DopplerBAS: Binaural Audio Synthesis Addressing Doppler Effect

Jinglin Liu[*1] , **Zhenhui Ye**[*1]**, Qian Chen**[2]**, Siqi Zheng**[2]**, Wen Wang**[2]**, Qinglin Zhang**[2]
Zhou Zhao[1]
[1]Zhejiang University
[2]Speech Lab of DAMO Academy, Alibaba Group

Abstract

Recently, binaural audio synthesis (BAS) has emerged as a promising research field for its applications in augmented and virtual realities. Binaural audio helps users orient themselves and establish immersion by providing the brain with interaural time differences reflecting spatial information. However, existing BAS methods are limited in terms of phase estimation, which is crucial for spatial hearing. In this paper, we propose the **DopplerBAS** method to explicitly address the Doppler effect of the moving sound source. Specifically, we calculate the radial relative velocity of the moving speaker in spherical coordinates, which further guides the synthesis of binaural audio. This simple method introduces no additional hyperparameters and does not modify the loss functions, and is plug-and-play: it scales well to different types of backbones. DopplerBAS distinctly improves the representative WarpNet and BinauralGrad backbones in the phase error metric and reaches a new state of the art (SOTA): 0.780 (versus the current SOTA 0.807). Experiments and ablation studies demonstrate the effectiveness of our method.

1 Introduction

Binaural audio synthesis (BAS), which aims to render binaural audio from the monaural counterpart, has become a prominent technology in artificial spaces (e.g. augmented and virtual reality) (Richard et al., 2021, 2022; Leng et al., 2022; Lee and Lee, 2022; Parida et al., 2022; Zhu et al., 2022; Park and Kim, 2022). Binaural rendering provides users with an immersive spatial and social presence (Hendrix and Barfield, 1996; Gao and Grauman, 2019; Huang et al., 2022; Zheng et al., 2022), by producing stereophonic sounds with accurate spatial information. Unlike traditional single channel audio synthesis (van den Oord et al., 2016; Chen et al., 2021), BAS places more emphasis on

accuracy over sound quality, since humans need to interpret accurate spatial clues to locate objects and sense their movements consistent with visual input (Richard et al., 2021; Lee et al., 2022).

Currently, there are three types of neural networks (NN) to synthesize binaural audio. Firstly, Richard et al. (2021) collects a paired monaural-binaural speech dataset and provides an end-to-end baseline with geometric and neural warping technologies. Secondly, to simplify the task, Leng et al. (2022) decompose the synthesis into a two-stage paradigm: the common information of the binaural audio is generated in the first stage, based on which the binaural audio is generated in the second stage. They also propose to use the generative model DDPM (Ho et al., 2020) to improve the audio naturalness. Thirdly, to increase the generalization capability for the out-of-distribution audio, Lee and Lee (2022) renders the speech in the Fourier space. These non-linear NN-based methods outperform the traditional digital signal processing systems based on a linear time-invariant system (Savioja et al., 1999; Zotkin et al., 2004; Sunder et al., 2015).

However, these NN methods still have room for improvement in accuracy, especially phase accuracy. Richard et al. (2022) claims that the correct phase estimation is crucial for binaural rendering [1]. Actually, the previous works tend to view the scene "statically", and only take into account the series of positions and head orientations. This motivates us to propose **DopplerBAS**, which facilitates phase estimation by explicitly introducing the Doppler effect (Gill, 1965; Giordano, 2009) into neural networks. Specifically, 1) we calculate the 3D velocity vector of the moving sound source in the Cartesian coordinates and then decompose this 3D velocity vector into a velocity vector in the spherical coor-

---

[1]Our ears can discriminate interaural time differences as short as 10μs (Brown and Duda, 1998; Richard et al., 2021; johansson et al., 2022).

dinates relative to the listener; 2) According to the Doppler effect, we use the radial relative velocity as an additional condition of the neural network, to incentivize the model to sense the moving objects. We also analyze the efficacy of different types of velocity conditions through extensive experiments.

Naturally, DopplerBAS can be applied to different neural binaural renderers without tuning hyperparameters. We pick two typical recent backbones to demonstrate the effectiveness of our method: 1) WarpNet (Richard et al., 2021), a traditional neural network optimized by reconstruction losses; 2) BinauralGrad (Leng et al., 2022), a novel diffusion model optimized by maximizing the evidence bound of the data likelihood. Experiments on WarpNet and BinauralGrad are representative and could show the generalizability of our proposed DopplerBAS on other conditions based on gains on these two models. The contributions of this work can be summarized as follows:

- We propose DopplerBAS, which distinctly improves WarpNet and BinauralGrad in the phase error metric and produces a new state of the art performance: 0.780 (vs. the current state of the art 0.807).

- We conduct analytical experiments under various velocity conditions and discover that: 1) NN does not explicitly learn the derivative of position to time (velocity); 2) The velocity condition is beneficial to binaural audio synthesis, even the absolute velocity in the Cartesian coordinates; 3) The radial relative velocity is the practical velocity component, which obeys the theory of the Doppler effect.

2 Method

In this work, we focus on the most basic BAS scenario where only the monaural audio, the series of positions and head orientations are provided (Richard et al., 2022; Leng et al., 2022), rather than other scenarios where extra modalities (Xu et al., 2021) are present. Note that scenarios with extra modalities present are different tasks. Also, as demonstrated in this paper, our proposed DopplerBAS is plug-and-play and can be easily integrated into other more complex scenarios. In this section, we will introduce the Doppler Effect as the preliminary knowledge, and then introduce the proposed method DopplerBAS. We will describe how to calculate and decompose the velocity vec-

tor, and how to apply this vector to two different backbones.

2.1 Doppler Effect

The Doppler effect (Gill, 1965) is the change in frequency of a wave to an observer, when the wave source is moving relative to it. This effect is originally used in radar systems to reveal the characteristics of interest for the target moving objects (Chen et al., 2006). It can be formulated as:

$$f = \left(\frac{c}{c \pm v_r} \right) f_0,$$

(1)

where c, v_r, f_0 and f are the propagation speed of waves, the radial relative velocity of the moving sound source, the original frequency of waves and the received frequency of waves, respectively.
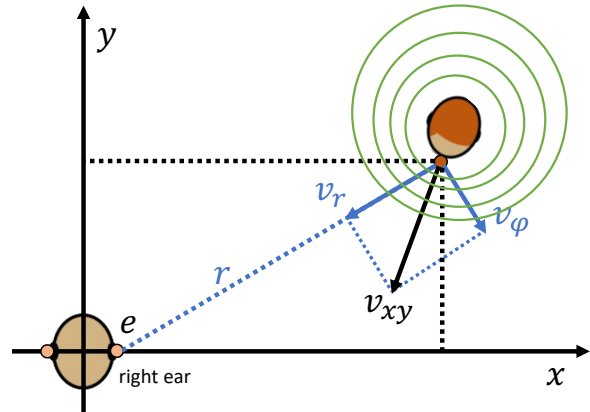

Figure 1: We illustrate the top view where the height dimension is omitted for simplicity. The sound source is moving in the x-y plane with the velocity v_{xy}. This velocity is decomposed into the radial velocity v_r relative to one ear (e.g., the right ear).

2.2 DopplerBAS

We do not directly apply Eq. (1) in the frequency domain of audio, because some previous works (Lee and Lee, 2022) show that modeling the binaural audio in the frequency domain degrades the accuracy although it could benefit the generalization ability. Different from modeling the Doppler effect in the frequency domain, we calculate the velocity of interest and use it as a condition to guide the neural network to synthesize binaural audio consistent with the moving event. In the receiver-centric Cartesian coordinates, we define $\vec{p}_s$ and $\vec{p}_e$ as the 3D position of the moving sound source s and one ear of the receiver e respectively (e.g., the right ear, as shown in Figure 1). The

Model	Wave L2 ($\times 10^{-3}$) $\downarrow$	Amplitude L2 $\downarrow$	Phase L2 $\downarrow$	PESQ $\uparrow$	MRSTFT $\downarrow$
DSP (Leng et al., 2022)	1.543	0.097	1.596	1.610	2.750
WaveNet (Leng et al., 2022)	0.179	0.037	0.968	2.305	1.915
NFS (Lee and Lee, 2022)	0.172	0.035	0.999	1.656	1.241
*WarpNet** (Richard et al., 2021)	0.164	0.040	0.805	1.935	2.051
*WarpNet** + *DopplerBAS*	**0.154**	**0.036**	**0.780**	**2.161**	**2.039**
*BinauralGrad** (Leng et al., 2022)	0.133	0.031	0.889	2.659	1.207
*BinauralGrad** + *DopplerBAS*	**0.131**	**0.030**	**0.869**	**2.699**	**1.202**

Table 1: The comparison regarding binaural audio synthesis quality. For *WarpNet** and *BinauralGrad**, we reproduced the results using their official codes (Section 3.1).

position vector $\vec{p} = (p_x, p_y, p_z)$ of s relative to e is:

$$\vec{p} = (p_x, p_y, p_z) = \vec{p}_s - \vec{p}_e.$$

Then s's velocity [2] can be calculated as:

$$\vec{v} = (v_x, v_y, v_z) = (\frac{\mathrm{d}p_x}{\mathrm{d}t}, \frac{\mathrm{d}p_y}{\mathrm{d}t}, \frac{\mathrm{d}p_z}{\mathrm{d}t}).$$

Next, we build the spherical coordinate system using the ear as the origin, and decompose $\vec{v}$ into the radial relative velocity $\vec{v}_r$ by:

$$\vec{v}_r = \frac{\vec{p} \cdot \vec{v}}{\|\vec{p}\|} \cdot \hat{\mathbf{r}}, \tag{2}$$

where $\hat{\mathbf{r}} \in \mathcal{R}^1$ is the radial unit vector.

Finally, we add $\vec{v}_r$ as the additional condition to the network: The original conditions in monaural-to-binaural speech synthesis are $C_o \in \mathcal{R}^7 = (x, y, z, qx, qy, qz, qw)$, of which the first 3 represent the positions and the last 4 represent the head orientations. We define the new condition $C \in \mathcal{R}^9 = (x, y, z, qx, qy, qz, qw, v_{r-left}, v_{r-right})$, where v_{r-left} and $v_{r-right}$ represent the radial velocity of source relative to the left and right ear respectively, which are derived from Eq. (2). We then apply C to WarpNet and BinauralGrad backbones, as follows.

2.2.1 WarpNet

WarpNet consists of two blocks: 1) The Neural Time Warping block to learn a warp from the source position to the listener's left ear and right ear while respecting physical properties (Richard et al., 2021). This block is composed of a geometric warp and a parameterized neural warp. 2) The Temporal ConvNet block to model subtle effects such as room reverberations and output the final binaural

audio. This block is composed of a stack of hyper-convolution layers. We replace the original C_o with C for the input of parameterized neural warp and for the condition of hyper-convolution layers.

2.2.2 BinauralGrad

BinauralGrad consists of two stages: 1) The "Common Stage" generates the average of the binaural audio. The conditions for this stage include the monaural audio, the average of the binaural audio produced by the geometric warp in WarpNet (Richard et al., 2021), and C_o. 2) The "Specific Stage" generates the final binaural audio. The conditions for this stage include the binaural audio produced by the geometric warp, the output of the "Common Stage", and C_o. BinauralGrad adopts diffusion model for both stages, which is based on non-causal WaveNet blocks (Oord et al., 2016) with a conditioner block composed of a series of 1D-convolutional layers. We replace C_o with C as the input of the conditioner block for both stages.

3 Experiments

In this section, we first introduce the commonly used binaural dataset, and then introduce the training details for WarpNet-based and BinauralGrad-based models. After that, we describe the evaluation metrics that we use to evaluate baselines and our methods. Finally, we provide the main results with analytical experiments on BAS.

3.1 Setup

Dataset We evaluate our methods on the standard binaural dataset released by Richard et al. (2021). It contains 2 hours of paired monaural and binaural audio at 48kHz from eight different speakers. Speakers were asked to walk around a listener equipped with binaural microphones. An OptiTrack system track the positions and orientations of the speaker and listener at 120Hz, which

---

[2]This velocity is the same in all the Cartesian coordinate systems relatively stationary to the receiver.

are aligned with the audio. We follow the original train-validation-test splits as Richard et al. (2021) and Leng et al. (2022) for a fair comparison.

Training Details We apply DopplerBAS on two open-source BAS systems WarpNet and BinauralGrad. We train 1) WarpNet and WarpNet+DopplerBAS on 2 NVIDIA V100 GPUs with batch size 32 for 300K steps, and 2) BinauralGrad and BinauralGrad+DopplerBAS on 8 NVIDIA A100 GPUs with batch size 48 for 300K steps [3].

Evaluation Metrics Following the previous works (Leng et al., 2022; Lee and Lee, 2022), we adopt 5 metrics to evaluate baselines and our methods: 1) **Wave L2**: the mean squared error between waveforms; 2) **Amplitude L2**: the mean squared errors between the synthesized speech and the ground truth in amplitude; 3) **Phase L2**: the mean squared errors between the synthesized speech and the ground truth in phase; 4) **PESQ**: the perceptual evaluation of speech quality; 5) **MRSTFT**: the multi-resolution spectral loss.

3.2 Main Results and Analysis

Main Results We compare the following systems: 1) *DSP*, which utilizes the room impulse response (Lin and Lee, 2006) to model the room reverberance and the head-related transfer functions (Cheng and Wakefield, 2001) to model the acoustic influence of the human head; 2) *WaveNet* (Richard et al., 2021; Leng et al., 2022), which utilizes the WaveNet (Oord et al., 2016) model to generate binaural speech; 3) *NFS*, which proposes to model the binaural audio in the Fourier space; 4) *WarpNet* (Richard et al., 2021), which proposes a combination of geometry warp and neural warp to produce coarse binaural audio from the monaural audio and a stack of hyper-convolution layers to refine coarse binaural audio; 5) *WarpNet + DopplerBAS*, which applies *DopplerBAS* to *WarpNet*; 6) *BinauralGrad* (Leng et al., 2022), which proposes to use diffusion model to improve the audio naturalness; 7) *BinauralGrad + DopplerBAS*, which applies *DopplerBAS* to *BinauralGrad*.

The results are shown in Table 1. "+ *DopplerBAS*" could improve both *WarpNet* and *BinauralGrad* in all the metrics, especially in the Phase L2 metric. *WarpNet + DopplerBAS* performs best in the Phase L2 metric and reaches a new state of the

No.	Model	W. L2	Amp. L2	Phase L2
1	*WarpNet*	0.164	0.040	0.805
2	+*Spherical $\vec{v}$*[†]	**0.154**	**0.036**	**0.780**
3	+*Cartesian $\vec{v}$*	0.164	0.038	<u>0.790</u>
4	+*Zeros*	0.159	0.038	0.806
5	+*Time series*	0.163	0.039	0.822

Table 2: Analysis Experiments. "W. L2" means Wave L2 $\cdot 10^3$; "Amp. L2" means Amplitude L2; [†] means our method: *DopplerBAS*. Best scores over the corresponding baseline are marked in bold.

art **0.780**. *BinauralGrad + DopplerBAS* obtains the best Wave L2, Amplitude L2, PESQ and MRSTFT score among all the systems. These results show the effectiveness of *DopplerBAS*.

Analysis We conduct analytical experiments for the following four velocity conditions. "*Spherical $\vec{v}$*": the velocity conditions introduced in Section 2.2 are calculated in the spherical coordinate system; "*Cartesian $\vec{v}$*": the velocity conditions are calculated in the Cartesian coordinate system; "*Zeros*": the provided conditions are two sequences of zeros; "*Time series*": the provided conditions are two sequences of time. The results are shown in Table 2, where we place WarpNet in the first row as the reference. We discover that: 1) Radial relative velocity is the practical velocity component, which obeys the theory of the Doppler effect (row 2 vs. row 1); 2) The velocity condition is beneficial to binaural audio synthesis, even for the absolute velocity in the Cartesian coordinates (row 3 vs. row 1); 3) Just increasing the channel number of the condition C_o (Section 2.2) by increasing the parameters in neural networks without providing meaningful information could not change the results (row 4 vs. row 1); 4) The neural networks do not explicitly learn the derivative of position to time (row 5 vs. row 1). These points verify the rationality of our proposed method.

4 Conclusion

In this work, we proposed DopplerBAS to address the Doppler effect of the moving sound source in binaural audio synthesis, which is not explicitly considered in previous neural BAS methods. We calculate the radial relative velocity of the moving source in the spherical coordinate system as the additional conditions for BAS. Experimental results show that DopplerBAS scales well to different types of backbones and reaches a new SOTA.

---

[3]Following the recommended training steps in their official repository.

Limitations

The major limitation is that we test our method only on a binaural speech dataset, in which there is a person moving slowly while speaking. Because this person moves slowly, the Doppler effect is not so obvious. We will try to find or collect a sound dataset of a source moving at high speed, such as a running man, flying objects, or vehicles, and further, analyze the experimental phenomena at different speeds of the moving source.

Ethics Statement

The immersive experience brought by space audio may make people indulge in the virtual world.

Acknowledgements

References

C.P. Brown and Richard O. Duda. 1998. A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2021. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Victor C. Chen, F. Li, Shen-Shyang Ho, and Harry Wechsler. 2006. Micro-doppler effect in radar: phenomenon, model, and simulation study. *IEEE Transactions on Aerospace and Electronic Systems*, 42:2–21.

Corey I. Cheng and Gregory H. Wakefield. 2001. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *Journal of The Audio Engineering Society*, 49:231–249.

Ruohan Gao and Kristen Grauman. 2019. 2.5d visual sound. In *CVPR*.

Thomas P. Gill. 1965. The doppler effect : an introduction to the theory of the effect. In *Logos Press, Limited*.

N. Giordano. 2009. *College Physics: Reasoning and Relationships*. Cengage Learning.

Claudia M. Hendrix and Woodrow Barfield. 1996. The sense of presence within auditory virtual environments. *Presence: Teleoperators & Virtual Environments*, 5:290–301.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *neural information processing systems*.

Wen-Chin Huang, Dejan Markovic, Alexander Richard, Israel Dejene Gebru, and Anjali Menon. 2022. End-to-end binaural speech synthesis. In *INTERSPEECH*.

jaan johansson, aki mäkivirta, matti malinen, and ville saari. 2022. interaural time difference prediction using anthropometric interaural distance. *journal of the audio engineering society*, 70(10):843–857.

Jingeun Lee, SungHo Lee, and Kyogu Lee. 2022. Global hrtf interpolation via learned affine transformation of hyper-conditioned features. *ArXiv*, abs/2204.02637.

Jinkyu Lee and Kyogu Lee. 2022. Neural fourier shift for binaural speech rendering. *ArXiv*, abs/2211.00878.

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, sheng zhao, and Tie-Yan Liu. 2022. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In *Advances in Neural Information Processing Systems*.

Yuanqing Lin and Daniel D. Lee. 2006. Bayesian regularization and nonnegative deconvolution for room impulse response estimation. *IEEE Transactions on Signal Processing*, 54:839–847.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125.

Kranti K. Parida, Siddharth Srivastava, and Gaurav Sharma. 2022. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2151–2160.

Sang-Min Park and Young-Gab Kim. 2022. A metaverse: Taxonomy, components, applications, and open challenges. *IEEE Access*, 10:4209–4251.

Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. 2022. Deep impulse responses: Estimating and parameterizing filters with deep networks. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3209–3213.

Alexander Richard, Dejan Markovic, Israel Dejene Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando De la Torre, and Yaser Sheikh. 2021. Neural synthesis of binaural speech from mono audio. In *ICLR*.

Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and R. Väänänen. 1999. Creating interactive virtual acoustic environments. *Journal of The Audio Engineering Society*, 47:675–705.

Kaushik Sunder, Jianjun He, Ee-Leng Tan, and Woonseng Gan. 2015. Natural sound rendering for headphones: Integration of signal processing techniques. *IEEE Signal Processing Magazine*, 32:100–113.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA.

Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. 2021. Visually informed binaural audio generation without binaural audios. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15480–15489.

Tao Zheng, Sunny Verma, and W. Liu. 2022. Interpretable binaural ratio for visually guided binaural audio generation. *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yin Zhu, Qiuqiang Kong, Junjie Shi, Shilei Liu, Xuzhou Ye, Ju-Chiang Wang, and Junping Zhang. 2022. Binaural rendering of ambisonic signals by neural networks. *ArXiv*, abs/2211.02301.

Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis. 2004. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6:553–564.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the "Limitations" section.

☑ A2. Did you discuss any potential risks of your work?
In the "Ethics Statement" section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

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

B ☑ Did you use or create scientific artifacts?

3.1

☑ B1. Did you cite the creators of artifacts you used?
3.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The original author did not provide this information.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ 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.
3.1

C ☑ Did you run computational experiments?

3.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.1

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?
3.1

☑ 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?
3

☑ 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.)?
3

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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.