Responsible NLP Checklist

Paper title: *Discourse-Driven Code-Switching: Analyzing the Role of Content and Communicative Function in Spanish-English Bilingual Speech*
Authors: *Debasmita Bhattacharya, Juan Junco, Divya Tadimeti, Julia Hirschberg*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
Our work is mainly in the exploratory phase and thus has few associated risks at the current stage. Our work is not currently tied to any particular applications and we do not presently see any paths to negative applications.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
We cite the creator of the dataset we use throughout the paper, particularly in the Introduction (Section 1) and Method (Section 3) sections. We also cite other artifacts like tag sets and model class balancing techniques used throughout Sections 2 and 3.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We state the license information of the data set we use in a footnote in Section 3 about the corpus. We comply with its conditions of use by referring to it by name appropriately, and providing a link to the website from which we accessed it (embedded in Section 3).

☒ 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)?
The data set we use did not specify its intended use. Nonetheless, we use it only for research purposes. Similarly, we do not explicitly note the consistency of our use of other existing artifacts with creator intentions, since these were developed for research purposes such as ours. For the artifacts we have created, we have added intended use information in the README associated with our data release, linked in the paper.

☑ 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?

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

We state in the Ethics Statement that we did not access any information that uniquely identifies individuals in the data set we used. We mention that the original authors had already de-identified the corpus, as outlined in the original data set documentation.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We state the coverage of domains, languages, and demographic groups represented in our work in Section 3 about the corpus.

☑ 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?
We report relevant statistics about the data set and information about train/test splits in Section 3 about the corpus and methodology.

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We report the number of parameters in each model in Section 3 about the method. We do not use GPUs, but state in a footnote in Section 3 that all models were trained in under an hour on a Mac M1 chip machine.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We describe the experimental setup in detail in the Method (Section 3) and in the Appendix. This includes information about data preprocessing. We report that we used default hyperparameter settings for both our models in the Appendix. We did not perform hyperparameter tuning, manually or otherwise, or model selection beyond using the default hyperparameter settings.

☑ 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?
It is transparent in Table 4 and Section 4.3 that we are reporting performance metrics from a single run. We qualify these model results by reporting statistically significant comparisons to baseline performance via z-tests of proportions, as stated in Table 4 and Section 4.3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
We report package version numbers in the Method and Results sections (Sections 3 and 4).

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☒ 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.?
The two human annotators in this work were the second author and a volunteer. We include an abridged version of the annotation instructions followed in the Method (Section 3) and Appendix. There were no risks to annotators; we labeled secondary non-offensive data that did not collect personal identifying information.

☑ 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)?
We mention in the Limitations section that one of our annotators was a high school student who volunteered for the labeling task and thus was not paid for annotation. We do not report additional

information about the other annotator, who is the second author of the work, and was not paid for annotation.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
We mention in the Limitations section that our volunteer annotator was briefed on how their annotations would be used prior to starting the labeling task. We did not explicitly do the same with the other human annotator in this work, the second author, as they volunteered to perform annotations and were aware of how the annotations would be used because they planned and executed the data analysis themself.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics review was not applicable to our data annotation.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We mention the relevant language proficiency of both annotators in the Method (Section 3) and Limitations sections. The geographic characteristics of the two annotators can be inferred from our institution information.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)