Automatic Named Entity Obfuscation in Speech

Judita Preiss
University of Sheffield, Information School
The Wave, 2 Whitham Road
Sheffield S10 2AH
judita.preiss@sheffield.ac.uk

Abstract

Sharing data containing personal information often requires its anonymization, even when consent for sharing was obtained from the data originator. While approaches exist for automated anonymization of text, the area is not as thoroughly explored in speech. This work focuses on identifying, replacing and inserting replacement named entities synthesized using voice cloning into original audio thereby retaining prosodic information while reducing the likelihood of deanonymization. The approach employs a novel named entity recognition (NER) system built directly on speech by training HuBERT (Hsu et al., 2021) using the English speech NER dataset (Yadav et al., 2020). Name substitutes are found using a masked language model and are synthesized using text to speech voice cloning (Eren and Team, 2021), upon which the substitute named entities are re-inserted into the original text. The approach is prototyped on a sample of the LibriSpeech corpus (Panayotov et al., 2015) with each step evaluated individually.

1 Introduction

Privacy concerns, particularly where an individual could be identified, preclude sharing and therefore automatic exploitation of many data sources. Anonymization, the removal of identifying information, has been automated for text (Lison et al., 2021), including large scale applications such as in clinical (Hartman et al., 2020) or legal settings (Oksanen et al., 2022), with off-the-shelf systems having reported performance of 90+% (Hartman et al., 2020). To minimize the risk of re-identification, obfuscation – replacing identifying information with a different substitute of the same type – has been explored as an alternative to replacing identifying information with a generic marker (Sousa and Kern, 2022). The main focus in speech has been on voice anonymization, which may not be a problem with speaker consent, with the removal of identifying information receiving less attention. To our knowledge, this is the first prototype to perform named entity obfuscation directly, in the original speaker's voice. Aside from voice cloning, it explores a named entity recognition approach based directly on audio signal and uses language model masking to find appropriate substitutions.

Recent advances in speech models, particularly the inclusion of language models within the speech model itself (e.g. HuBERT (Hsu et al., 2021)) gives models greater insight into expected contexts. Previous work on named entity recognition (NER) in speech frequently employs a two step approach, transcribing speech first, followed by the application of existing named entity techniques (Yadav et al., 2020). However, this process has the potential to compound errors as errors in transcription will increase the probability of error in NER. We suggest that the addition of language models into the speech model gives these sufficient power to perform NER directly, and therefore that transcribing (automatic speech recognition, ASR) and NER can be separated, and used to provide a confidence measure in their performance. Divided, the two do not propagate errors in the same way; in fact, treating ASR and NER separately allows one to fix (some of the) errors of the other. The proposed second (final) ASR pass merely produces a confidence value in the result to decide whether a manual check should be performed.

The success of few shot learning, where a limited number of examples is used to generalize a pre-trained deep learning model to a new situation, for text-to-speech – and specifically voice cloning (Zhang and Lin, 2022) – enables an alternative, equivalent but different, entity to be inserted in the audio signal in place of the original while preserving the prosody information throughout. While large databases of potential replacement entities can be used to select a substitution, these may not preserve necessary properties (such as gender). Al-

Findings of the Association for Computational Linguistics: ACL 2023, pages 615–622
July 9-14, 2023 ©2023 Association for Computational Linguistics

ternatively, word embeddings have been used to suggest close (in the multi-dimensional space) alternatives (Abdalla et al., 2020), however these can suffer from the same drawback. We propose using a more contextualized alternative to word embeddings, a masked language model (Devlin et al., 2019), where the model is trained by hiding (masking) words and predictions of the original word are made based on their context.

This work makes the following contributions: (1) a complete obfuscation pipeline for names in speech[1], (2) a named entity recognizer built directly on speech without requiring text transcription first, (3) alternative (obfuscated) entity replacement selection via masking language model, and (4) confidence annotated system output, allowing for manual correction and / or selection of shareable instances. Section 2 contains the methodology with results in Section 3. Section 4 presents the conclusions and future work.

2 Methodology

The steps of the overall pipeline, which takes in an audio file and produces an obfuscated audio file along with a confidence value, can be found in Figure 1. The approach comprises of three main parts: 1) identification of named entities (NEs) in the audio, 2) finding an equivalent alternative for the original NEs, and 3) reconstructing the original audio to incorporate the replacement NEs. The reconstructed audio can further be used to obtain a confidence value.

2.1 Identification of named entities

To enable the direct use of a language model on speech input for the purpose of named entity recognition (NER), a dataset of audio recordings with annotated NEs is required. The English speech NER dataset (Yadav et al., 2020), which consists of 70,769 waveforms with transcripts annotated with person, location and organization NEs, is used for fine-tuning the Hidden-Unit BERT speech model (HuBERT) (Hsu et al., 2021). HuBERT was selected over other speech models since it learns both accoustic and language models from its inputs and therefore has an increased awareness of context. The success of language models on text NER has demonstrated how crucial context is for this task, and using a model which incorporates both an acoustic and a language model (over acoustic only) allows the approach to exploit the information used in text NER, while managing to avoid the need for a transcript.

For training, NE annotations need to be converted to a suitable format, indicating the presence or absence of a NE in each position. Following the inside-outside(-beginning) chunking common to many NER approaches (Tjong Kim Sang and De Meulder, 2003), three formats were explored: 1) character level annotation, mapping each character to either *o* for a character outside of a named entity, space, or *n, l, e* for characters within person, location or organization entities respectively, 2) the same character level annotation with separate characters added to denote the beginning of each type of NE (mapping the sentence *TELL JACK* to *oooo mnnn* with *m* denoting the start of a person NE), 3) and, for completeness, annotation was also explored at word level.

With the training parameters shown in Appendix A.1, the best NE performance was obtained from the first annotation approach, where NE beginnings were not explicitly annotated. The lower performance of the second annotation approach can be attributed to the low quantity of training data for the beginning marker annotations. While word level annotation was explored, it is likely to need a far greater quantity of data to enable mapping of different length inputs to a single label.

Separately, HuBERT was also fine-tuned for automatic speech recognition (ASR), i.e. for transcribing text from audio. Identical training data was used, with annotation being the transcription provided as part of the NE annotation (with NE annotation removed). The same parameters were employed for its training. Alongside the predicted (NE or ASR) annotation, prediction output also yields an offset which can be converted to a time offset. This can be used to identify the position of the NE(s) to be replaced, and after a greedy alignment of the two outputs, the original transcription of the original NE(s) can be extracted.

2.2 Finding an alternative NE

Once a person NE is identified, a suitable equivalent substitution needs to be obtained, i.e. we want to find the word which could replace the NE in the text if the NE was hidden. This is precisely the concept behind masked language models (MLMs):

[1] The code for the obfuscation pipeline is available from: https://github.com/juditapreiss/speech_ne_obfuscation

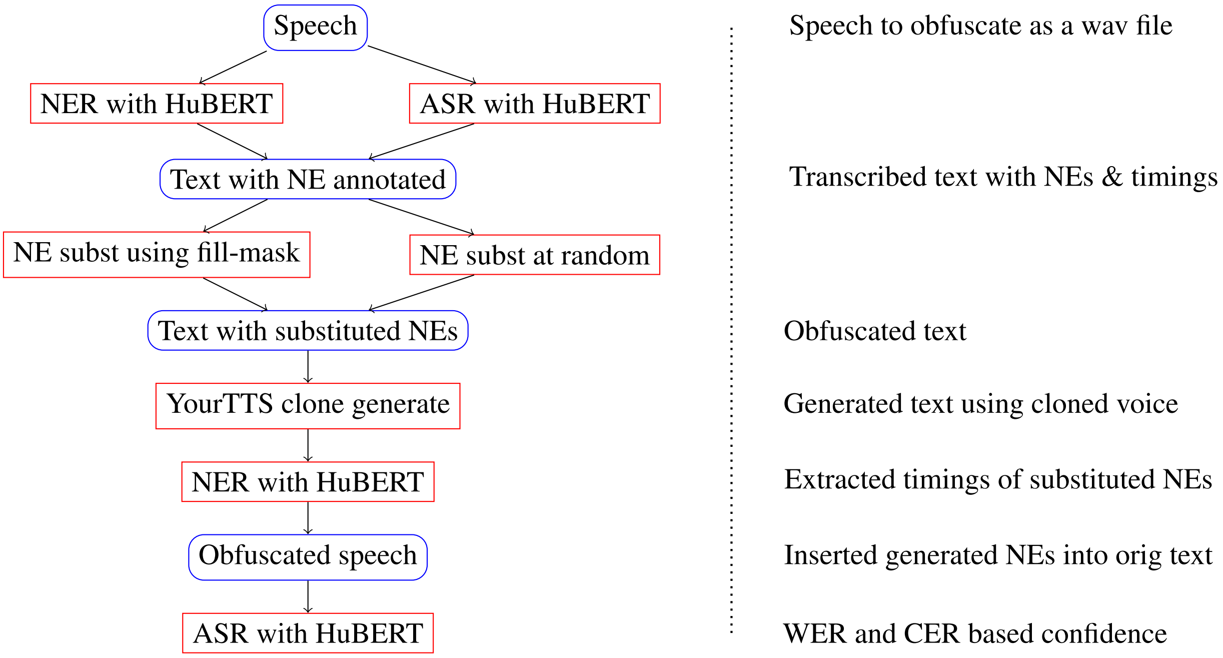
Speech
NER with HuBERT
ASR with HuBERT
Text with NE annotated
NE subst using fill-mask
NE subst at random
Text with substituted NEs
YourTTS clone generate
NER with HuBERT
Obfuscated speech
ASR with HuBERT
Speech to obfuscate as a wav file
Transcribed text with NEs & timings
Obfuscated text
Generated text using cloned voice
Extracted timings of substituted NEs
Inserted generated NEs into orig text
WER and CER based confidence

Figure 1: Obfuscation pipeline

these models learn their weights so that given a sentence with a hidden (masked) word, the model will output the complete original sentence. The (ASR extracted) original sentences with NEs (as identified by the NE tuned model) masked were passed to a MLM. Three MLM models were explored: BERT, `bert-large-uncased` model (Devlin et al., 2019), ALBERT, `albert-xxlarge-v2`, model (Lan et al., 2019) and the distilled RoBERTa base, `distilroberta-base`, model (Sanh et al., 2019). Each model, with no additional tuning, results in a (pre-specified) number of predictions for each NE in the sentence. Since the models used different datasets in training, their predictions are expected to be different: for example, some may suggest pronouns rather than names.

Given the propensity of the MLM to return substitutions which are not names (for example, for the sentence *you should call Stella*, the model returns *you should call him*, *you should call them*, *you should call 911* etc), an external list of people names is used for the validation of the proposed suggestions[2] and the highest scoring substitution is returned. Heuristically, the original name is matched against the list to identify whether it is a first or a last name (where possible) and names of the same type suggested by the MLM are returned. Simple rules are employed (last of a sequence of names is a last name, a single name without a title is a first name etc) to decide on a substitution when the original name does not appear in either the first or last name list. Given the nature of MLMs, suggested alternatives are likely to be more common words: as a positive side effect, this should make them easier to render with voice cloning as they may already appear in the reference speech. Should MLM fail to propose any suitable substitutions, one is selected at random from the first & last name lists, subject to the same heuristic rules.

2.3 Reconstruction of original audio

In this work, the substitute NE is to be re-inserted into the original audio. To reduce the risk of de-identification via the extraction of entities which failed to be identified and therefore stayed in their original form, the substitute entity needs to be produced in the speaker's voice. The YourTTS (Casanova et al., 2021) model, which offers the ability for fine-tuning with less than one minute of speech while achieving good results with reasonable quality, can be used to generate the substitute sentence with all available speech of the speaker provided as reference. Note that it is not necessary to remove the original sentence from the reference data: in fact, its presence may result in more accurate rendering of the substitute sentence. The pre-trained model used in this work (`tts_models/multilingual/multi-dataset/your_tts`)

[2] In this work, `https://github.com/dominictarr/random-name` are used.

was trained on the the voice cloning toolkit (VCTK) dataset (Yamagishi et al., 2019) which contains approximately 400 sentence, selected from newspaper text, uttered by 108-110 different speakers, giving it its generalization power. Aside from the reference passed to the model on the command line, no tuning or training of the YourTTS model is done in this work.

The ASR transcribed text with the substituted NE is generated, rather than the substitution alone, to ensure that the intonation as closely matches the substitution's position in the sentence. The average amplitude of the generated audio is matched to that of the original segment using the Python pydub library. The generated audio is again pased through the HuBERT based NE recognizer, to identify the location of the substituted NE in the generated audio and allow its extraction (note that in this pass, it is not necessary to perform ASR – only the offsets of the replacement NE are required). Should the NE recognizer not identify the same number of NEs as were present in the original, the instance is flagged for manual review.

For each NE in the text, a pair of start and end offsets are available: one pair extracted by the HuBERT based NE extraction from the original audio and a second pair from the audio generated from the substituted text. This allows the new NEs to be inserted in place of the original NEs. The splicing and concatenation of the waveforms is also performed using the pydub library.

A second HuBERT based ASR pass over the newly constructed (substituted) audio, and its comparison against the substituted text using word error rate (WER) and character error rate (CER) gives measures of confidence. Both the metrics, commonly used for evaluation of ASR, allow for sequences of different length to the target – the further the reconstructed audio is from the target sentence, the less likely it is that the substitution will go unnoticed.

3 Results and discussion

3.1 Identification of named entities

The 70,769 training corpus, sampled at 16kHz, is divided up into 70% for training (49,540 instances), and 15% for both validation and evaluation (10,615 examples). The `hubert-base-ls960` model is used with parameters listed in Appendix A.1. The performance in training, indicated via WER and CER, is shown in Table 1 for both ASR and NER.

	Eval WER	Eval CER
ASR	0.142	–
NE	0.199	0.063

Table 1: Metric results of the ASR and NE HuBERT based models

MLM	Avg ASR	NE	Avg confidence
ALBERT	0.980	13/20	0.109
BERT	0.980	13/20	0.098
RoBERTa	0.980	13/20	0.106

Table 2: Evaluation of individual steps

For the purpose of the demonstrating the viability of the prototype, no hyperparameter optimization was performed, and the larger HuBERT models were not employed, however improvement in performance of both models are expected should this be pursued.

3.2 Finding an alternative NE

A small scale evaluation is performed on a sample of 20 sentences selected at random from the LibriSpeech corpus (Panayotov et al., 2015) across 6 speakers. Sentence selection was subject to them containing a person named entity. While detailed results for the individual steps can be found in Table 2, it should be noted that – for the purposes of this work – the focus is the accuracy of the extraction of the correct NE. The stated accuracy is therefore somewhat misleading: in a number of cases, such as the word *Raphael*, the named entity is divided into two separate words, suggesting two consecutive named entities. However, this issue is corrected when the NE output is aligned with ASR output and the two separate NE instances are (correctly) merged. Cases with NEs which cannot be aligned are flagged up for manual intervention. The average ASR and (exact match) NE identification do not vary when a different MLM is employed, as this only effects the selection of the substituted name, resulting in different average confidence values.

3.3 Reconstruction of original audio

The voice cloning model requires some reference audio for the speaker: for the 6 selected speakers, 4 have less than 5 audio files (two having 3, and one having only 2 files) in the dataset. The quantity of data used as reference is likely to impact the quality (in terms of its similarity to the original speaker) of the generated text. Given the likely scenarios of deployment, such as dialogues where more than 2 sentences of speech per speaker are available, this may not be representative of the results obtainable with the pipeline. However, it should be noted that even if all substituted instances can be identified as substitutions, the system is equal to a masking technique (where an entity is replaced with a fixed entity, such as a bleep).

4 Conclusion

The prototype described shows the steps of an obfuscation pipeline for speech, which results in substituted person named entities uttered in the original speakers voice and replaced in the original audio signal. The prototype makes use of a named entity recognizer built directly on top of audio input, and employs masked language models to generate the substituted entity. It offers an end-to-end automatic solution enabling the sharing of speech with identifying information removed.

The resulting obfuscated speech remains in the original speaker's voice, allowing for the application of traditional speaker anonymization approaches to mask the speaker's identity. The original prosody can be protected by applying a transformation such as waveform change, offering a significant advantage over a technique which generates a complete obfuscated transcription (instead of splicing an obfuscated entity into original speech).

Limitations

The cloning model used, YourTTS, is trained on the VCTK dataset which consists of high-quality speech signal. It is therefore unclear whether the same accuracy would be obtained with lower quality signal which may contain some background noise. (However, it should again be noted that even if all substituted instances are identifiable in the output, the system is equivalent to a masking model.)

The selection of a person NE replacement does not currently account for continuity: if the same person entity is referred to later, it may be substituted with a different entity to the previous occasion. In addition, the back-off strategy ignores aspects such as gender.

To show the approach feasible, very little optimization was performed. Further training and parameter optimization is likely to lead to improved performance for both ASR and NER models.

The approach is currently only implemented for person NEs but it could be extended very simply to other types of NEs. However, the degree to which other entity types require obfuscation in speech is not clear to us as mentions of organizations may well not be identifying at all.

Ethics Statement

Aside from the ethical concerns regarding voice cloning (covered in e.g. YourTTS (Casanova et al., 2021)), deployment would require a detailed evaluation of risk of de-identification. It is believed that the final confidence and the accuracy of each step can be combined to significantly reduce this risk. The voice itself also offers options for identification: the value of yielding substitutions in the original speaker's voice (and keeping the original prosody) would need to be weighed up against approaches which anonymize voice but preserve prosodic information.

References

Mohamed Abdalla, Moustafa Abdalla, Frank Rudzicz, and Graeme Hirsto. 2020. Using word embeddings to improve the privacy of clinical notes. *J Am Med Inform Assoc*, 27(6):901–907.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir Antonelli Ponti. 2021. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. *CoRR*, abs/2112.02418.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Gölge Eren and The Coqui TTS Team. 2021. Coqui TTS. `https://github.com/coqui-ai/TTS`.

Tzvika Hartman, Michael D. Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, Ming Jack Po, Jutta Williams, Scott Ellis, Gavin Bee, Avinatan Hassidim, Rony Amira, Genady Beryozkin, Idan Szpektor, and Yossi Matias. 2020. Customization scenarios for de-identification of clinical notes. *BMC Med Inform Decis Mak*, 20(14).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Arttu Oksanen, Eero Hyvönen, Minna Tamper, Jouni Tuominen, Henna Ylimaa, Katja Löytynoja, Matti Kokkonen, and Aki Hietanen. 2022. An anonymization tool for open data publication of legal documents. In *Joint Proceedings of the 3th International Workshop on Artificial Intelligence Technologies for Legal Documents (AI4LEGAL 2022) and the 1st International Workshop on Knowledge Graph Summarization (KGSum 2022) co-located with the 21st International Semantic Web Conference (ISWC 2022), Virtual Event, Hangzhou, China, October 23-24, 2022*, volume 3257 of *CEUR Workshop Proceedings*, pages 12–21. CEUR-WS.org.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Samuel Sousa and Roman Kern. 2022. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artif Intell Rev*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from English speech. *CoRR*, abs/2005.11184.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Haitong Zhang and Yue Lin. 2022. Improve few-shot voice cloning using multi-modal learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8317–8321.

A Model training details

A.1 HuBERT parameters

HubertForCTC hubert-base-ls960 model was used with frozen feature encoder, training size of 49540 examples, and validation and evaluation size of 10615, alongside the following parameters:

Parameter	Value
group_by_length	True
per_device_train_batch_size	8
per_device_eval_batch_size	batch_size
evaluation_strategy	"steps"
num_train_epochs	num_epochs
fp16	True
gradient_checkpointing	True
save_steps	500
eval_steps	500
learning_rate	1e-4
weight_decay	0.005
warmup_steps	1000
num_epochs	30

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Unnumbered, follows conclusion

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement discusses potential risks of the work. This follows "Limitations".

☑ 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?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ 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)?
No response.

□ 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?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ 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.
No response.

C ☑ Did you run computational experiments?

Sections 2 & 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?
Appendix A1. The infrastructure was chosen by the HPC and therefore wasn't known - if needed, the experiments can be rerun with this constrained.

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?
Experimental setup discussed, hyperparameter optimization not performed as stated as the paper only checks viability.

☑ 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?
Average used, clearly denoted. Section 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.)?
Not applicable. Left blank.

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.