NollySenti: Leveraging Transfer Learning and Machine Translation for Nigerian Movie Sentiment Classification

Iyanuoluwa Shode[†] **David Ifeoluwa Adelani**[‡] **Jing Peng**[†] **Anna Feldman**[†]
[†]Montclair State University, USA, and [‡]University College London, United Kingdom
{shodei1,pengj,feldmana}@montclair.edu, d.adelani@ucl.ac.uk

Abstract

Africa has over 2000 indigenous languages but they are under-represented in NLP research due to lack of datasets. In recent years, there have been progress in developing labelled corpora for African languages. However, they are often available in a single domain and may not generalize to other domains. In this paper, we focus on the task of sentiment classification for cross-domain adaptation. We create a new dataset, NollySenti—based on the Nollywood movie reviews for five languages widely spoken in Nigeria (English, Hausa, Igbo, Nigerian-Pidgin, and Yorùbá). We provide an extensive empirical evaluation using classical machine learning methods and pre-trained language models. Leveraging transfer learning, we compare the performance of cross-domain adaptation from Twitter domain, and cross-lingual adaptation from English language. Our evaluation shows that transfer from English in the same target domain leads to more than 5% improvement in accuracy compared to transfer from Twitter in the same language. To further mitigate the domain difference, we leverage machine translation (MT) from English to other Nigerian languages, which leads to a further improvement of 7% over cross-lingual evaluation. While MT to low-resource languages are often of low quality, through human evaluation, we show that most of the translated sentences preserve the sentiment of the original English reviews.

1 Introduction

Nigeria is the sixth most populous country in the world[1] and the most populous in Africa with over 500 languages (Eberhard et al., 2021). These languages are spoken by millions of speakers, and the four most spoken indigenous languages (Hausa, Igbo, Nigerian-Pidgin (Naija), and Yorùbá) have more than 25 million speakers but they are still under-represented in NLP research (Adebara and Abdul-Mageed, 2022; van Esch et al., 2022). The development of NLP for Nigerian languages and other African languages is often limited by a lack of labelled datasets (Adelani et al., 2021b; Joshi et al., 2020). While there have been some progress in recent years (Eiselen, 2016; Adelani et al., 2022b; NLLB-Team et al., 2022; Muhammad et al., 2023; Adelani et al., 2023), most benchmark datasets for African languages are only available in a single domain, and may not transfer well to other target domains of interest (Adelani et al., 2021a).

One of the most popular NLP tasks is sentiment analysis. In many high-resource languages like English, sentiment analysis datasets are available across several domains like social media posts/tweets (Rosenthal et al., 2017), product reviews (Zhang et al., 2015; He and McAuley, 2016) and movie reviews (Pang and Lee, 2005; Maas et al., 2011). However, for Nigerian languages, the only available dataset is NaijaSenti (Muhammad et al., 2022) - a Twitter sentiment classification dataset for four most-spoken Nigerian languages. It is unclear how it transfers to other domains.

In this paper, we focus on the task of sentiment classification for cross-domain adaptation. We create the first sentiment classification dataset for Nollywood movie reviews known as **NollySenti** — a dataset for five widely spoken Nigerian languages (English, Hausa, Igbo, Nigerian-Pidgin, and Yorùbá). Nollywood is the home for Nigerian movies that depict the Nigerian people and reflect the diversities across Nigerian cultures. Our choice of this domain is because Nollywood is the second-largest movie and film industry in the world by annual output[2], and the availability of Nollywood reviews on several online websites. However, most of these online reviews are only in English. To cover more languages, we asked professional translators to translate about 1,000-1,500 reviews

[1] https://www.census.gov/popclock/print.php?component=counter

[2] https://www.masterclass.com/articles/nollywood-new-nigerian-cinema-explained

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 986–998

from English to four Nigerian languages, similar to Winata et al. (2023). Thus, **NollySenti** is a **parallel multilingual sentiment corpus** for five Nigerian languages that can be used for both *sentiment classification* and *evaluation of machine translation (MT) models* in the user-generated texts domain — which is often scarce for low-resource languages.

Additionally, we provide several supervised and transfer learning experiments using classical machine learning methods and pre-trained language models. By leveraging transfer learning, we compare the performance of cross-domain adaptation from the Twitter domain to the Movie domain, and cross-lingual adaptation from English language. Our evaluation shows that transfer from English in the same target domain leads to more than 5% improvement in accuracy compared to transfer from the Twitter domain in the same target language. To further mitigate the domain difference, we leverage MT from English to other Nigerian languages, which leads to a further improvement of 7% over cross-lingual evaluation. While MT to low-resource languages are often of low quality, through human evaluation, we show that most of the translated sentences preserve the sentiment in the original English reviews. For reproducibility, we have released our datasets and code on Github[3].

2 Related Work

African sentiment datasets There are only a few sentiment classification datasets for African languages such as Amharic dataset (Yimam et al., 2020), and NaijaSenti (Muhammad et al., 2022)—for Hausa, Igbo, Nigerian-Pidgin, and Yorùbá. Recently, Muhammad et al. (2023) expanded the sentiment classification dataset to 14 African languages. However, all these datasets belong to the social media or Twitter domain. In this work, we create a new dataset for the Movie domain based on human translation from English to Nigerian languages, similar to the NusaX parallel sentiment corpus for 10 Indonesia languages (Winata et al., 2023).

MT for sentiment classification In the absence of training data, MT models can be used to translate texts from a high-resource language like English to other languages, but they often introduce errors that may lead to poor performance (Refaee and Rieser, 2015; Poncelas et al., 2020). However, they do have a lot of potentials especially when translating between high-resource languages like European languages, especially when combined with English (Balahur and Turchi, 2012, 2013). In this paper, we extend MT for sentiment classification to four low-resource Nigerian languages. This paper is an extension of the YOSM paper (Shode et al., 2022) – A Yorùbá movie sentiment corpus.

3 Languages and Data

3.1 Focus Languages

We focus on four Nigerian languages from three different language families spoken by 30M-120M.

Hausa belongs to the Afro-Asiatic/Chadic language family with over 77 million speakers (Eberhard et al., 2021). It is a native to Nigeria, Niger, Chad, Cameroon, Benin, Ghana, Togo, and Sudan. However, the significant population for the language reside in northern Nigeria. Hausa is an agglutinative language in terms of morphology and tonal with two tones — low and high. It is written with two major scripts: Ajami (an Arabic-based script) and Boko script (based on Latin script) — the most widely used. The Boko script make use of all the Latin letters except for "p,q,v, and x" including the following additional letters "ɓ, ɗ, ƙ, ƴ, kw, ƙw, gw, ky, ƙy, gy, sh, and ts".

Igbo belongs to the Volta–Niger sub-group of the Niger-Congo language family with over 31 million speakers (Eberhard et al., 2021). It is native language to South-Eastern Nigeria, but also spoken in Cameroon and Equatorial Guinea in Central Africa. Igbo is an agglutinative language in terms of its sentence morphology and tonal with two tones - high and low. The language utilizes 34 Latin letters excluding "c,q and x", however, it includes additional letters "ch, gb, gh, gw, kp, kw, nw, ny, ọ, ọ́, ụ and sh".

Nigerian-Pidgin aka Naija is from the English Creole Atlantic Krio language family with over 4 million native speakers and 116 million people second language speakers. It is a broken version of Nigerian English that is also a creole because it is used as a first language in certain ethnic communities (Mazzoli, 2021). It serves as a common language for all as it facilitates communication between several ethnicities. Naija has 26 letters similar to English with an analytical sentence morphology.

[3] https://github.com/IyanuSh/NollySenti

Yorùbá belongs to the Volta–Niger branch of the Niger-Congo language family with over 50 million speakers (Eberhard et al., 2021) thus making it the third most spoken indigenous African language. Yorùbá is native to South-Western Nigeria, Benin and Togo, and widely spoken across West Africa and Southern America like Sierra Leone, Côte d'Ivoire, The Gambia, Cuba, Brazil, and some Caribbean countries. Yorùbá is an isolating language in terms of its sentence morphology and tonal with three lexical tones - high, mid and low - that are usually marked by diacritics which are used on syllabic nasals and vowels. Yorùbá orthography comprises 25 Latin letters which excludes "c, q, v, x, and z" but includes additional letters "gb, ẹ, ṣ and ọ".

3.2 NollySenti creation

Unlike Hollywood movies that are heavily reviewed with hundreds of thousands of reviews all over the internet, there are fewer reviews about Nigerian movies despite their popularity. Furthermore, there is no online platform dedicated to writing or collecting movie reviews written in the four indigenous Nigerian languages. We only found reviews in English. Here, we describe the data source for the Nollywood reviews and how we created parallel review datasets for four Nigerian languages.

3.2.1 Data Source

Table 1 shows the data source for the NollySenti review dataset. We collected 1,018 positive reviews (POS) and 882 negative reviews (NEG). These reviews were accompanied with ratings and were sourced from three popular online movie review platforms - **IMDB**, **Rotten Tomatoes** and, **Letterboxd**. We also collected reviews and ratings from four Nigerian websites like **Cinemapointer**, **Nollyrated**. Our annotation focused on the classification of the reviews based on the ratings that the movie reviewer gave the movie. We used a rating scale to classify the POS or NEG reviews and defined ratings between 0-4 to be in the NEG category and 7-10 as POS.

3.2.2 Human Translation

We hire professional translators in Nigeria and ask them to translate 1,010 reviews randomly chosen from the 1,900 English reviews. Thus, we have a parallel review dataset in English and other Nigerian languages and their corresponding ratings. For quality control, we ask a native speaker per language to manually verify the quality of over 100 randomly selected translated sentences, and we confirm that they are good translations, and they are not output of Google Translate (GT).[4] All translators were properly remunerated according to the country rate[5]. In total, we translated 500 POS reviews and 510 NEG reviews. We decided to add 10 more NEG reviews since they are often shorter – like one word e.g. ("disappointing").

4 Experimental Setup

Data Split Table 2 shows the data split into **Train**, **Dev** and **Test** splits. They are 410/100/500 for `hau`, `ibo` and `pcm`. To further experiment with the benefit of adding more reviews, we translate 490 more reviews for `yor`. The ratio split for `yor` is 900/100/500, while for `eng` is 1,300/100/500. We make use of the same reviews for **Dev** and **Test** for all languages. For our experiments of transfer learning and machine translation, we make use of all the training reviews for English (i.e 1,300). We make use of a larger test set (i.e. 500 reviews) for `hau`, `ibo` and `pcm` because the focus of our analysis is on zero-shot transfer, we followed similar data split as XCOPA (Ponti et al., 2020), COPA-HR (Ljubesic and Lauc, 2021) and NusaX datasets. The small training examples used in NollySenti provides an opportunity for researchers to develop more data efficient cross-lingual methods for under-resourced languages since this is a more realistic scenario.

4.1 Baseline Models

Here, we train sentiment models using classical machine learning models like Logistic regression and Support Vector Machine (SVM) and *fine-tune* several pre-trained language models (PLMs). Unlike classical ML methods, PLMs can be used for cross-lingual transfer and often achieve better results (Devlin et al., 2019; Winata et al., 2023). We fine-tune the following PLMs: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mDeBERTaV3 (He et al., 2021), AfriBERTa (Ogueji et al., 2021), and AfroXLMR (Alabi et al., 2022). The last two PLMs have been pre-trained or adapted to all the focus languages. For XLM-R and AfroXLMR, we make use of the base versions. The classical ML methods were implemented using Scikit-Learn (Pedregosa et al., 2011). Appendix B provides more details.

[4] Easy to verify for languages with diacritics like Yorùbá since GT ignores diacritics. GT does not support Naija

[5] $450 per language except for `yor` with more reviews

Sentiment	No. Reviews	Ave. Length (No. words)	Data source IMDB	Rotten Tomatoes	LetterBoxd	Cinemapoint	Nollyrated	Others
positive	1018	35.0	493	107	81	154	181	2
negative	882	20.7	292	140	101	269	74	6
Total	1900	–	785	247	182	423	255	8

Table 1: **Data source, number of movie reviews per source, and average length of reviews**

Language	Train pos	neg	all	Dev all	Test all
English (eng)	1018	882	1300	100	500
Hausa (hau)	200	210	410	100	500
Igbo (ibo)	200	210	410	100	500
Naija (pcm)	200	210	410	100	500
Yorùbá (yor)	450	450	900	100	500

Table 2: **Dataset split.** The DEV and TEST split have equal number samples in positive and negative classes

4.2 Zero-shot Adaptation

4.2.1 Transfer Learning

Cross-domain adaptation We train on the Twitter domain and perform cross-domain adaptation to the Nollywood movie domain. We make use of the NaijaSenti dataset for training. The datasets consist of between 12k-19k tweets for each of the Nigerian languages, 30 folds larger than our dataset.

Cross-lingual adaptation We train on two English datasets: (1) IMDB (Maas et al., 2011) – with 25,000 reviews and (2) NollySenti English with 1,300 reviews. The resulting models are evaluated on the test set of the remaining Nigerian languages.

4.2.2 Machine Translation

Lastly, we make use of MT to mitigate the domain difference. We make use of NLLB (NLLB-Team et al., 2022)[6] for `hau`, `ibo`, and `yor` languages. NLLB is a multilingual MT trained on 200 languages and dialects. It includes the three Nigerian languages except for Nigerian-Pidgin. For Nigerian-Pidgin, we make use of a pre-trained `eng`→`pcm` MT model by Adelani et al. (2022a) – trained on both religious and news domain.

5 Results

5.1 Baseline Results

Table 3 provides the baseline results using both logistic regression, SVM, and several PLMs. All baselines on average have over 80% accuracy. However, in all settings (i.e. all languages and number of training samples, `N=400,900`, and `1300`), PLMs exceed the performance of classical machine learning methods by over $5-7\%$. In general, we find Africa-centric PLMs (AfriBERTa-large and AfroXLMR-base) have better accuracy than massively multilingual PLMs pre-trained on around 100 languages. Overall, AfriBERTa achieves the best result on average, but slightly worse for English and Nigerian-Pidgin (an English-based creole language) since it has not been pre-trained on the English language.

5.2 Zero-shot Evaluation Results

We make use of AfriBERTa for the zero-shot evaluation since it gave the best result in Table 3 (see avg. excl. `eng`). Table 4 shows the zero-shot evaluation.

Performance of Cross-domain adaptation We obtained an impressive zero-shot result by evaluating a Twitter sentiment model (i.e. `Twitter (lang)`) on movie review (73.8 on average). All have over 70 except for `yor`.

Performance Cross-lingual adaptation We evaluated two sentiment models, trained on either `imdb` or `NollySenti (eng)` English reviews. Our result shows that the adaptation of `imdb` has similar performance as the cross-domain adaptation, while the `NollySenti (eng)` exceeded the performance by over $+6\%$. The `imdb` model (i.e `imdb (eng)`) was probably worse despite the large training size due to a slight domain difference between Hollywood reviews and Nollywood reviews — may be due to writing style and slight vocabulary difference among English dialects (Blodgett et al., 2016). An example of a review with multiple indigenous named entities including a `NEG` sentiment is *"**'Gbarada'** is a typical **Idumota** 'Yoruba film' with all the craziness that come with that subsection of Nollywood. "* that may not frequently occur in Hollywood reviews. Another observation is that the performance of `pcm` was unsurprisingly good for both setups (84.0 to 86.2) because it is an English-based creole.

Machine Translation improves adaptation To mitigate the domain difference, we found that by

[6] https://huggingface.co/facebook/nllb-200-distilled-600M

Model	Parameter size	eng N=410	eng N=1300	hau N=410	ibo N=410	pcm N=410	yor N=410	yor N=900	avg	avg (excl. eng)
LogisticReg	<20K	79.2	84.2	78.8	81.8	83.4	78.8	80.1	$81.0_{\pm 0.2}$	$80.8_{\pm 0.2}$
SVM	<20K	79.0	85.2	79.0	80.6	83.6	79.7	81.9	$81.3_{\pm 0.6}$	$81.0_{\pm 0.6}$
mBERT	172M	90.3	92.6	80.0	82.4	89.1	84.8	87.8	$87.0_{\pm 0.5}$	$85.2_{\pm 0.5}$
XLM-R-base	270M	93.2	**94.1**	76.8	83.6	90.8	83.9	86.0	$86.9_{\pm 0.5}$	$84.2_{\pm 0.5}$
mDeBERTaV3	276M	**94.2**	**95.1**	83.7	87.1	**91.8**	82.2	87.4	$\mathbf{88.8_{\pm 0.5}}$	$86.4_{\pm 0.5}$
AfriBERTa-large	126M	86.2	89.5	**87.2**	**88.4**	88.3	**85.9**	**90.9**	$88.1_{\pm 0.3}$	$\mathbf{88.1_{\pm 0.3}}$
AfroXLMR-base	270M	92.3	94.1	84.2	85.6	91.0	83.8	88.4	$88.5_{\pm 0.8}$	$86.6_{\pm 0.8}$

Table 3: **Baseline result using classical machine learning and pre-trained language models.** We make use of the number of training examples, $N = 410, 900$, and 1300. We report accuracy. Average performed over 5 runs.

	hau	ibo	pcm	yor	ave
Twitter (lang)	76.7	78.4	74.1	66.0	$73.8_{\pm 0.6}$
IMDB (eng)	71.3	71.2	84.0	66.4	$73.2_{\pm 2.2}$
NollySenti (eng)	80.2	78.9	86.2	72.8	$79.5_{\pm 2.9}$
machine translation (en → lang)					
IMDB (lang, N=25k)	86.8	83.8	86.8	82.0	$83.0_{\pm 1.0}$
NollySenti (lang, N=410)	84.0	86.3	81.2	83.0	$83.6_{\pm 0.6}$
NollySenti (lang)	88.3	86.5	87.0	**84.0**	$86.4_{\pm 0.2}$
NollySenti (eng+lang)	**89.5**	**86.8**	**87.2**	83.8	$\mathbf{86.8_{\pm 0.3}}$
Supervised	87.2	88.4	88.3	90.9	$88.7_{\pm 0.3}$

Table 4: **Zero-shot scenario using AfriBERTa-large:** cross-domain (Twitter -> Movie), cross-lingual experiments (eng -> lang) and review generation using machine translation (Meta's NLLB and MAFAND (Adelani et al., 2022a) eng→pcm model)

Lang.	BLEU	CHRF	Adequacy	sentiment preservation
hau	13.6	40.8	4.4	92.0%
ibo	9.8	33.4	3.8	92.0%
pcm	26.4	53.0	4.6	96.0%
yor	3.53	16.9	4.0	89.5%

Table 5: **Automatic** (N=410) and **Human evaluation** (N=100) of the MT generated reviews from TRAIN split.

automatically translating N=410 reviews using a pre-trained MT model improved the average zero-shot performance by over $+4\%$. With additional machine translated reviews (N=1300), the average performance improved further by $+3\%$. Combining all translated sentences with English reviews does not seem to help. Our result is quite competitive to the supervised baseline (-1.9%). As an additional experiment, we make use of MT to translate 25k IMDB reviews, the result was slightly worse than NollySenti (lang). This further confirms the slight domain difference in the two datasets.

Sentiment is often preserved in MT translated reviews Table 5 shows that despite the low BLEU score (< 15) for hau, ibo and yor, native speakers (two per language) of these languages rated the machine translated reviews in terms of content preservation or adequacy to be much better than average (3.8 to 4.6) for all languages on a Likert scale of 1-5. Not only does the MT models preserve content, native speakers also rated their output to preserve more sentiment (i.e. achieving at least of 90%) even for some translated texts with low adequacy ratings. Appendix C provides more details on the human evaluation and examples.

6 Conclusion

In this paper, we focus on the task of sentiment classification for cross-domain adaptation. We developed a new dataset, **NollySenti** for five Nigerian languages. Our results show the potential of both transfer learning and MT for developing sentiment classification models for low-resource languages. As a future work, we would like to extend the creation of movie sentiment corpus to more African languages.

Limitations

One of the limitations of our work is that we require some form of good performance of machine translation models to generate synthetic reviews for sentiment classification. While our approach seems to work well for some low-resource languages like yor with BLEU score of 3.53, it may not generalize to other sequence classification tasks like question answering where translation errors may be more critical.

Ethics Statement

We believe our work will benefit the speakers of the languages under study and the Nollywood industry. We look forward to how this dataset can be used to improve the processes of the Nollywood industry and provide data analytics on movies.

We acknowledge that there maybe some bias introduced due to manually translating the dataset from English, but we do not see any potential harm in releasing this dataset. While the texts were

crawled online, they do not contain personal identifying information.

Acknowledgements

This material is partly based upon work supported by the National Science Foundation under Grant Numbers: 2226006, 1828199, and 1704113. We appreciate Aremu Anuoluwapo for coordinating and verifying the translation of the reviews to the Nigerian languages. We appreciate the collective efforts of the following people: Bolutife Kusimo, Oluwasijibomi Owoka, Oluchukwu Igbokwe, Boluwatife Omoshalewa Adelua, Chidinma Adimekwe, Edward Agbakoba, Ifeoluwa Shode, Mola Oyindamola, Godwin-Enwere Jefus, Emmanuel Adeyemi, Adeyemi Folusho, Shamsuddeen Hassan Muhammad, Ruqayya Nasir Iro and Maryam Sabo Abubakar for their assistance during data collection and annotation, thank you so much. David Adelani acknowledges the support of DeepMind Academic Fellowship programme. Finally, we thank the Spoken Language Systems Chair, Dietrich Klakow at Saarland University for providing GPU resources to train the models.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions*

of the Association for Computational Linguistics, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris C. Emezue, Sana Al-Azzawi, Blessing K. Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Oluwaseyi Ajayi, Tatiana Moteu Ngoli, Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka C. Obiefuna, Shamsuddeen Hassan Muhammad, Saheed Salahudeen Abdullahi, Mesay Gemeda Yigezu, Tajuddeen Rabiu Gwadabe, Idris Abdulmumin, Mahlet Taye Bame, Oluwabusayo Olufunke Awoyomi, Iyanuoluwa Shode, Tolulope Anu Adelani, Habiba Abdulganiy Kailani, Abdul-Hakeem Omotayo, Adetola Adeeko, Afolabi Abeeb, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Raphael Ogbu, Chinedu E. Mbonu, Chiamaka Ijeoma Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge, Sakayo Toadoum Sari, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Ussen Abre Kimanuka, Kanda Patrick Tshinu, Thina Diko, Siyanda Nxakama, Abdulmejid Tuni Johar, Sinodos Gebre, Muhidin A. Mohamed, S. A. Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for african languages. *ArXiv*, abs/2304.09972.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea. Association for Computational Linguistics.

Alexandra Balahur and Marco Turchi. 2013. Improving sentiment analysis in Twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 49–55, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2021. Ethnologue: Languages of the world. twenty-third edition.

Roald Eiselen. 2016. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portorož, Slovenia. European Language Resources Association (ELRA).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nikola Ljubesic and Davor Lauc. 2021. Bertić - the transformer language model for bosnian, croatian, montenegrin and serbian. *ArXiv*, abs/2104.09243.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Maria Mazzoli. 2021. The ideological debate on naijá and its use in education. *English World-Wide*, 42(3):299–323.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Djouhra Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino D'ario M'ario Ant'onio Ali, Davis C. Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Rabiu Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *ArXiv*, abs/2302.08956.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alberto Poncelas, Pintu Lohar, James Hadley, and Andy Way. 2020. The impact of indirect machine translation on sentiment classification. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Eshrag Refaee and Verena Rieser. 2015. Benchmarking machine translated sentiment analysis for Arabic tweets. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 71–78, Denver, Colorado. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. yosm: A new yoruba sentiment corpus for movie reviews.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz,

and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

A Focus Languages

We focus on four Nigerian languages from three different language families. **Hausa** (hau) is from the Afro-Asiatic/Chadic family spoken by over 77 million (M) people. **Igbo** (ibo) and **Yorùbá** (yor) are both from Niger-Congo/ Volta-Niger family spoken by 30M and 46M respectively. While **Nigerian-Pidgin** (pcm) is from the English Creole family, spoken by over 120M people. The Nigerian-Pidgin is ranked the 14th most spoken language in the world[7]. All languages make use of the Latin script. Except for Nigerian-Pidgin, the remaining are tonal languages. Also, Igbo and Yorùbá make extensive use of diacritics in texts which are essential for the correct pronunciation of words and for reducing ambiguity in understanding their meanings.

B Hyper-parameters for PLMs

For fine-tuning PLMs, we make use of HuggingFace transformers (Wolf et al., 2019). We make use of maximum sequence length of 200, batach size of 32, number of epochs of 20, and learning rate of $5e-5$ for all PLMs.

C Human Evaluation

To verify the performance of the MT model, we hire at least two native speakers of each Nigerian indigenous languages - three native Igbo speakers, four native Yorùbá speakers, four native speakers of Nigerian Pidgin and two Hausa native speakers. The annotators were individually given 100 randomly selected translated reviews in Excel sheets to report the adequacy and sentiment preservation (1: if they preserve sentiment, 0:otherwise) of the MT outputs. Alongside the sheets, the annotators are given an annotation guideline to guide them during the course of the annotation. Asides that the annotators are of the Nigerian descent as well as native speakers of the selected languages, their minimum educational experience is a bachelor's degree which qualifies them to efficiently read, write and comprehend the annotation materials and data to be annotated.

To measure the consistency of our annotators, we added repeated 5 examples out of the 100 examples. Our annotators were consistent with their annotation. We measure the inter-agreement among the two annotators per task. For adequacy, the annotators achieved Krippendorff's alpha scores of 0.675, 0.443, 0.41, 0.65 for Hausa, Igbo, Nigerian-Pidgin, and Yorùbá respectively. Similarly, for sentiment preservation, Krippendorff's alpha scores of 1.0, 0.93, 0.48, and 0.52 for Hausa, Igbo, Nigerian-Pidgin, and Yorùbá respectively. In general, annotators reviewed the translated texts to have adequacy of 3.8 and 4.6. Nigerian-Pidgin (4.6) achieved better adequacy result as shown in Table 5 because of her closeness to English language, Igbo was rated to have a lower adequacy score (3.8). Overall, all annotators rated the translated sentences to preserve sentiment at least in 90% of the time i.e 90 out of 100 translations preserve the original sentiment in the English sentence.

C.1 Qualitative analysis

The human evaluation is to verify the manually verify the quality of over 100 randomly selected translated sentences manually. Also, the reports from the annotators were automatically computed to support our claim that sentiment is usually preserved in MT outputs. The examples listed in Table 6 are extracted during the annotation process. The examples illustrate the noticeable mistakes in MT outputs. The annotators are expected to give a rating scale between 1-5 if the randomly selected machine translated review is adequately translated and a binary 0-1 rating scale if the sentiment of the original review is retained in the the randomly selected machine translated review.

The examples that are listed in Table 6 buttress our claim that MT outputs are not completely accurate as some translations in the target languages are missing thereby affecting the complete idea and meaning of the movie review that is originally

[7]https://www.ethnologue.com/guides/ethnologue200

English Translation	Target Language Translation	Literal Translation of Target language
Target Language: Yorùbá **Incorrect translation, sentiment not preserved.**		
In the absence of such a perfect storm, avoid stabbing your wallet in the heart with this 'Dagger'. Definitely not recommended	Níwòn bí k'o ti sí 'ijì líle tó dára, má ṣe fi "Dagger" yìí pa owó rẹ ní ọkàn rẹ.	In the absence of a great storm, do not use this "Dagger" to kill your money in the heart
Incorrect translation, sentiment preserved.		
Citation the movie. Perfect Movie. Loved every second of the movie. Wished it didn't end	Mo fẹ́rà gbogbo ìṣẹ́jú tí mo fi ń ṣe fîmù náà, mo fẹ́ kí ó máà parí	I enjoyed every second that I used to make this movie. Wished it did not end
Incorrect and Incomplete translation, sentiment not preserved		
Funny Funny Funny. Oh mehn, this movie is super funny. if you are looking for a movie to lift your mood up then this is the right movie for you .	Orinrinrinrinrinrin...	song..... (MT output is nonsensical)
Target Language: Igbo **Incorrect translation, sentiment not preserved.**		
Fifty minutes is spent advertising a holiday resort in Lagos, Movie closes. Money down the drain. Not recommended.	Ọ bụrụ na ị na-eme ihe ndị a, ị ga-enwe ike ịhapụ ya.	Do these things to leave it
Incorrect translation, sentiment preserved.		
Temi Otedola's performance was truly stunning. I thoroughly enjoyed the layers that the story had and the way that each key piece of information was revealed.	Ihe a o mere tọrọ m ezigbo ụtọ, ọ na-atọkwa m ụtọ otú e si kọwaa ihe ndị dị mkpa.	I thoroughly enjoyed the layers that the story had and the way that each key piece of information was revealed.
Incorrect and Incomplete translation, sentiment not preserved		
Nice cross-country movie. The only thing that I don't like about this movie is the way there was little or no interaction with the Nigerian or Indian environment. Beautiful romantic movie .	Ihe m na-adịghị amasị na fim a bụ na ọ dịghị ihe jikọrọ ya na ndị Naịjirịa ma ọ bụ ndị India.	The only thing that I don't like about this movie is the way there was little or no interaction with the Nigerian or Indian environment
Target Language: PCM - Nigerian Pidgin		
Incorrect translation, sentiment preserved.		
Nice cross-country movie . The only thing that I don't like about this movie is the way there was little or no interaction with the Nigerian or Indian environment. Beautiful romantic movie .	The only thing wey I no like about this film na because e no too get interaction with Nigerian or Indian people.	The only thing that I don't like about this movie is the way there was little or no interaction with the Nigerian or Indian people.
Incorrect translation, sentiment preserved.		
A flawed first feature film , but it shows a great deal of promise	Fear first feature film, but e show plenti promise.	Fear was featured in the film firstly but it shows a great deal of promise
Incorrect and Incomplete translation, sentiment not preserved		
Spot On!!! Definitely African movie of the year, enjoyed every minute of the 2hours 30minutes	Na almost every minute of the 2hours 30minutes wey dem take play for Africa film dem dey play.	It is almost every minute of the 2hours 30minutes that they play African movie they play

Table 6: **Examples of translation mistakes observed and impact on the sentiment**. The Gray color identifies the sentiment portion of the review

written in English, which eventually could lead to losing the sentiment of the movie review. Also, as shown in Table 6, the sentiments of some reviews are preserved regardless of the incorrect or missing translations and the idea or meaning of the review is not totally lost.

C.2 Annotation Guideline

We provide the annotation guideline on Github[8].

[8]

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6 (Limitation)

☑ A2. Did you discuss any potential risks of your work?
6 (Ethics Statement)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract; 1 - Introduction

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

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3,4,5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
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)?
3, 5, 6 (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?
6 (Ethics Statement)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ 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

C ☑ Did you run computational experiments?

4,5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4,5

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?
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?
5

☑ 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.)?
4, 5

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

3, Appendix (C)

☑ 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.?
3, Appendix (C)

☑ 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)?
3, Appendix (C)

☑ 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?
6 (Ethics Statement)

☐ 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?
3, Appendix (C)