectors acquired in Text Encoder and output a probability of being AD for each transcript and its corresponding augmented samples. We will discuss more details in the following subsections.

3.1 Text Encoder and Classifier

For fair comparisons with previous work (Woszczyk et al., 2022), the input text is encoded using the pre-trained BERT (bert-base-uncased) and represented by `[CLS]` after `bert_pooler`. Given a text sequence $\boldsymbol{x}_i$, we can get the encoded representations $\boldsymbol{h}_i$ through the encoder.

$$\boldsymbol{h}_i = \text{BERT}(\boldsymbol{x}_i) \tag{1}$$

After obtaining the embedding of the transcript, we pass it through a simple linear classifier (Eq. 2) to get final prediction scores, we use the commonly used binary cross-entropy (BCE) as our classification loss function, and the classification loss is denoted as $\mathcal{L}_{BCE}$ (Eq. 3).

$$\hat{y}_i = \sigma(\boldsymbol{W}\boldsymbol{h}_i + b) \tag{2}$$

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

, where y_i is the golden label for $\boldsymbol{x}_i$, $\boldsymbol{W}$ and b are trainable parameters in classifier.

3.2 Contrastive Data Augmentation

The performance of previous work is limited due to a lack of data availability. To alleviate this, we propose the contrastive data augmentation approach (CDA) to replicate the cognitive decline associated with AD to expand the data size and improve the model robustness.

Negative Sample Generation Assuming that the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ contains N training samples. We randomly delete a proportion of $p \in [0, 1]$

words from each sample for n_{neg} times to create n_{neg} negative samples. After that we can get an augmented set $\{\boldsymbol{x}_i, y_i, \mathcal{X}^i_{neg}\}^N_{i=1}$, where $\mathcal{X}^i_{neg} = \{\tilde{\boldsymbol{x}}^j_i\}^{n_{neg}}_{j=1}$ are from $\boldsymbol{x}_i$. We can further augment the training set by repeating the whole process for n_{aug} times to get $\{\boldsymbol{x}_i, y_i, \mathcal{X}^i_{neg}\}^{N\times n_{aug}}_{i=1}$ and expand the data size by n_aug.

Positive Sample Generation Inspired by Gao et al. (2021), we resort to the randomness of dropout to construct positive samples. Dropout is a popular regularization technique due to its simplicity, but the randomness it introduces may hinder further improvements in the model's generalization performance. R-Drop (Wu et al., 2021) is proposed to fix the aforementioned problem by ensuring consistency between the outputs of two forward-pass with the same data. We deploy the R-Drop algorithm as a regularization method for generating positive instances. More specifically, the original sample $\boldsymbol{x}_i$ is fed to the model twice at each step, and two corresponding predictions, denoted as $\hat{y}^1_i$ and $\hat{y}^2_i$, are obtained. Then we try to minimize the bidirectional Kullback-Leibler (KL) divergence between them, which is denoted as $\mathcal{L}_{KL}$ (Eq. 4):

$$\mathcal{L}_{KL} = \sum_{i=1}^{N} \frac{1}{2}[\mathcal{D}_{KL}(\hat{y}^1_i \| \hat{y}^2_i) + \mathcal{D}_{KL}(\hat{y}^2_i \| \hat{y}^1_i)] \quad (4)$$

Contrastive Loss It is reasonable to assume that the negative samples are more likely to have AD than the original ones in view of the degradation in semantic coherence and integrity. To achieve this, we regularize their differences to be larger than a margin m.

Particularly, the encoder receives $\boldsymbol{x}_i$ and $\mathcal{X}^i_{neg}$ as input and outputs their corresponding embedding representations $\boldsymbol{h}_i$ and $\mathcal{H}^i_{neg}$. Then, their representations are fed to the classifier to get a final score $\hat{y}_i$ and $\tilde{y}^j_i$ for $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}^j_i$, respectively. Their differences becomes Eq. 5:

$$\mathcal{L}_{margin} = \sum_{i=1}^{N} max(0, m - \hat{y}_i + \frac{\sum_{j=1}^{n_{neg}} \tilde{y}^j_i}{n_{neg}}) \quad (5)$$

, where m is the margin between positive and negative samples. The final loss is a combination of the above three loss terms $\mathcal{L}_{BCE}$, $\mathcal{L}_{margin}$ and $\mathcal{L}_{KL}$.

$$\mathcal{L} = \mathcal{L}_{BCE} + \alpha\mathcal{L}_{margin} + \mu\mathcal{L}_{KL} \quad (6)$$

, where α and μ are hyperparameters that control the impact of positive and negative samples, and we set $\alpha = 0.5$ and $\mu = 0.5$ in our model.

4 Experiments

We employ 10-fold cross-validation to estimate the generalization error and adjust the model's parameter settings. The best setting is used to retrain models on the whole train set with five different random seeds and is then applied to the test set. The results reported in this paper are the average of these models. The accuracy is used as the primary metric of task performance since the dataset is balanced. Recall, precision, and F1 are also reported for the AD class to provide a more comprehensive assessment. The hyperparameters in our model are: learning rate=1e-04, batch size=8, epoch=5, n_{aug}=3, n_{neg}=3, p=0.3, $margin$=0.1.

4.1 Baselines

We compare our method with: 1) LDA, which is the challenge baseline linear discriminant analysis (LDA) (Luz et al., 2020); 2) BERT, Balagopalan et al. (2021) compared BERT models with feature-based Models and obtained relatively better results using the former; 3) Fusion, Campbell et al. (2021) fused the features of language and audio for classification; 4) SVM(BT RU)(Woszczyk et al., 2022), is the SVM model using Back-translation from Russian that achieves the best results over the BERT model using Back-translation from German (BT DE); 5) Ensemble methods, Sarawgi et al. (2020) take a majority vote between three individual models. ERNIE0p and ERNIE3p are based on ERNIE-large (Sun et al., 2020) that use original transcripts and transcripts with pauses manually inserted for AD classification, respectively.

4.2 Results

The main experimental results are shown in Table 2. We can observe that the performance significantly improves when BERT is applied. Back-translation data augmentation results in consistent improvements in both BERT (BT DE) and SVM (BT RU), suggesting that data argumentation is a promising strategy. Our method achieves accuracy (87.5%), precision (88.1%), and F1 score (86.9%), outperforming the baseline method by a substantial margin, suggesting the effectiveness of cognitive impairment simulation in our method. By ensembling our models on five models with a majority vote mechanism, the performance improves significantly (4.2% absolute improvements in accuracy and 4% absolute improvements in F1 score, respectively) and achieves the best results among all

Methods	Accuracy%	Precision%	Recall%	F1%
LDA (Luz et al., 2020)	75.0	83.0	62.0	71.0
BERT (Balagopalan et al., 2021)	83.3	83.9	83.3	83.3
Fusion (Campbell et al., 2021)	83.3	80.1	**87.5**	84.0
BERT(BT DE) (Woszczyk et al., 2022)	84.0	–	75.0	–
SVM(BT RU) (Woszczyk et al., 2022)	85.0	–	79.0	–
CDA (single-model, ours)	**87.5**	**88.1**	83.3	**86.9**
Ensemble Methods				
Sarawgi et al. (2020)	83.0	83.0	83.0	83.0
ERNIE0p (Yuan et al., 2020)	85.4	94.7	75.0	83.7
ERNIE3p (Yuan et al., 2020)	89.6	95.2	**83.3**	88.9
CDA (ensembled, ours)	**91.7**	**100.0**	**83.3**	**90.9**

Table 2: Results of our method and the baselines on the test set.

methods, outperforming even ERINE, a larger and knowledge-richer pre-trained model.

4.3 Ablation Study

To determine the effectiveness of the main modules, namely random deletion (RD) and regularized dropout (R-Drop), we removed them from the model one by one and tested their impact on performance in 10-fold cross-validation.

Methods	Accuracy%	Recall%
BERT	72.3	71.9
CDA (ours)	**77.5**	75.2
- *w/o* RD	72.3	74.2
- *w/o* R-Drop	76.7	**76.5**

Table 3: Result of the AD classification in cross-validation. The average accuracy(%) and recall(%) is reported. *w/o* RD indicates CDA without the random deletion module and *w/o* R-Drop indicates CDA without the R-Drop module.

As shown in Table 3, by combining the contrastive data augmentation strategy with the base BERT, our model outperforms it by a large margin. However, when either module is removed, the model experiences a significant loss of performance, suggesting their positive contributions to the performance.

4.4 Parameter Analysis

We also perform parameter analysis under the same experimental settings. As illustrated in Figure 2, we can see that a lower deletion rate leads to relatively higher accuracy, as the more words deleted, the less informative the transcript is. But a large margin negatively impacts both recall and accuracy. As for n_{aug}, the model performs better regarding recall and accuracy when it is set to 3, and lower or higher values will affect the performance. The same conclusion applies to n_{neg}, where a breakdown of the model is observed when n_{neg}=7. The model performance also improves as the number of negative samples increases. However, this will take more computing resources.

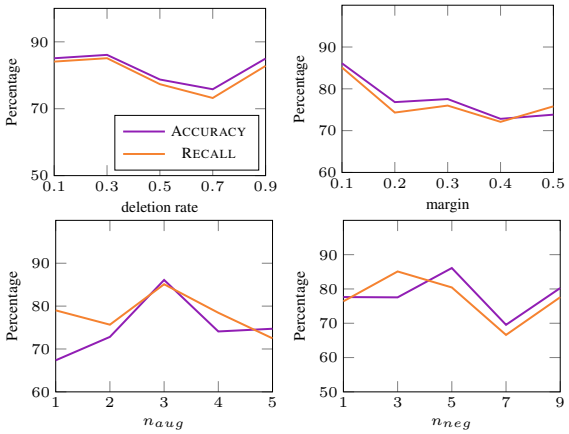

Figure 2: Accuracy and recall scores at different deletion rate, margin, n_{aug} and n_{neg}.

5 Conclusion

Our experiments show the potential of contrastive data argumentation in improving the accuracy of models for Alzheimer's disease diagnosis. As a comparison to large, complex multimodal models, and other techniques of data augmentation, we obtain the best results by simulating cognitive impairment caused by AD. Despite the small size of the dataset, the results of this study provide a basis for further research into more complex issues.

Limitations

The limitation of our study is that we only evaluated our model on a limited set of spoken language transcripts. We believe that additional attention should be given to features specific to AD patients, such as pauses and filler words in speech.

Furthermore, the lack of diversity in the data may also adversely affect the model's performance on unseen samples. Our model would benefit from further testing on a wider range of data, including different languages and different modalities, to see if it is capable of generalizing to other domains in the future.

Ethics Statement

The dataset we use in this paper is from the public ADReSS challenge, which contains the minimum amount of personal information and restricts unauthorized access. Data usage and data sharing for ADReSS data has been conducted in accordance with the Ground Rules and Code of Ethics. Furthermore, it is important to note that the study does not include all possible diagnoses of Alzheimer's disease since it is based on transcript text data from an English-speaking cultural context. As this model was designed primarily for academic research, it is unlikely to provide a valid diagnosis in every situation and will be risky if applied to real-world clinical diagnosis situations.

Acknowledgements

We thank anonymous reviewers for their helpful feedback. We thank Jiang Han and Guo Huai for their initial review and feedback for the earlier version of the paper. This work is supported in part by National Natural Science Foundation of China (Grant No.62172444, 62006251, U22A2041) and the Natural Science Foundation of Hunan Province (No. 2021JJ40783), Central South University Innovation-Driven Research Programme under Grant 2023CXQD018. We are grateful for resources from the High Performance Computing Center of Central South University.

References

Alzheimers Association. 2020. What is dementia? https://www.alz.org/alzheimers-dementia/what-is-dementia, Last accessed on 2023-01-03.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Edward L Campbell, Laura Docío Fernández, Javier Jiménez Raboso, and Carmen García-Mateo. 2021. Alzheimer's dementia detection from audio and language modalities in spontaneous speech. In *IberSPEECH*.

Junwen Duan, Huai Guo, Min Zeng, and Jianxin Wang. 2022. Asnet: An adversarial sparse network for multi-task biomedical named entity recognition. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 416–421.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. *arXiv preprint arXiv:2203.13397*.

Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. 2021. Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7323–7327. IEEE.

Saturnino Luz, Sofia de la Fuente, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk*. Psychology Press.

Matej Martinc and Senja Pollak. 2020. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. In *INTERSPEECH*, pages 2157–2161.

Gokul Prabhakaran, Rajbir Bakshi, et al. 2018. Analysis of structure and cost in a longitudinal study of alzheimer's disease. *Journal of Health Care Finance*, 44(3).

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech. *arXiv preprint arXiv:2106.09668*.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. 2020. Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity. In *Proc. Interspeech 2020*, pages 2212–2216.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. 2021. Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling. *IEEE Access*, 9:88377–88390.

Dominika Woszczyk, Anna Hedlikova, Alican Akman, Soteris Demetriou, and Björn Schuller. 2022. Data Augmentation for Dementia Detection in Spoken Language. In *Proc. Interspeech 2022*, pages 2858–2862.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Ying Yu, Junwen Duan, and Min Li. 2023. Fusion model for tentative diagnosis inference based on clinical narratives. *Tsinghua Science and Technology*, 28(4):686–695.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *INTERSPEECH*, volume 2020, pages 2162–6.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitaion.

☑ A2. Did you discuss any potential risks of your work?
Section Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2.

☑ B1. Did you cite the creators of artifacts you used?
Section 2.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section Ethics Statement.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section Ethics Statement.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 2.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.

C ☑ Did you run computational experiments?

Section 3.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Because the data size is small and the overall computational expenditure is minimal.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.