How effective is machine translation on low-resource code-switching?  
A case study comparing human and automatic metrics

Li Nguyen\textsuperscript{1,2}, Christopher Bryant\textsuperscript{1}, Oliver Mayeux\textsuperscript{3}, Zheng Yuan\textsuperscript{1,4}  
\textsuperscript{1}ALTA Institute, University of Cambridge, UK  
\textsuperscript{2}Linguistics and Language Technology Lab, FPT University, Vietnam  
\textsuperscript{3}Trinity College, University of Cambridge, UK  
\textsuperscript{4}Department of Informatics, King’s College London, UK  
\{nhbn2, cjb255, ofm23\}@cam.ac.uk, zheng.yuan@kcl.ac.uk

Abstract
This paper presents an investigation into the differences between processing monolingual input and code-switching (CSW) input in the context of machine translation (MT). Specifically, we compare the performance of three MT systems (Google, mBART-50 and M2M-100\textsuperscript{big}) in terms of their ability to translate monolingual Vietnamese, a low-resource language, and Vietnamese-English CSW respectively. To our knowledge, this is the first study to systematically analyse what might happen when multilingual MT systems are exposed to CSW data using both automatic and human metrics. We find that state-of-the-art neural translation systems not only achieve higher scores on automatic metrics when processing CSW input (compared to monolingual input), but also produce translations that are consistently rated as more semantically faithful by humans. We further suggest that automatic evaluation alone is insufficient for evaluating the translation of CSW input. Our findings establish a new benchmark that offers insights into the relationship between MT and CSW.

1 Introduction

Code-switching (CSW) is the linguistic phenomenon where two or more languages are mixed within a discourse or utterance. This is illustrated in the following example which mixes English and Vietnamese.

(1) and mỗi group phải có a different focus each  
‘and each group must have a different focus’  
(from CanVEC, Nguyen and Bryant, 2020)

Code-switching occurs frequently and naturally among bilingual speakers and has recently become increasingly visible in social media data (Doğruöz et al., 2021; Winata et al., 2022). Despite its prevalence however, Natural Language Processing (NLP) applications are typically designed to process monolingual data and so often struggle with CSW input (Solorio et al., 2021; Sitaram et al., 2020; Doğruöz et al., 2021; Nguyen et al., 2021, 2022). For machine translation (MT), no current system is designed to support code-switched text (Çetinoğlu et al., 2016; Menacer et al., 2019); and despite increasing research attention in recent years (see e.g. Chen et al., 2022 for an overview), work in this area remains sparse.

In this paper, we explore the limits of three off-the-shelf state-of-the-art machine translation systems in terms of their ability to translate Vietnamese/English CSW data, using both automatic and human evaluation metrics. As far as we are aware, this is the first study to investigate the efficacy of machine translation on CSW data involving a low-resource language which is also structurally vastly different from English. In fact, existing work has mainly focused on comparatively better resourced and/or typologically similar languages, such as Spanish/English (Xu and Yvon, 2021), French/English (Xu and Yvon, 2021; Weller et al., 2022) or Hindi/English (Appicharla et al., 2021). Vietnamese/English, or Vietnamese in particular, remains severely under-represented in NLP.

We conduct our analysis using a variety of both automatic and human metrics in order to i) better understand the strengths and weaknesses of different systems, and ii) gain some insight into the relationship between automatic and human metrics with respect to CSW input. We find that systems not only achieve higher scores on CSW input (compared to monolingual input) according to automatic metrics, but also produce translations that are considered more semantically faithful by humans. Automatic metrics furthermore fail to correlate with human judgements, which suggests that automatic evaluation alone is not enough for evaluating MT output of CSW input. We release our annotations to facilitate future research.
2 Experimental Setup

2.1 Data

We conduct our experiments using the Canberra Vietnamese English natural speech corpus\(^1\) (CanVEC), which consists of 23 self-recorded conversations among 45 Vietnamese immigrants living in Canberra, Australia (Nguyen and Bryant, 2020). One advantage of CanVEC is that it contains transcribed CSW produced by bilingual speakers in an informal speech setting – an environment that has been found to be most conducive to natural CSW behaviour (Poplack, 1980, 1993; Labov, 2004; Torres Cacoullos and Travis, 2018; Nguyen, 2018, 2020). This differs to other NLP work in this domain, which has explored either scripted CSW speech corpora (Chan et al., 2005; Shen et al., 2011; Modipa et al., 2013; Yilmaz et al., 2017) or social media text (Doğruöz and Skantze, 2021; Winata et al., 2022).

The full CanVEC corpus consists of 14,047 clauses,\(^2\) of which 3,313 contain CSW.\(^3\) From these 3,313, we then selected a random sample of 100 clauses, which i) contained at least 5 tokens, and ii) represented the maximum number of unique speakers. The first condition was set to ensure clauses were of a minimum length to aid contextual translation, while the second condition was set to ensure the data was diverse and did not overly represent individual speakers. Various statistics about CanVEC and our test set are shown in Table 1.

Having selected 100 CSW clauses, we gave them to two bilingual annotators with complementary language competencies; i.e. L1 English/L2 Vietnamese and L1 Vietnamese/L2 English. Each annotator then translated the CSW clauses into monolingual English and monolingual Vietnamese respectively. The monolingual English translations were used as references, while the monolingual Vietnamese translations were used as source text, which allowed us to compare CSW translations against a more typical monolingual baseline.

### Table 1: Descriptive statistics of the Mixed part of CanVEC in relation to our random sample. Avg. EN Toks is the average proportion of English tokens in a mixed clause; i.e. most CSW clauses are majority Vietnamese with English mixed in.

<table>
<thead>
<tr>
<th></th>
<th>Clauses</th>
<th>Avg. Len.</th>
<th>Tokens</th>
<th>Avg. EN Toks (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CanVEC (Mixed)</td>
<td>3,313</td>
<td>7.5</td>
<td>24,807</td>
<td>30.90%</td>
</tr>
<tr>
<td>CanVEC (Sample)</td>
<td>100</td>
<td>8.6</td>
<td>862</td>
<td>29.13%</td>
</tr>
</tbody>
</table>

2.2 MT systems

We employ three widely used multilingual NMT models, which support both English and Vietnamese, and represent the cutting edge in both commercial and academic research.

**Google Translate\(^4\)** is one of the world’s most popular translation services that supports 133 languages. We access it using the translatepy\(^5\) v2.3 Python API.  

**mBART-50** is an extension of pre-trained multilingual BART (Liu et al., 2020) that has been fine-tuned on 50 languages (Tang et al., 2021). We use the mbart-50-many-to-many model.\(^7\)

**M2M-100** is another multilingual model that has been trained to translate between any pair of 100 different languages (Fan et al., 2021). It has been noted to perform better on non-English translations than other models and produce fluent translations with high semantic accuracy. We use the large 1.2B parameter model.\(^8\)

3 Evaluation

3.1 Automatic evaluation

Robust evaluation is still an unsolved problem in machine translation and many metrics have been proposed (Chatzikoumi, 2020). In our experiments, we compare five different automatic metrics, which evaluate translation output quality in different ways.

**BLEU** is the most widely used metric for automatic MT evaluation. It estimates similarity between system output and human reference translations in terms of precision of word \(n\)-gram overlap,\(^4\)

\(^1\)https://github.com/Bak3rLi/CanVEC  
\(^2\)The corpus was originally segmented into finite clauses to test a specific theoretical model of code-switching, the Matrix Language Framework (Myers-Scotton, 1997), but many of these clauses are equivalent to short sentences (Nguyen, 2020).  
\(^3\)The original CanVEC paper reports 2,721 mixed clauses because it redistributed non-clause utterances (e.g. noun phrases and interjections) to the language neutral tag.  
\(^4\)https://translate.google.com/  
\(^5\)https://pypi.org/project/translatepy/  
\(^6\)This free API may not be as good as the paid API.  
\(^7\)https://github.com/facebookresearch/fairseq/tree/main/examples/multilingual  
\(^8\)https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100
weighted by a brevity penalty to punish overly short translations (Papineni et al., 2002).

**chrF** computes an F-score using character n-grams (Popović, 2015). This helps reduce penalties when matching morphological variants of words. In our experiments, we used the default chrF2 which weights recall twice as much as precision.

**TER** evaluates a system in terms of the number of edit operations (i.e. insertions, deletions, shifts and substitutions) required to change a hypothesis sentence into a reference sentence (Snover et al., 2006).

**METEOR** is a token-based metric that additionally rewards semantic similarity in terms of exact string match, stem match and synonym match (Denkowski and Lavie, 2014).

**COMET** is a trained metric that is designed to output a score that correlates with the human perception of translation quality (Rei et al., 2020). It uses a cross-lingual encoder, XLM-R (Conneau et al., 2020), and pooling operations to obtain sentence-level representations of the source, hypothesis, and reference. These sentence embeddings are combined and then passed through a feed-forward network to produce a score.

We use the implementation in sacrebleu\(^9\) for the first three metrics (case agnostic, ignoring punctuation), and the pre-trained wmt20-comet-da model for COMET.\(^10\) METEOR is available separately.\(^11\)

### 3.2 Human evaluation

In addition to automatic metrics, we also manually rated system output according to three human metrics: Fluency, Grammaticality, and Semantic Faithfulness (Koehn, 2009; Dorr et al., 2011). These metrics are defined as follows.

- **Fluency**: does the translation sound natural/idiomatic in the target language?
- **Grammaticality**: is the translation grammatical, independent of the source?\(^12\)
- **Semantic Faithfulness**: does the translation retain the intended meaning of the source?

We trained two bilingual, domain-expert annotators to assign judgements for each metric on a binary scale (0: bad, 1: good). We used a binary scale because the input clauses in our experiments are short and there were unlikely to be a lot of translation errors that would warrant a more granular scale (Koehn, 2009, p.218). It is nevertheless worth mentioning that robust human evaluation of machine translation output is still an active area of research and alternate methodologies exist (van der Lee et al., 2019; Freitag et al., 2021; Licht et al., 2022; Saldías Fuentes et al., 2022).

### 4 Experiments

We evaluated our three chosen MT systems in two settings: code-switching to English (csw-en)\(^13\) and monolingual Vietnamese to English (vi-en). Recall that the sentences in the vi-en setting are the same as the csw-en setting except all English words and phrases were manually translated to Vietnamese by a human translator (Section 2.1). This enabled us to directly compare the effect of CSW against a highly controlled baseline.

Altogether, we obtained 200 translations from each system (100 clauses x 2 settings) and 600 translations in total (3 systems). We then asked our bilingual annotators to manually assign binary judgements to each translation based on the three human metrics (1800 judgements). Specifically, after training,\(^14\) we asked the L1 English annotator to assign judgements for Fluency and Grammaticality, and the L1 Vietnamese annotator to assign judgements for Semantic Faithfulness. We believe judgements for Fluency and Grammaticality require native assessment of the English translation irrespective of the source, while Semantic Faithfulness also requires native assessment of the Vietnamese source. In all cases, a positive judgement was only awarded if the translation fully met the criteria of the given metric; this conservative approach ensured greater confidence that positive judgements truly reflected a more competent translation.

### 5 Results and discussion

Results from all experiments are shown in Table 2. In terms of automatic metrics, we can see that

---

\(^9\)https://github.com/mjpost/sacrebleu

\(^10\)https://github.com/Unbabel/COMET

\(^11\)https://www.cs.cmu.edu/~alavie/METEOR/

\(^12\)For an example of how we distinguish Fluency and Grammaticality in particular, see Appendix A.

\(^13\)We also explored code-switching to Vietnamese (csw-vi) but found the output was very noisy so ultimately discounted these results. This was because the csw-vi setting is operationalised as an en-vi model even though the majority of input tokens are not English.

\(^14\)Both annotators doubly annotated 10% of the sample on all metrics. The average inter-annotator agreement rate across all metrics and settings was 91.7%. 
In this work, we compared the performance of three state-of-the-art MT systems on CSW input, using both automatic and human metrics. We found that systems not only achieved higher scores on automatic metrics when processing CSW input (com-

7 Table 2: Performance of all systems translating code-switching to English (csw-en) and monolingual Vietnamese to English (vi-en) in terms of automatic metrics and human metrics (Fluency, Grammaticality, Semantic Faithfulness) compared to a do-nothing code-switching baseline. The best scores are highlighted in bold.

<table>
<thead>
<tr>
<th>Setting</th>
<th>System</th>
<th>BLEU</th>
<th>chrF</th>
<th>TER</th>
<th>METEOR</th>
<th>COMET</th>
<th>Fluency</th>
<th>Gram.</th>
<th>Sem.</th>
</tr>
</thead>
<tbody>
<tr>
<td>csw-en</td>
<td>Google</td>
<td>27.159</td>
<td>52.861</td>
<td>51.171</td>
<td>0.315</td>
<td>0.098</td>
<td>92</td>
<td>98</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>mBART-50</td>
<td>25.935</td>
<td>49.916</td>
<td>56.792</td>
<td>0.292</td>
<td>0.182</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M2M-100big</td>
<td>24.362</td>
<td>48.101</td>
<td>54.333</td>
<td>0.287</td>
<td>0.271</td>
<td>82</td>
<td>97</td>
<td>69</td>
</tr>
<tr>
<td>vi-en</td>
<td>Google</td>
<td>15.639</td>
<td>41.346</td>
<td>64.169</td>
<td>0.257</td>
<td>-0.011</td>
<td>57</td>
<td>78</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>mBART-50</td>
<td>10.658</td>
<td>36.433</td>
<td>73.185</td>
<td>0.214</td>
<td>-0.158</td>
<td>92</td>
<td>100</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>M2M-100big</td>
<td>12.216</td>
<td>35.862</td>
<td>68.150</td>
<td>0.218</td>
<td>-0.115</td>
<td>75</td>
<td>95</td>
<td>55</td>
</tr>
</tbody>
</table>

8Table 3: System output for an example clause showing how CSW input may be more favourably constrained towards a reference compared to monolingual input.

<table>
<thead>
<tr>
<th>CSW-Input</th>
<th>Google</th>
<th>mBART-50</th>
<th>M2M-100big</th>
<th>VI-Input</th>
<th>Google</th>
<th>mBART-50</th>
<th>M2M-100big</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>hôi mà con made eye contact với Jimmy</td>
<td>I was made Eye Contact with Jimmy.</td>
<td>I made eye contact with Jimmy.</td>
<td></td>
<td>hôi mà con bất gặp ánh mắt của Jimmy</td>
<td>That’s when I saw Jimmy's eyes.</td>
<td>That’s when I met Jimmy’s eyes.</td>
<td></td>
<td>When I made eye contact with Jimmy</td>
</tr>
</tbody>
</table>

9 Google outperforms mBART-50 and M2M-100big on all metrics except for COMET in csw-en; this suggests that Google is the best of the three MT systems on our CSW/monolingual test sets. It is furthermore noteworthy that performance on csw-en translation for all three MT systems is consistently and significantly higher than monolingual vi-en translation. In fact, a do-nothing CSW baseline seems to outperform mBART-50 and M2M-100big at monolingual vi-en translation in terms of BLEU.

10 We hypothesise that this is because the translation might be considered ‘easier’ when CSW fragments only need to be copied to the output. For example, Table 3 shows that all systems generate output containing the phrase “made eye contact” when that same phrase is present in the CSW input, but generate synonymous output “caught Jimmy’s gaze”, “met Jimmy’s eyes” and “saw Jimmy’s eyes” from the monolingual Vietnamese input. BLEU thus benefits more from this exact word match compared to other automatic metrics. Consequently, CSW translation is more constrained than monolingual translation, which might make it ‘easier’ to achieve higher scores.

11 In contrast, system performance on human metrics is more varied, and different systems performed better and worse on different metrics. For example, mBART-50 achieved near-perfect scores for both Fluency and Grammaticality regardless of whether the input was CSW or monolingual, while Google achieved higher scores in the monolingual setting and M2M-100big achieved higher scores in the CSW setting on the same metrics. Holistically, this suggests that mBART-50 may be the most stable and effective of the three systems in terms of processing CSW input in relation to these metrics. Google, in contrast, appears to be the weakest system, which contradicts our findings from the automatic evaluation. This lack of agreement is not entirely surprising however, given that it is already challenging to develop automatic metrics that correlate with human judgements in monolingual settings (Fomicheva and Specia, 2019), let alone CSW settings where languages are mixed.

12 Among the three human metrics, we also observe that the scores for Semantic Faithfulness were consistently higher given CSW input compared to monolingual input. While this is again likely due to the constraining nature of CSW input, this result potentially suggests a specific aspect of MT where CSW input can contribute to enhancing system output. We direct readers to Appendix B for some detailed examples. Ultimately, we consider this finding worthy of further investigation, especially in relation to the development of models involving the understanding and/or generation of code-switching texts.

13 6 Conclusion

In this work, we compared the performance of three state-of-the-art MT systems on CSW input, using both automatic and human metrics. We found that systems not only achieved higher scores on automatic metrics when processing CSW input (com-
pared to monolingual input), but also produced translations that were consistently rated as more semantically faithful by humans. We furthermore observed that automatic and human metrics do not agree, which again highlights the need for more sophisticated, robust metrics, especially in non-monolingual tasks. Our findings establish a new benchmark in the relationship between MT and CSW, and motivate further research into how CSW might be used to improve future systems.

**Limitations**

The main limitation of our work is that 100 clauses is a small test set, but this was necessary to keep our human evaluation experiments manageable. We furthermore believe this was sufficient to be able to draw meaningful conclusions about the capabilities of different systems.

Another limitation is that we were only able to evaluate low-resource CSW in the context of Vietnamese and English. Future work might explore whether the same observations hold with CSW involving other low-resource languages, but this would require access to more suitable corpora and annotators.

**Ethics Statement**

We made every effort to make sure the work described in this paper adheres to the ACL Code of Ethics.

**Acknowledgements**

We would like to thank Professor Paula Buttery (Director, ALTA Institute, University of Cambridge) for her continuing support.

**References**


Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and


---

### GLOSSARY

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First person</td>
</tr>
<tr>
<td>2</td>
<td>Second person</td>
</tr>
<tr>
<td>CLF</td>
<td>Classifier</td>
</tr>
<tr>
<td>DET</td>
<td>Determiner</td>
</tr>
<tr>
<td>PL</td>
<td>Plural</td>
</tr>
<tr>
<td>POSS</td>
<td>Possessive</td>
</tr>
<tr>
<td>SG</td>
<td>Singular</td>
</tr>
<tr>
<td>Q</td>
<td>Question marker</td>
</tr>
</tbody>
</table>

### A Distinguishing Fluency and Grammaticality

We specified in Section 3.2 the three metrics that we used for human judgement in this work, namely *Fluency, Grammaticality* and *Semantic Faithfulness*. We consider the distinction between *Grammaticality* and *Fluency* an especially important aspect of languages in contact as it is likely to involve non-standard or hybrid features that may not be easily translated into the target language. Despite some overlap, there are cases in the dataset where these two criteria are clearly separated. Example (2) illustrates.

(2) *chính vì dùng mirror* [M2M]

*main because use*

nó mới có điểm chết đấy mấy 3SG then have point death PRT 2SG

Translation: ‘Because in the mirror, you have a point of death’

*Intended meaning:* ‘The blind spot is precisely because of using the mirror, you[2SG, VOCATIVE]’

Here, the machine translation is grammatically correct, but not fluent to a native’s ear. An expected fluent output in this case would be ‘You have a blind spot precisely because of the mirror.’ The use of the non-idiomatic expression ‘point of death’ and the topicalisation of the prepositional phrase ‘in the mirror’, therefore, while not wrong, could not be marked as fluent.

### B Analysis of Semantic Faithfulness

We reported near the end of the Discussion (section 5) that the scores for *Semantic Faithfulness* are always higher in the code-switching data (vi-en) compared to monolingual data (vi-en). This difference is confirmed as statistically significant (*p* < 0.05) using a bootstrap resampling test (Efron and Tibshirani, 1993). Here, we provide some qualitative examples.
As we can see, even when the code-switching part of the source only comprises a single word (highlighted in purple), the translation output is noticeably enriched. In (3) for example, while the Google system could not capture either the correct possessor (‘of those people’) or the precise meaning of the infinitive (‘become’), it was able to do so on both occasions in a csw-en setting. This is particularly striking considering that the source sentence is long (which should give sufficient context) and that the only difference between (3a) and (3b) is the language of one lexical item (‘bạn-gái’ vs ‘girlfriend’). Similarly, examples (4) and (5) show comparable behaviour for mBart-50 and M2M-100big, where a single code-switch noticeably adds to the output’s semantics.
ACL 2023 Responsible NLP Checklist

A For every submission:

✓ A1. Did you describe the limitations of your work?
   *Section 7: Mandatory limitations section*

☐ A2. Did you discuss any potential risks of your work?
   *Not applicable. No known risks*

✓ 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 2.1: New human translations Section 4: New human judgments*

✓ B1. Did you cite the creators of artifacts you used?
   *Section 2.1: Data Section 2.2: Machine translation models Section 3.1: Machine translation metrics*

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
   *Is it necessary to include a license for, e.g., the BLEU score in the paper?*

☐ 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?
   *Anonymization of the original corpus was described in the original corpus paper. We added new annotations to an already anonymized, ethically-created corpus.*

✓ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
   *Section 2.1: Corpus description*

✓ 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.
   *Table 1*

C ☒ Did you run computational experiments?

*We evaluated existing models on new data and carried out an automatic and human evaluation.*

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
   *No response.*

*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?

No response.

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?

No response.

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.)?

No response.

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

Section 2.1: Human translators Section 4: Human judgments of machine translation output

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.?

Section 3.2: Definitions of metrics on a binary scale.

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. Annotators are co-authors.

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. Annotators are co-authors.

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?

Section 2.1: L1/L2 information about annotators.