Temporal and Second Language Influence on Intra-Annotator Agreement and Stability in Hate Speech Labelling

Temporal and Second Language Influence on Intra-Annotator Agreement and Stability in Hate Speech Labelling

Gavin Abercrombie
Heriot-Watt University
g.abercrombie@hw.ac.uk

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Vinodkumar Prabhakaran
Google Research
vinodkpg@google.com

Abstract

Much work in natural language processing (NLP) relies on human annotation. The majority of this implicitly assumes that annotator's labels are temporally stable, although the reality is that human judgements are rarely consistent over time.

As a subjective annotation task, hate speech labels depend on annotator's emotional and moral reactions to the language used to convey the message. Studies in Cognitive Science reveal a 'foreign language effect', whereby people take differing moral positions and perceive offensive phrases to be weaker in their second languages. Does this affect annotations as well?

We conduct an experiment to investigate the impacts of (1) time and (2) different language conditions (English and German) on measurements of intra-annotator agreement in a hate speech labelling task. While we do not observe the expected lower stability in the different language condition, we find that overall agreement is significantly lower than is implicitly assumed in annotation tasks, which has important implications for dataset reproducibility in NLP.

1 Introduction

While *inter*-annotator agreement is commonly used in natural language processing (NLP) research to measure annotation *reliability* (how well annotators agree with each other) (Carletta, 1996), *intra*-annotator agreement (the extent to which individuals provide the same responses for the same prompts when asked repeatedly) is rarely reported (Abercrombie et al., 2023).

However, measurements of intra-annotator agreement are essential for NLP datasets as they indicate the consistency of each human annotator and thus the stability of the data they generate (Teufel et al., 1999). Intra-annotator measures can be used to control the quality of the annotation process (e.g. Akhbardeh et al., 2021; Cao et al., 2022; Hengchen and Tahmasebi, 2021) or to assess the difficulty and subjectivity of a particular task (Abercrombie et al., 2023). The field's continuing failure to measure and report intra-annotator agreement, though, suggests that it is implicitly assumed (by omission) that annotators' responses are 100% stable—even when this is intuitively and empirically not the case. In fact, there is widespread evidence from Psychology that the same people often make wildly inconsistent judgments depending on seemingly unrelated factors such as mood, time of day, the weather, or even how well their preferred sports team is performing (Kahneman et al., 2021). Here, we consider the following two aspects:

Time: There is some evidence that annotator inconsistency increases as a function of time (Kiritchenko and Mohammad, 2017; Li et al., 2010). However, in the majority of cases in which intra-annotator agreement *is* reported, repeat annotations are collected in the same session that annotators label the items in the first instance (Abercrombie et al., 2023). In those circumstances, annotators' responses are likely influenced by the recency effects of priming on memory (Vriezen et al., 1995). In this study, we therefore re-examine annotators after substantial temporal intervals of between two and eight weeks.

Language: The kind of language to annotate is likely to affect annotations. Annotating or producing abusive language, such as hateful speech, can be understood as "morally motivated behavior[s] grounded in people's moral values and perceptions of moral violations" (Hoover et al., 2019). However, there is evidence that people take different moral positions when presented with dilemmas in their first or second languages—the '*foreign language effect*' (Costa et al., 2019; Stankovic et al., 2022).

Furthermore, Dewaele (2004) shows that bilingual people perceive the emotional force of swearwords and taboo words to be weaker in their second

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 96–103

languages, suggesting that they may judge toxic language differently when observed in different languages. We can therefore expect annotators to respond differently to text examples from hateful language datasets that feature moral issues and toxic slurs in their first (L1) or second language (L2). In this work, we investigate the stability of labels produced by bilingual annotators in response to near-identical (i.e., carefully translated) examples presented in both English and German.

We ask the following **Research questions**:

R1: Are annotators' responses **stable over time** when labelling hateful language?

R2: Is label stability lower when repeated annotation items are presented **in a different language** than in the same language?

2 Bilingual hateful speech data

We use the `XHate999` corpus (Glavaš et al., 2020), the test set of which consists of (999) abusive and non-abusive texts that have been translated from English to five other target languages. Translations were made by experts with an emphasis on maintaining the level and nuance of abuse, hatefulness, and aggression present in the texts. We use the English and German language versions of the '*Gao*' hatefulness subset (Gao and Huang, 2017) (the data is originally sourced from three separate datasets). We chose German as it is the most widely-spoken of the target languages, which we expect to expediate annotator recruitment. We chose the '*Gao*' subset as we judged the domain or language and topics of the other two to be somewhat esoteric for primarily Europe-based annotators: Wulczyn et al. (2017) consists of disputes on the content of Wikipedia pages among their authors; and Kumar et al. (2018) is comprised of Hindi-English political discussions. While many examples from the *Gao* subset concern specific events, we expected the subject matter (e.g. the #BlackLivesMatter movement, Israel-Palestine conflict) to be more well-known internationally. The test set comprises 99 items.[1]

3 Experimental Setup

We recruited 30 billingual German (L1) and English (L2) speaking annotators from the Prolific crowdsourcing platform,[2] chosen for its capacity to facilitate longitudinal studies, ethical participant payment policies, and data quality (Peer et al., 2022). We presented the particpants with 96 examples from the test data: 48 in English and 48 in German. With these, we interspersed four items taken from HATECHECK (Röttger et al., 2021), which we used as attention check questions (Abbey and Meloy, 2017), as they were designed to be clearcut examples of hateful language.

We use a 'descriptive dataset paradigm' (Rottger et al., 2022), with annotators provided with minimal instructions to encourage the emergence of subjective perspectives. As such, annotators were presented with the original definition of hateful language from Gao and Huang (2017):

> *We define hateful speech to be language which explicitly or implicitly threatens or demeans a person or a group based upon a facet of their identity such as gender, ethnicity, or sexual orientation.*

We also provided one example each of hateful and non-hateful items taken from the three unused test set items. In order to maintain concentration and to regularly provide the option of withdrawing participation, we split the task into 20 pages with five items to be annotated on each. The instructions, definition, and examples were repeated on each page, and are available in Appendix A.

We made the task available to all German-speakers on the platform that are also fluent in English, as managed by Prolific's in-built participant filtering functions. Based on the findings of Kiritchenko and Mohammad (2017) and Li et al. (2010), we then waited two weeks before opening a second round of the task to the same annotators, filtering by their Prolific identification codes. Here, the annotators were presented with a further four attention check items and the 96 items from the test set. Again, half of these were in English and half in German. 50% were presented in the same language as in round one, and 50% in the alternative language. To control for order-effect bias, we split participants into two groups, and presented the items to each in a different order, also shuffling within-item response options (*hateful/not hateful*).

Following the principles of *perspectivist* data practices (Abercrombie et al. (2022); Cabitza et al. (2023)) and the recommendations of Prabhakaran et al. (2021), full set of collected labels is available at `https://github.com/`

[1] Available at `https://github.com/codogogo/xhate/tree/main/test/en`

[2] `https://www.prolific.co/`

GavinAbercrombie/XHateStability. For reproducibility, we include the question item order and full instructions. We also provide a full data statement and annotator demographic information in Appendix B.

4 Analysis

Of the 28 participants that completed both rounds of annotation, 22 labelled all attention check items in agreement with the original HATECHECK labels, and we report results for these annotators only.[3]

4.1 Reliability

To evaluate reliability, we report Fleiss' *kappa*, which can account for multiple annotators to show overall agreement, as well as average pair-wised Cohen's *kappa* and raw percentage agreement for completeness. We also evaluate agreement between the labels most commonly assigned by our participants (majority vote) with the original labels collected by Gao and Huang (2017).

	Billingual participants			**Majority vote *v.* Original labels**	
	Fleiss	Cohen	%	Cohen	%
All	0.28	0.29	64.2	0.44	71.7
EN	0.29	0.29	64.6	0.48	74.0
DE	0.27	0.27	63.4	0.40	68.9

Table 1: Reliability as measured by inter-annotator agreement (Fleiss' and Cohen's κ and raw percentage agreement). Cohen's κ and % are calculated pairwise.

As shown in Table 1, the participants do not agree with each other to a high degree. *Kappa* scores for agreement between them are below 0.3, suggesting that the task is highly subjective.

Aggregating their responses by majority vote and comparing to the original labels, produces similarly modest agreement ($\kappa < 0.5$). This somewhat calls into question the reliability of the original *Gao* labels, on which the authors reported almost perfect agreement between two annotators. All agreement measurements are poorer still on the German examples, which also casts some doubt on the feasability/validity of translating text and keeping the labels applied to items annotated in a different language, as was the case for XHate999.

[3]Although, agreement scores are, in fact, comparable when including all annotators, to ensure quality, we do not include those that did not pass all attention checks.

4.2 Stability

As we argue in section 1, most dataset developers implicitly assume annotator consistency to be 100% stable. We therefore use raw percentage agreement as the primary metric for stability and examine deviations from full agreement. For completeness, we also report Cohen's *kappa* (Table 2).

		κ	%
All items		0.49	74.5
Same language	All	0.44	72.3
	EN	0.43	71.6
	DE	0.45	72.9
Different language	All	0.53	76.9
	EN→DE	0.54	77.2
	DE→EN	0.53	76.6

Table 2: Stability as measured by intra-annotator agreement (Cohen's κ and raw percentage agreement).

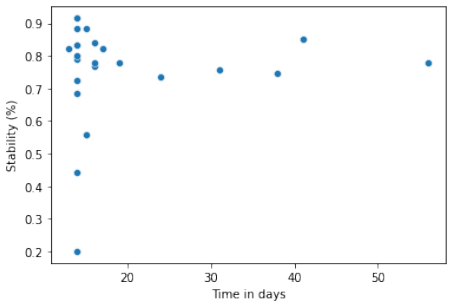

Figure 1: Stability of individual annotators over time measured by raw percentage intra annotator agreement.

Overall stability over time At under 75% and $\kappa = 0.49$, stability is low overall.[4] This is considerably worse than the only reported intra-annotator agreement we are aware of on a similar task: $\kappa = 0.89$ on abusive language detection (Cercas Curry et al., 2021), where those annotations were made by experts under higher levels of supervision.

Consistency varies considerably among the annotators ($max = 91.6\%$, $min = 20.0\%$, $\mu = 74.5\%$, $\sigma = 15.7\%$), with even the most consistent falling considerably below 100%. After the minimum two week interval, we do not see a pattern of further deterioration in intra-annotator agreement, as shown (from limited datapoints) in Figure 1. This lends further support to the findings of Kiritchenko and Mohammad (2017) and Li et al. (2010), that this interval may be sufficient for re-annotation.

[4]Stability of the majority vote on each item is somewhat more stable: $\kappa = 0.77$, 88.4%

Feature	Examples	Prevalence (%)	Reliability	Stability
Length in tokens (normalized $[0, 1]$)	—	100.0	−0.02	−0.05
Different language	—	50.0	n/a	0.03
Identity terms	*feminists*, *Juden*	37.9	0.06	0.07
Named entities	*Africa*, *Hillary*	33.7	0.00	0.03
Nature terms	*alligator*, *Lagune*	21.1	0.01	0.04
Offensive terms	*Blödmann*, *scumbags*	12.6	0.00	0.03
Political terms	*feminists*, *Liberalismus*	32.6	0.03	−0.01
Quote	—	11.6	0.06	0.03
Original label = *hateful*	—	41.4	0.00	0.00

Table 3: Regression coefficients for hand-crafted features with example terms and their prevalence in the data by percentage of text examples they feature in. The dependent variables are *reliability* and *stability*.

Language effect We do not see the expected difference between items re-annotated in the different conditions. Indeed, stability is actually slightly higher in the different language condition. However, this effect is not uniform across the participants, with around a third (8 of the 22) exhibiting more consistency for the same language condition.

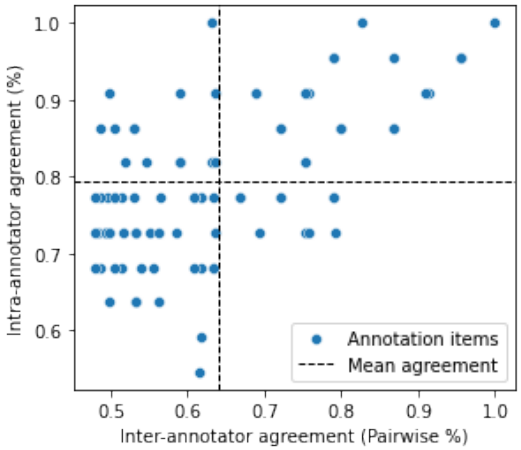

Figure 2: Inter- and intra-annotator agreement by-item.

4.3 By-item agreement

Intuitively, some texts are more straightforward than others to label. We would therefore expect to see variation in intra- (and inter-) annotator agreement between annotation items, and this is indeed the case. Pearson's r for the correlation between raw inter- and intra-annotator agreement is 0.62, indicating a fairly strong, but not perfect relationship between reliability and stability. Following Abercrombie et al. (2023), we can interpret the items furthest towards the top-right of Figure 2 as *straightforward*, those near the top-left as *subjective*, and those in the bottom-left as *ambiguous/difficult*.

To investigate which factors contribute to stability and reliability in this data, we manually labelled each item with a set of hand-crafted features designed to capture the linguistic and world knowledge that intuitively seem necessary to infer whether these texts are hateful. These include the ratio variable *length* in tokens, as well as binary variables based on: the original label assigned to the item; inclusion of quotations; and presence of certain unigram tokens (such as identity terms) in the text. We ran regression analyses on these features with the by-item inter- and intra-annotator agreement scores as the dependent variables.

Table 3 shows the coefficients of each hand-crafted feature for both reliability and stability. None of the features are strongly indicative of either, although several, such as *identity terms* and *text length* do have slightly larger coefficients (positive or negative) than the different language condition. Ultimately, the data sample is not large enough to surface the feature patterns indicative of the ambiguities that provoke annotator disagreements and inconsistencies.

5 Conclusion

While attention has been paid to noise in linguistic annotations caused by factors such as subjectivity and ambiguity (e.g. Aroyo and Welty, 2015; Basile et al., 2021; Prabhakaran et al., 2021), and the level and quality of annotator attention (e.g. Hovy et al., 2013; Klie et al., 2023), this study represents an initial foray into a hitherto understudied aspect of the human labeling work that most NLP research and systems are built upon: intra-annotator agreement and label stability. For this hateful language annotation task, we find that label stability is far lower than common practise implicitly implies (**R1**).

In this study, presenting the items for re-annotation in a different language does not lead to lower stability overall (**R2**), with L1 German speakers no less consistent—and often more so—than when re-annotating items in the same language. We suspect that our data sample of 96 items is too

small to disentangle any L2 effects from other factors that may affect label stability. However, we see lower agreement overall (inter- and intra-) on the German language items, suggesting that the translation process adds some ambiguity. Future work should investigate the linguistic and cultural factors that influence annotators' judgments more closely and—on a larger set of items.

Despite the limitations of our study, we have shown that annotator stability, along with reliability, is necessary for the repeatability and reproducibility of annotation studies (Teufel et al., 1999). We therefore recommend that researchers and practitioners measure and report *intra-* (as well as *inter-*) annotator agreement scores for the labeled NLP datasets they create. The fact that this measure is still rarely reported adds to the emerging reproducibility issues in the field (Belz et al., 2023).

Ethical Considerations

We received approval to conduct these experiments from the institutional review board (IRB) of Heriot-Watt University (ref. 2022-3336-7139).

As annotators were exposed to potentially upsetting language, we took the following mitigation measures:

- Participants were warned about the content (1) before accepting the task on the recruitment platform, (2) in the Information Sheet provided at the start of the task, and (3) in the Consent Form where they acknowledged the potential risks.
- Participants were required to give their consent to participation.
- They were able to leave the study at any time on the understanding that they would be paid for any completed work.
- The task was kept short (all participants completed each round in under 30 minutes) to avoid lengthy exposure to upsetting material.

Following the advice of Shmueli et al. (2021) we paid participants at a rate that was above both the living wage in our jurisdiction and Prolific's current recommendation of at least £9.00 GBP/$12.00 USD.

Acknowledgements

Gavin Abercrombie was supported by the EPSRC projects 'Gender Bias in Conversational AI' (EP/T023767/1) and 'Equally Safe Online' (EP/W025493/1). His visit to Bocconi University was funded by a Scottish Informatics and Computer Science Alliance (SICSA) PECE travel grant. Dirk Hovy received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). He is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. This work was also partly funded by Google Research.

The authors would like to thank Erlint Alitsani, Sabrina McCallum, Elisabetta Pique', and Nikolas Vitsakis for their assistance in developing the annotation task, as well as the anonymous annotators. We are also grateful to Verena Rieser and to Lora Aroyo, Aida Davani, Chris Homan, Chris Welty and others at Google for helpful and insightful discussions on this topic, and to the LAW programme committee for their useful feedback.

References

James D. Abbey and Margaret G. Meloy. 2017. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53-56:63–70.

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki

Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Federico Cabitza, , Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Washington DC, USA.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Costa, Jon Andoni Duñabeitia, and Boaz Keysar. 2019. Language context and decision-making: Challenges and advances. *Quarterly Journal of Experimental Psychology*, 72(1):1–2. PMID: 30803348.

Jean-Marc Dewaele. 2004. The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of Multilingual and Multicultural Development*, 25(2-3):204–222.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Simon Hengchen and Nina Tahmasebi. 2021. SuperSim: a test set for word similarity and relatedness in Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 268–275, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Joseph Hoover, Mohammad Atari, Aida M. Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, Drew Kogon, and Morteza Dehghani. 2019. Bound in hatred: The role of group-based morality in acts of hate.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

D. Kahneman, O. Sibony, and C.R. Sunstein. 2021. *Noise*. Harper Collins Publishers.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, 49(1):157–198.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, and Kazuaki Maeda. 2010. Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Michelle Stankovic, Britta Biedermann, and Takeshi Hamamura. 2022. Not all bilinguals are the same: A meta-analysis of the moral foreign language effect. *Brain and Language*, 227:105082.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Ellen R. Vriezen, Morris Moscovitch, and Sandy A. Bellos. 1995. Priming effects in semantic classification tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4).

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

A Annotator guidelines

The following guidelines were provided to annotators at the beginning of the task, and the definition and examples were repeated at the top of each page of five items. To avoid reprinting potentially offensive text, here we provide the row numbers of the examples from `XHate999-EN-Gao-test`.[5]

Instructions

We define hateful speech to be the language which explicitly or implicitly threatens or demeans a person or a group based upon a facet of their identity such as gender, ethnicity, or sexual orientation.

Read the following 100 posts, which are written in either English or German.

Do you think that they are **Hateful** or **Not hateful**?

If you're not sure, select the option that seems most likely to you.

Examples:

`EN-Gao-test row 100.`
Hateful

`EN-Gao-test row 95.`
Not hateful

B Data statement

We provide a data statement, as recommended by Bender and Friedman (2018).

[5] Available at https://github.com/codogogo/xhate/blob/main/test/en/XHate999-EN-Gao-test.txt

Curation rationale Textual data is from the '*GAO*' subset of `XHate999`, selected for the reasons highlighted in section 2. For further details of the original data collection process, see Gao and Huang (2017). For information on the translations, see Glavaš et al. (2020).

Language variety: `en-US`, `de-DE`. Predominantly US English, as written in comments on the Fox News website. Translated to German by editing automatic outputs of Google Translate. Translators were 'expert' L1 speakers of German who were also fluent in English.

Author demographics: Unknown.

Annotator demographics: The original Gao and Huang (2017) labels were produced by 'two native English speakers', with no further information provided. Annotator demographics for the bilingual labelling are as follows.

- Age: $18 - 70$, $\mu = 33.1$, $\sigma = 12.9$
- Gender: Female: 12 (55%); Male: 10 (45%)
- Ethnicity: White: 19 (86%), Mixed: 3 (14%)
- Native language: German (`de`) 100%
- Socio-economic status:
  - Employment: N/A: 7, Full-Time: 10, Not in paid work: 4, Part-Time: 3, Other: 2
  - Student: Yes: 9, No: 8, N/A: 5
- Training in relevant disciplines: Unknown

Text production situation:

- Time and place: unknown.
- Modality: written, spontaneous, asynchronous interaction.
- Intended audience: other website users.

Text characteristics Comments on articles on the Fox News website. The articles appear to concern events in the United States of America and the wider world in c.2016: Black Lives Matter protests, the Israel-Palestine conflict, and the death of a child at Disney World feature prominently.

Provenance: Data statements were not provided with the original datasets.