Gender Identity in Pretrained Language Models: An Inclusive Approach to Data Creation and Probing

Urban Knupleš[1] and **Agnieszka Falenska**[12*] and **Filip Miletić**[1*]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
{urban.knuples,agnieszka.falenska,filip.miletic}@ims.uni-stuttgart.de

Abstract

Pretrained language models (PLMs) have been shown to encode binary gender information of text authors, raising the risk of skewed representations and downstream harms. This effect is yet to be examined for transgender and non-binary identities, whose frequent marginalization may exacerbate harmful system behaviors. Addressing this gap, we first create TRANSCRIPT, a corpus of YouTube transcripts from transgender, cisgender, and non-binary speakers. Using this dataset, we probe various PLMs to assess if they encode the gender identity information, examining both frozen and fine-tuned representations as well as representations for inputs with author-specific words removed. Our findings reveal that PLM representations encode information for all gender identities but to different extents. The divergence is most pronounced for cis women and non-binary individuals, underscoring the critical need for gender-inclusive approaches to NLP systems.

1 Introduction

Gender identity – an individual's sense of self, reflected in their experience and perception of their gender – is closely connected to language. Sociolinguistic research shows that speakers intentionally use language to construct their gender identity (Eckert and McConnell-Ginet, 1992); such linguistic practices evoke links with identity when perceived by others (Eckert, 2008). Crucially for NLP, this means that gender information is inherent in all linguistic data and may be inadvertently learned by computational systems.

In a stark illustration, Lauscher et al. (2022) conduct a probing study to show that pretrained language models (PLMs) encode binary gender in their representations. But are transgender and non-binary identities encoded to the same extent?[1] This question is vital because gender information may have a skewed distribution in the data, which may be exacerbated through pretraining and propagated to downstream tasks. PLM-based systems tend to make more errors for texts authored by underrepresented groups, particularly in tasks such as predicting psychometric characteristics and personality traits (Lalor et al., 2022). These biases may result in allocation harms and have significant real-world consequences (Blodgett et al., 2020; Hossain et al., 2023). Regarding gender identity, harmful system behaviors are likely to disproportionately affect transgender and non-binary people since they are generally underrepresented (Dev et al., 2021; Devinney et al., 2022). This issue reflects a broader

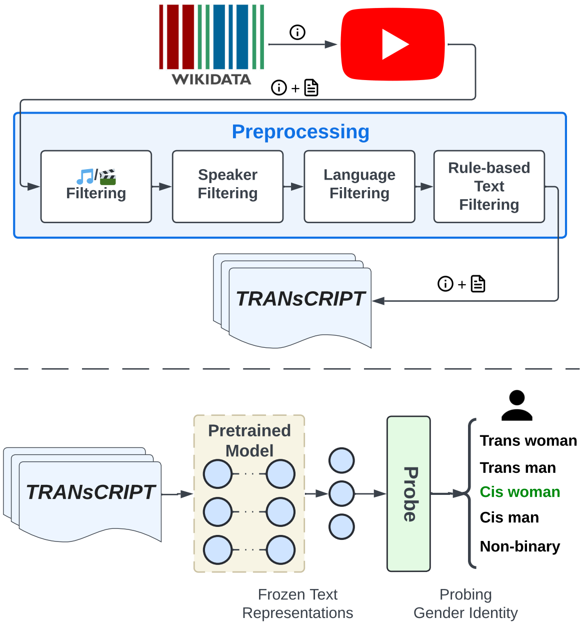


Figure 1: Overview of our corpus creation and probing.

[*] Equal contribution.

[1] Following Zimman and Hayworth (2020b), we use the term *transgender* to refer to individuals whose gender identity differs from the gender assigned to them at birth; *cisgender* to those who identify with their assigned gender; and *non-binary* to all those who do not identify as exclusively female or male.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11612–11631
November 12-16, 2024 ©2024 Association for Computational Linguistics

need for gender-inclusive approaches to language, centering the experiences of trans and non-binary people (Zimman, 2019, 2020).

We adopt precisely such an approach to investigate gender information in PLMs from an inclusive perspective. We create TRANSCRIPT, a corpus of texts by transgender, cisgender, and non-binary speakers (see Figure 1 for an overview). It contains transcripts of YouTube videos whose creators were sampled based on public sociodemographic information from Wikidata (see Section 8 for an explanation of the data licensing). We use our corpus to replicate the probing methodology by Lauscher et al. (2022) on a broader range of identities, and extend it to account for interactions with topic (Bamman et al., 2014; Dayanik and Padó, 2021). We pose the following research questions:

RQ1 Do PLM representations encode gender identity beyond the male–female binary?

RQ2 Does the encoded information reflect author identity rather than gender?

RQ3 Is gender identity encoded differently after fine-tuning on an explicit label?

Our approach uses various PLMs to encode segments of transcripts and then trains a probing classifier to predict the gender identity of the transcript's author. In this setting, gender classification is used as a proxy task to examine PLM representations: above-chance performance is indicative of encoded gender information, and relatively higher performance for a given class suggests a stronger bias towards it. We provide the following contributions: (1) A comprehensive pipeline to collect YouTube transcripts, resulting in 6,000 texts by 168 individuals across five gender identities.[2] (2) A battery of probing experiments which consistently show that PLM representations encode all gender identities, but to different extents. The divergence is strongest for cis women and non-binary individuals, confirming the relevance of our gender-inclusive approach. (3) A novel author-controlled probing method to distinguish the effect of author and gender identity. This approach confirms that author-related information, such as the use of specific words, is also encoded in PLM representations, but does not affect the encoding of gender.

2 Related Work

NLP and gender identity. Different linguistic expressions may convey the same meaning, but also reflect the social identity of the speaker; this is what sociolinguistics defines as the social meaning of linguistic variation (Eckert, 2008). Illustrating this pattern, Bamman et al. (2014) examine lexical choices on social media with respect to assigned binary gender. They identify linguistic features linked to gender identity, but also find that alignments between gender and language use are variable, not necessarily binary, and dependent on interaction. From a different perspective, such linguistic variation is exploited to predict gender on the task of authorship profiling (Argamon et al., 2009; Das and Paik, 2021; HaCohen-Kerner, 2022).

Another line of work focuses on biased gender representations learned from training data (Caliskan et al., 2017). Devinney et al. (2020) analyze corpora in terms of topics specific to masculine, feminine, and non-binary gender. Genders are treated differently in all datasets, pointing to a misrepresentation of disadvantaged groups and a risk for the biases to transfer to PLMs. Dev et al. (2021) investigate BERT's predictions of pronouns for non-binary individuals, noting its tendency to misgender even when given additional context. Cao and Daumé III (2021) observe that NLP research tends to make strong binary assumptions around gender identity, ignoring the existence of trans and non-binary individuals. This overall highlights the need to better understand the limitations of NLP models with respect to marginalized communities.

Probing PLMs. The knowledge encoded in PLMs' internal representations is generally analyzed using probing techniques (Hupkes et al., 2018; Tenney et al., 2019; Belinkov, 2022). A probe is usually a simple classifier trained to predict a property of interest on a model's frozen representations (Belinkov et al., 2017; Petroni et al., 2019). Its performance is taken to indicate the information encoded in the representations. However, probing classifiers may memorize patterns in the data which are unrelated to representational properties (Hewitt and Liang, 2019). Therefore, alternative probing techniques are based on an information-theoretic approach, such as the minimum description length (MDL) probe (Voita and Titov, 2020).

Different techniques have been applied to probe syntactic structures (Hewitt and Manning, 2019; Linzen and Baroni, 2021), lexical semantics (Vulić

[2] Code for reconstructing the corpus along with experiments: https://github.com/urbikn/inclusive-probing-of-gender-identity.

et al., 2020; Mickus et al., 2020), and factual knowledge (Petroni et al., 2019; Zhong et al., 2021). In directly relevant work, Lauscher et al. (2022) assess sociodemographic knowledge: age and binary gender. They use a probing classifier and an MDL probe to analyze PLMs of different sizes, comparing frozen and fine-tuned representations across layers. They find that sociodemographic information *is* encoded, but also note the binary framing of gender as a limitation due to a lack of gender-inclusive datasets. It is to be determined if trans and non-binary identity is encoded in the same way.

Gender-inclusive datasets. Existing datasets generally treat gender as a binary variable, leading the studies that use them to adopt the same framing (Dayanik and Padó, 2021; Lauscher et al., 2022; Orgad et al., 2022, among others). One exception is GiCoref, a dataset aimed at analyzing coreference resolution systems (Cao and Daumé III, 2021). It includes neopronouns (e.g. *ze/hir*, *xe/xem*) and articles about non-binary individuals. From a sociolinguistic perspective, Zimman and Hayworth (2020a) create the TransLiveCorpus using posts from four online communities aimed at transgender and non-binary people. These examples incorporate gender-inclusive data based on the topic or target audience; we are unaware of similarly inclusive datasets with author-level information.

3 TRANSCRIPT Corpus

In order to investigate the encoding of gender in PLM representations from an inclusive perspective, we require texts produced by authors of different gender identities. Faced with a lack of such corpora, we construct TRANSCRIPT: the **TRA**nsgender **N**on-binary **C**isgender transc**RIPT**s corpus.

Due to recent restrictions on sources of sociodemographically enriched text such as X and Reddit (Davidson et al., 2023), we devise an alternative approach (Figure 1). We use Wikidata to sample English-speaking, famous YouTubers across gender identities, collect transcripts of their YouTube videos, and then filter them to ensure data quality. YouTube transcripts are a unique domain, comprising segments of prepared speech (e.g., interviews) as well as spontaneous conversations, opening possibilities for analyzing various language variations, such as regional origin (Coats, 2023).

We now describe our pipeline in more detail and present a topic analysis of the collected content.

3.1 Corpus Creation Pipeline

Using Wikidata to sample authors. Wikidata is an open, collaborative database containing structured data about real-world entities (Vrandečić and Krötzsch, 2014); it has also been used to study PLMs (Petroni et al., 2019; Meng et al., 2022). Each item contains standardized property–value pairs describing its characteristics; e.g. the item "Chris Hadfield" has the property "occupation" with the value "astronaut". Importantly for us, Wikidata contains information on YouTubers.

We access the Wikidata Query Service[3] and select entries that (1) are an instance of *human*; (2) contain properties *sex or gender*, *date of birth*, *country of citizenship*, and *YouTube channel ID*; (3) are citizens of the *USA* or the *UK*, so as to prioritize English speakers; (4) contain properties *number of subscribers* and *number of viewers/listeners* of the YouTube channel, so as to control for online presence and influence; (5) for the *sex or gender* property, contain the value *trans woman*, *trans man*, *female*, *male*, or *non-binary*.[4] We find four instances where the *sex or gender* property has two values, which may occur when information predating gender transition is not removed; we retain the more recent value.

The cisgender groups are significantly overrepresented on Wikidata (Appendix A.1). Therefore, we apply a sampling procedure, in which transgender and non-binary individuals are the "target group" and cisgender individuals "control group". We retain control group candidates only if they can be matched with a target group individual based on (i) country of citizenship; (ii) age, in five-year increments; and (iii) level of YouTube fame (log-transformed subscriber and view counts in ±0.5 range). This ensures that gender identity (rather than e.g., age) remains the main distinguishing feature for the two groups.

Collecting YouTube transcripts. For the next step, we direct our attention to YouTube, a social media platform containing user-uploaded videos. YouTube includes vast amounts of conversational content such as interviews and vlogs.[5] We focus on videos which include transcripts uploaded by users, usually channel owners themselves (Lakomkin

[3] query.wikidata.org/

[4] See Appendix A.1 for corresponding Wikidata codes.

[5] Vlogs are video blogs in which creators document their lives or engage in discussions while facing the camera (Biel and Gatica-Perez, 2010).

et al., 2018), because they are likely reviewed for accuracy and alignment with dialogue.

We start from the YouTubers identified on Wikidata and gather their channels from the *YouTube channel ID* property; if multiple channels are provided, we select all of them. Next, we compile a list of all videos from these channels and download respective manually created transcripts with libraries `yt-dlp`[6] and YouTube Transcript/Subtitle API.[7] We retain only English transcripts with metadata language codes `en`, `en-US`, and `en-GB`.

A user-uploaded transcript is comprised of segments. Each segment is a verbatim transcription of a speaker's utterance in a specific part of the video, associated with the start time and duration in seconds. We additionally extract all metadata from the associated YouTube videos using the `pytube` library,[8] including the video title and description.

Preprocessing. After collecting the data, we apply four filtering steps to limit any noise. (1) We exclude transcripts of music videos and movie trailers, identified from video metadata (`musicVideoType` attribute) and string-based heuristics (video title containing *music video*, *cover*, or *trailer*). (2) Some transcripts include segments from multiple speakers. Therefore, we implement a speaker diarization system on the audio of the transcribed videos and leave only segments spoken by the YouTuber of interest (see Appendix A.2). (3) Our transcripts are filtered for English based on metadata, but out of precaution we also run `py3langid`[9] (Lui and Baldwin, 2012) and exclude transcripts predicted to be non-English. (4) Similarly to Lakomkin et al. (2018), we apply rule-based filters to exclude irrelevant information: transcripts with a duration less than 1 second; non-ASCII characters; URL-specific patterns *http** and *www**; speaker markings (*Speaker 1:*, *Person:*) and other annotations (*[laughs]*, *(laughs)*, \laughs\\).

As final preprocessing steps, we tokenize the corpus with `spaCy` (Honnibal and Montani, 2017) using the `en_core_web_md` model. We concatenate continuous transcript segments and slice them into sequences of up to 256 tokens, using a 240-token sliding window for context. These are the segments referenced in the remainder of the paper.

Gender Ident.	Users	Transc.	Segm.	Tokens
Trans woman	13	546	15,121	1,646,4121
Trans man	6	192	6,478	548,579
Cis woman	55	2,446	92,436	4,986,314
Cis man	79	2,474	80,902	4,309,761
Non-binary	15	514	9,397	960,224
Total	168	6,172	204,334	12,451,290

Table 1: Distribution of users, transcripts, segments and tokens across gender identities in TRANSCRIPT.

Corpus overview. The final corpus (Table 1) contains 6,172 transcripts by 168 speakers across five gender identities, for a total of 12.5 million tokens. The distribution of data is skewed towards the cisgender groups, possibly due to our sampling procedure: each target group speaker was matched to potentially multiple control group candidates, all of whom were retained. However, we reiterate that we included *all* potential target group speakers. The problem is therefore broader than a sampling issue and likely explained by the target group's demographic under-representation, further exacerbated in databases such as Wikidata (Zhang and Terveen, 2021).

3.2 Topic Analysis

Before deploying our corpus in probing experiments, we explore its content via topic analysis. On the one hand, topics mediate the relationship between language use and gender identity (Bamman et al., 2014); on the other, they may skew gender classifiers (Dayanik and Padó, 2021). Strong topical variation across gender identity groups would affect the robustness of our probing experiments.

We use the Latent Dirichlet Allocation topic model (Blei et al., 2003) and present the 15 most frequent topics in Figure 2 (see Appendix A.4 for implementation details). Nearly all topics are represented in all groups, but to different degrees. For example, trans and non-binary subcorpora include more discussions on issues surrounding the LGBTQIA+ community, such as gender identity, sexuality, and inclusion; this is in line with the community's use of YouTube to voice lived experiences (Miller, 2019). The cisgender groups have comparatively more discussions regarding general interests and personal experiences, including sports, travel, and food. Across the groups, there is a similar proportion of topics such as politics, entertainment, internet spaces, and healthcare.

[6] `github.com/yt-dlp/yt-dlp`

[7] `github.com/jdepoix/youtube-transcript-api`

[8] `github.com/pytube/pytube`

[9] `github.com/adbar/py3langid`

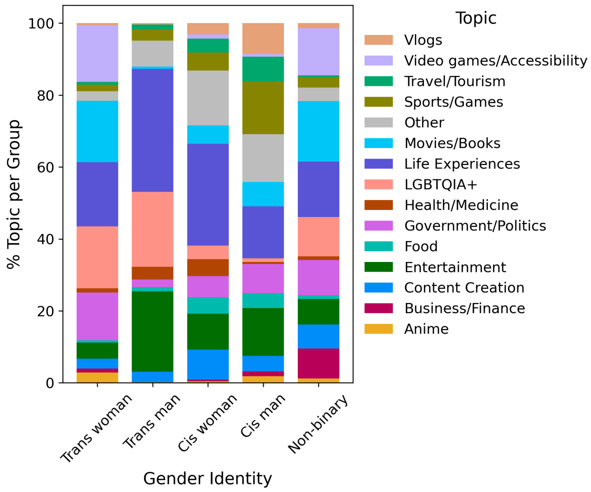


Figure 2: Top 15 topics in TRANSCRIPT and the distribution across gender identity groups.

These results indicate that our gender-specific subcorpora are overall comparable in topic; we therefore view them as a validation of our data collection pipeline. But we also take note of the variable relative importance of individual topics, which may also be due to idiosyncratic interests of individual YouTubers. We aim to account for it in our probing methodology, which we now present.

4 Probing Methodology

We replicate the probing methodology used by Lauscher et al. (2022), who investigated binary gender information encoded in PLMs. To ensure consistency, we reimplement all their experimental steps, including the probing methods and the selection of PLMs. The only variable that differs in our experiments is the dataset used for the analysis. Note more generally that our method sometimes involves direct comparisons of trans and cis gender identities. We stress that we do not imply a fundamental distinction between these groups, but rather use this perspective for a controlled assessment of disparities in model behavior.

4.1 Data

To ensure balanced representation across all classes[10], we randomly subsample each subcorpus within TRANSCRIPT to match the number of examples in the smallest one (i.e., the trans man part with 6,478 segments). The resulting dataset contains 32,390 segments across five gender identity groups. We divide it into training, validation, and test sets using an 80/10/10 split, resulting in 25,910, 3,240, and 3,240 segments, respectively. Each split maintains an equal proportion of segments from each gender identity group. We refer to this balanced subset as the **TRANSPROB** subcorpus.

4.2 Analyzed PLMs

We use PLMs from three model families: **RoBERTa** in the base and large configurations (Liu et al., 2019), **DeBERTa** base and large (He et al., 2021b), and **DeBERTaV3** xsmall, small, base, and large (He et al., 2021a). We refer to Table 7 in Appendix B.1 for sizes of these models. It is important to note that our study examines only Lauscher et al.'s (2022) subset of existing PLMs because we aim to evaluate if their findings extend to non-binary identities. Analyses of other models and architectures are reserved for future research.

4.3 Probing Methods

To ensure the robustness of our findings, we use two probing methods: traditional probing classifiers and Minimum Description Length (MDL).

Traditional probing classifier is a supervised model trained to predict a label (i.e., in our case, the gender identity of the author) from the representations of PLMs (Belinkov, 2022; Hewitt and Liang, 2019). Figure 1 (bottom panel) depicts an example of such a probe. The input dimension of the classifier matches the embedding size of PLM, and the output dimension corresponds to the number of classes in the probed task. Since the representations are frozen during training (i.e., they are not updated during back-propagation), the higher the classifier's accuracy, the more information it was able to decode from these representations.

We implement the probing classifier as a two-layer feed-forward network with ReLU activation and softmax output. We refer to Appendix B.2 for relevant hyperparameters.

MDL probing is an information-theoretic approach based on the idea that the efficiency of information encoding in the frozen representations correlates with the amount of data needed to extract this information. During training, the probing model estimates the minimum length (i.e., minimum amount of data) required to transmit the target property. Thus, the more effectively the representations encode the property, the more efficiently the probing model can compress and transmit it.

[10] Now that we switch to probing, we adopt ML terminology, using the terms "classes" or "labels" to refer to the variable extracted from PLMs - the gender identity of the authors.

	F1		MDL	
RoBERTa-large	.81	±.01	298	±2.3
DeBERTa-verLarge	.81	±.00	284	±7.1
DeBERTaV3-verLarge	.78	±.00	324	±3.2

Table 2: Pairwise probing results for frozen representations and two gender identities: cis men and cis women. Averages and standard deviation for 5 runs.

For estimating MDL, we use online coding (Voita and Titov, 2020). In general, a lower MDL score represents a higher extractability of the property from the given frozen PLM representations.

Preprocessing. We further tokenize segments from TRANSPROB using the default tokenizers provided with the models. We pad the given inputs to a consistent length of 512 tokens and extract word embeddings from the final hidden layer of the PLMs. To create a single vector representation for each input segment, we average the word embeddings, ignoring any special tokens. These averaged embeddings are then used as frozen representations to train and evaluate our two probing methods.

4.4 Evaluation

The performance of traditional probing classifiers is evaluated on a held-out test set using the F1 score. MDL probing is cross-evaluated on the training data using the online coding measurement (see Equation (2) in Appendix B.2). Performance scores are reported as the mean and standard deviation from five runs using different random initialization.

5 Gender Identity in PLMs

We now turn to the primary goal of this work – analyzing whether PLMs capture information about authors' gender identities. Before addressing our three main research questions, we replicate Lauscher et al.'s (2022) binary classification setup with our dataset. These experiments evaluate the positioning of TRANSPROB in relation to other domains analyzed by the authors and determine if our probing results are consistent with theirs.

5.1 Pairwise Classification

We perform pairwise classification, where the probes are tasked with distinguishing between only two classes. Table 2 presents a subset of results for the three best-performing models and the same two gender identities as Lauscher et al. (2022) – cis men and cis women (for results across all genders, see Table 8 in Appendix C.1). We observe that the classifiers are capable of predicting the gender identity of the authors with a robust F1 score of 0.81. This performance places TRANSPROB on par with the Lauscher et al.'s (2022) control task CoLA and their easiest dataset `fb_wiki` (Voigt et al., 2018) – Facebook posts from public figures – for which RoBERTa-large achieves 0.81 F1. Regarding MDL, our results are also the most similar to `fb_wiki`, where the three models receive between 260 to 300 MDL.[11] Most importantly, both probing methods consistently rank our three classifiers in the same order as Lauscher et al. (2022).

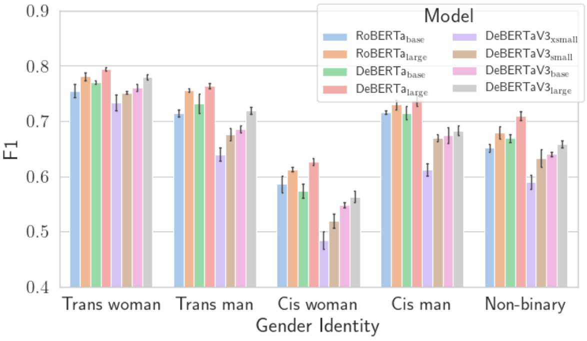


Figure 3: Results from the 5-way probing using frozen representations. Average per-class F1 scores and standard deviation for 5 runs with different random seeds.

5.2 5-way Probing of Gender Identity

Having established the validity of TRANSPROB with respect to the probing framework of Lauscher et al. (2022), we can address **RQ1** and ask if frozen PLM representations encode identities beyond the binary cis genders. To explore this, we conduct a 5-way classification, where models are trained to predict one out of the five labels from our dataset.

Figure 3 presents probing results for all gender groups. First, accuracy is lower than in Table 2 – an expected outcome given the more challenging setting of predicting one of five classes. Nevertheless, all classifiers perform well above the random baseline of 0.2 F1, and almost all exceed 0.5 F1. This result demonstrates a substantial signal about gender identities within the representations. Second, similarly to results from Lauscher et al. (2022), model size influences the encoding of sociodemographic information. Larger models, such as DeBERTa-large, exhibit higher information extractability than smaller models like DeBERTaV3-xsmall.

[11] The authors do not provide concrete numbers, so we estimate them from their plots.

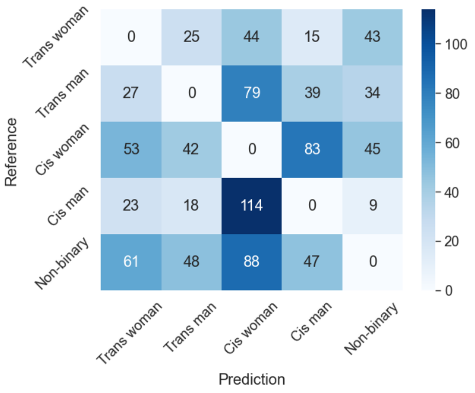


Figure 4: Number of errors (y-axis) from the 5-way probing of frozen representations from DeBERTa-large.

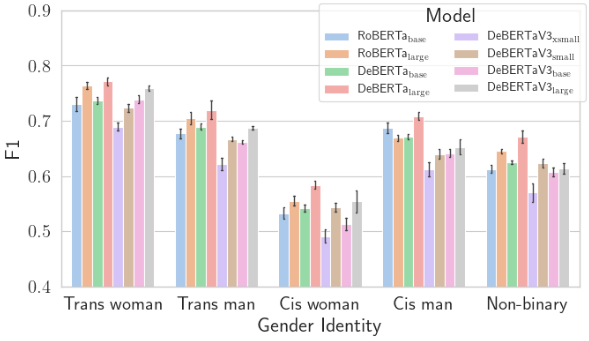


Figure 5: Results from the 5-way probing using author-controlled frozen PLM representation. Average per-class F1 scores and standard deviation from 5 runs.

Surprisingly, when comparing performance across groups, we observe notable variations in the difficulty of predicting gender identities from the representations. Detecting signals for cis women and non-binary individuals proves more challenging compared to those for trans women, trans men, and cis men. Figure 4 zooms into this result and presents the number of errors by the model with the highest probing score – DeBERTa-large. It frequently misclassifies examples from the cis woman subcorpora as cis man and vice versa. Moreover, examples from all subcorpora are predicted as belonging to the non-binary subcorpus, potentially highlighting the broad spectrum of identities under the non-binary umbrella. We expand upon these observations in a broader context in Section 6.

5.3 Controlling for Authors' Signals

As outlined in Section 3.2, our gender-specific subcorpora are comparable in topics. However, YouTube personalities often craft their online personas around specific matters or catchphrases – could these be the signals that probing methods capture? To address **RQ2**, we continue the probing experiments from the previous section, this time controlling for signals specific to authors.

Experimental setting. To identify words specific to individual YouTubers, we use the Sparse Additive Generative Model of Text (SAGE; Eisenstein et al., 2011). SAGE identifies terms that are significantly over- or underrepresented in a target dataset by comparing their frequency to the frequency in a general (i.e., background) corpus.

We apply SAGE for each author in a one-vs-all setting, where each target corpus comprises all author-specific transcripts from the full TRANSCRIPT, and the background corpus consists of all remaining texts. After calculating normalized SAGE coefficients, we select all lemmas with values exceeding 1.0. This process identifies 41,829 terms, i.e. approximately 14% of all tokens in TRANSPROB. Among these, we find phrases specific to individual YouTubers, such as *underachiever* used as an established way of addressing the audience or *teehee* used as an outro phrase for each video. However, we also find more general terms such as *Kotick* or *Activision*, which one YouTuber uses in multiple videos about gaming companies.[12] Therefore, while our method primarily aims to capture author-specific signals, it also serves as an implicit means of controlling for topic. Note moreover that individual YouTubers might follow template-like structures for their videos. While these could constitute an author-specific signal, we deliberately mitigate such potential influences by splitting the transcripts into smaller randomly subsampled segments (see Section 3.1).

Next, we mask with <MASK> all author-related tokens identified by SAGE in the corresponding author's transcripts in TRANSPROB. Afterwards, we extract the averaged vector embeddings from the selected PLMs and leave all the other preprocessing and training procedures unchanged.

Results. The results of our author-controlled 5-way probing experiments are depicted in Figure 5. Given that the patterns observed from the pairwise classifications and MDL probing are consistent with the 5-way classification, we consider them as validation experiments and provide details in appendix (see Table 9 in Appendix C.1).

Comparing Figures 3 and 5, we notice decreased probing accuracy for all models. While their rank

[12] Robert Kotick is the CEO of Activision Blizzard, one of the largest video game publishers.

remains unchanged – with DeBERTa-large achieving the highest F1 score and DeBERTaV3-xsmall the lowest – their accuracy consistently drops by approx. 0.05 F1 for all subcorpora. The impact of model size also remains strong, with larger models continuing to demonstrate greater extractability of information. However, despite lower accuracy, all models achieve F1 well above the majority baseline. This suggests that the signals remaining after author-specific information is removed are still sufficiently robust to enable the decoding of information related to those authors' gender identities. Furthermore, similarly to the previous section, we observe trends in these signals specific to the five subcorpora, with the cis woman and non-binary identities showing the lowest extractability.

5.4 Fine-tuning Representations

So far, we observed that the frozen representations of PLMs encode gender identity information sufficiently to achieve probing accuracy up to 0.7 F1. This finding prompts the question: Is this the upper limit? To explore this, we shift our focus to **RQ3** and evaluate how the capabilities of PLMs change if we fine-tune them using the same training objectives as in probing. These experiments are designed to assess the change in the models' representations when provided with explicit supervised signals as a comparison to the probing experiments.

Experimental setting. We use the same architecture as in the probing classifiers, using the averaged word embeddings as inputs to the classification heads. The main change is that the PLM representations are not frozen anymore and can learn new signals from the training data. All models are fine-tuned for 3 epochs using a batch size of 8 and a learning rate 1×10^{-5} with Adam optimizer. We use early stopping based on validation loss, with a patience of 5. During fine-tuning, the models are evaluated ten times and the best-performing model is saved. Once training is complete, we load the model with the overall best score.

Results. Figure 6 presents the results of the 5-way fine-tuning experiments (see Table 10 in Appendix C.1 for the validation pairwise results). As expected, fine-tuning improves accuracy by up to 0.1 F1 with the largest models reaching nearly 0.8 F1 for 4 out of 5 subcorpora. These results align with other gender prediction findings that, according to HaCohen-Kerner (2022), vary widely in accuracy (52% to 91%) across different domains and architectures. Additionally, fine-tuning reduces the variance between the results of different PLMs, with smaller models scoring closer to the base and larger variants. However, despite these improvements, disparities in F1 scores persist among gender identities, with the cis woman and non-binary subcorpora continuing to yield the lowest scores.

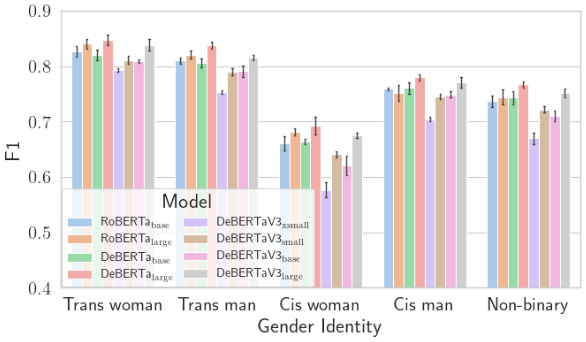


Figure 6: Results from 5-way fine-tuned classification. Average F1 scores and standard deviation from 5 runs.

6 Discussion and Conclusion

In this paper, we adopted an inclusive perspective to analyze whether gender identity is encoded in the representations of PLMs. Our study began with the creation of TRANSCRIPT, which, to the best of our knowledge, is the first dataset to include language samples from individuals across five gender identities. The proposed data collection pipeline and topic analysis contribute to ongoing research on language use in trans and queer communities.

Do representations encode gender identity? With two distinct probing methods, we found a substantial amount of gender identity information encoded within PLM representations. The information extends beyond binary gender and is extractable from frozen representations even after masking author-specific words, which may benefit the probes. Moreover, model size is one of the strongest factors indicating final probing accuracy.

Are there gender biases in representations? We consistently observed error discrepancies in model behaviors (refer to Shah et al. (2020) for an overview of types of biases in NLP models). These disparities persisted even when controlling for author-specific words or fine-tuning with a direct signal. Figure 7 presents an overview on the disparities across all three experiments from the highest scoring model DeBERTa-large. They were particularly visible in model behavior for cis

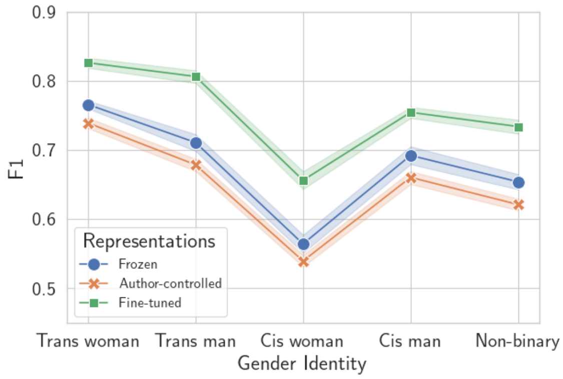


Figure 7: Overview of the 5-way classification results from all three experiments on DeBERTa-large.

women and non-binary individuals. A detailed sociolinguistic analysis of why this is the case (e.g., stronger expressivity of some individuals or identity-establishing language variation) represents important future work. Analogously, a relevant direction is the use of interpretability methods such as SHAP (Lundberg and Lee, 2017) to identify individual tokens that lead to such model behaviours. However, we can already conclude that even in balanced data the ease of extracting gender identity from representations differs across the analyzed populations.

The observed biases in representations can pose significant challenges in downstream applications. For example, applications like YouTube's automatic captions may have disparate performance for different gender identities due to biased encodings (Tatman, 2017). Perhaps more critically, some identity verification systems incorrectly interpret non-binary identities as fake or non-human (Dev et al., 2021). Similarly, tools predicting users' gender on social media platforms can misgender individuals from already marginalized groups, negatively impacting their self-esteem (Fosch-Villaronga et al., 2021). Future work should directly examine how the patterns identified in our findings affect such downstream tasks. Current debiasing methods are incapable of fully removing authorship signals from PLMs (Lalor et al., 2022), so it is all the more critical to identify the tasks which might use information related to gender identities and the populations which might be harmed by them. Placed within this context, our findings underscore the vital importance of including transgender and non-binary people in the development and assessment of language technologies.

7 Limitations

This paper provides two main contributions: a pipeline for creating TRANSCRIPT and the findings from our probing analysis. Regarding the pipeline, there are several points to consider. Firstly, the pipeline strongly relies on the representation of diverse gender identities on Wikidata. Given that the trans and queer community is already demographically underrepresented, publicly available databases such as Wikidata exacerbate this underrepresentation (Zhang and Terveen, 2021). Consequently, our approach to population sampling reflects and potentially perpetuates this issue.

Secondly, the sampling method for cisgender individuals, which relies on the same nationality, similar age, and level of YouTube fame, can introduce a selection bias in our corpus. Restricting the country of citizenship allows us to focus on English-speaking YouTubers, but including both American and British English speakers may introduce minor linguistic variations. Finally, we perform an automatic diarization step to filter segments spoken by the target speaker. While we carefully design the methodology, it might still misidentify some of the speakers in collaborative videos, leading to noise in our data.

When it comes to the probing experiments, it is important to emphasize that our results are limited to the domain of TRANSCRIPT: YouTube transcripts. As such, the analyzed segments are samples of spoken, but possibly scripted, language. Therefore, our result do not directly generalize to other domains. Secondly, while we account for author-specific words and implicitly topic-related signals, our probing experiments do not control for other latent aspects that could influence our results.

8 Ethical Considerations

Our research involves collecting and predicting sensitive personal information, specifically gender identity. We acknowledge the potential for dual use of our work, and specifically the fact that predictive methods may be used for harmful applications, such as author profiling, which may perpetuate biases and discrimination. However, we believe that this risk is outweighed by obtaining findings that promote the inclusion of LGBTQIA+ individuals and assess the risks that NLP systems may pose to already marginalized communities. We strive to acknowledge associated risks and to account for the sensitive nature of research related to the

LGBTQIA+ community, thereby fostering more inclusive and equitable NLP technologies.

We developed TRANsCRIPT by collecting material available on YouTube in accordance with YouTube's Terms of Service (YouTube, 2024), which allow for automated download of content "as permitted by applicable law". Article 3 of the EU directive on copyright and related rights in the Digital Single Market (Directive 2019/790) sets out an exception for collecting and analyzing otherwise copyright-protected material within the scope of text and data mining for scientific research. The work presented here falls under such a scope since it was conducted as a non-commercial project in a public research institution in the European Union.

To regulate the possible dual-use issues associated with the data, TRANsCRIPT will be made available only upon direct request. Consequently, we are releasing only the code needed to reproduce our dataset and analysis – an approach that aligns with previous work on YouTube (Ko et al., 2023, among others). Moreover, we store the collected data securely and do not directly disclose the authors' identities. We note however that the source data remains publicly accessible by virtue of the authors' presence on YouTube.

The information we collect does not include self-reported identity labels. Instead, we use a collaborative database, Wikidata, which contains information about public figures, including their gender identity labels. This approach has limitations, as gender identities can overlap and change over time. Relying on externally assigned labels may not always accurately represent an individual's self-identified gender.

Finally, our findings pertain specifically to the corpus and models used in this study and should not be generalized to broader populations. The results do not represent the entire spectrum of gender identities or sociodemographic groups and communities. We recognize the complexity and diversity of gender identities and acknowledge that our work captures only a subset of this spectrum within the context of our resources.

9 Acknowledgements

Agnieszka Falenska was supported by the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9-19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. Filip Miletić was supported by DFG Research Grant SCHU 2580/5-1.

References

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Joan-Isaac Biel and Daniel Gatica-Perez. 2010. Vlogcast yourself: Nonverbal behavior and attention in social media. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pages 1–4, New York, NY, USA. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *INTERSPEECH 2023*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661.

Steven Coats. 2023. Dialect Corpora from YouTube. In *Dialect Corpora from YouTube*, pages 79–102. De Gruyter.

Sudeshna Das and Jiaul H Paik. 2021. Context-sensitive gender inference of named entities in text. *Information Processing & Management*, 58(1):102423.

Brittany I. Davidson, Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen, and Alicia G. Cork. 2023. Platform-controlled social media apis threaten open science. *Nature Human Behaviour*, 7(12):2054–2057.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling Document Topic and Author Gender in Multiple Languages: Lessons for Adversarial Debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online). Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

Directive 2019/790. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. `https://eur-lex.europa.eu/eli/dir/2019/790/oj`.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*, 21:461–490.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1041–1048, Madison, WI, USA. Omnipress.

Eduard Fosch-Villaronga, Adam Poulsen, Roger Andre Søraa, and BHM Custers. 2021. A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, 58(3):102541.

Yaakov HaCohen-Kerner. 2022. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, 199:117140.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and "diagnostic classifiers" reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can language models laugh at YouTube short-form videos? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2897–2916, Singapore. Association for Computational Linguistics.

Egor Lakomkin, Sven Magg, Cornelius Weber, and Stefan Wermter. 2018. KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition from YouTube Videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 90–95, Brussels, Belgium. Association for Computational Linguistics.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, When, and Where Language Models Learn about Sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(Volume 7, 2021):195–212.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? Assessing BERT as a distributional semantics model. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Jordan F. Miller. 2019. Youtube as a site of counternarratives to transnormativity. *Journal of Homosexuality*, 66(6):815–837.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How Gender Debiasing Affects Internal Model Representations, and Why It Matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, pages 3222–3226.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender:

A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

YouTube. 2024. Terms of service. https://www.youtube.com/static?template=terms. (accessed on 15 July 2024).

Charles Chuankai Zhang and Loren Terveen. 2021. Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In *Proceedings of the 17th International Symposium on Open Collaboration*, OpenSym '21, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019(256):147–175.

Lal Zimman. 2020. Transgender Language, Transgender Moment: Toward a Trans Linguistics. In Kira Hall and Rusty Barrett, editors, *The Oxford Handbook of Language and Sexuality*. Oxford University Press.

Lal Zimman and Will Hayworth. 2020a. How we got here: Short-scale change in identity labels for trans, cis, and non-binary people in the 2000s. *Proceedings of the Linguistic Society of America*, 5(1):499–513.

Lal Zimman and Will Hayworth. 2020b. Lexical change as sociopolitical change in talk about trans and cis identity labels: New methods for the corpus analysis of internet data. In *Selected Papers from NWAV47*, pages 143–152.

Property		Value	
PID	Label	QID	Label
P31	instance of	Q5	human
P27	country of citizenship	Q30	United States of America
		Q145	United Kingdom
P21	sex or gender	Q1052281	trans woman
		Q2449503	trans man
		Q6581072	female
		Q6581097	male
		Q48270	non-binary
P569	date of birth		
P2397	YouTube channel ID		
P3744	number of subscribers		
P5436	number of viewers/listeners		

Table 3: Wikidata property (PID) and value (QID) codes, accompanied by their corresponding labels.

A Corpus

A.1 Wikidata Information

Wikidata codes. The properties and values used in the construction of TRANSCRIPT, along with their corresponding Wikidata codes, are listed in Table 3.

Distribution of gender identities. As outlined in Table 4, there is a significant discrepancy of gender identities among YouTubers on Wikidata, where less than 0.03% of the individuals on Wikidata represent queer identities. The underrepresentation of gender identities is reflected in the skewed distribution in TRANSCRIPT towards the control group.

A.2 Speaker Filtering

YouTube transcripts can include segments spoken by different speakers. Moreover, commentary videos can feature clips of other speakers followed by a discussion of their statements. Therefore, as a preprocessing step, we perform speaker filtering (Figure 8) using speaker diarization to identify the person we sample from Wikidata (target speaker) and remove segments in the transcripts produced by other speakers (background speakers).

Speaker diarization is an unsupervised process that identifies and labels specific segments of audio or video recordings with a speaker identity la-

Wikidata QID	Wikidata Label	Total	YouTubers	# Youtubers
Q6581097	Male	60.9%	0.21%	23,093
Q6581072	Female	21.1%	0.12%	12,861
Q1052281	Trans woman	0.02%	$1 \times 10^{-3}\%$	161
Q48270	Non-binary	$1 \times 10^{-2}\%$	$1 \times 10^{-3}\%$	140
Q2449503	Trans man	$4 \times 10^{-3}\%$	$2 \times 10^{-4}\%$	27
Q18116794	genderfluid	$9 \times 10^{-4}\%$	$2 \times 10^{-4}\%$	21
Q12964198	Genderqueer	$9 \times 10^{-4}\%$	$1 \times 10^{-4}\%$	13
Q189125	Transgender	$7 \times 10^{-4}\%$	$8 \times 10^{-5}\%$	9

Table 4: Top 8 frequently occurring *sex or gender* (P21) values on Wikidata that are *human* (Q5). We display the overall percentage of occurrence along with the overall percentage and actual number of individuals with a *YouTube channel ID* (P2397).

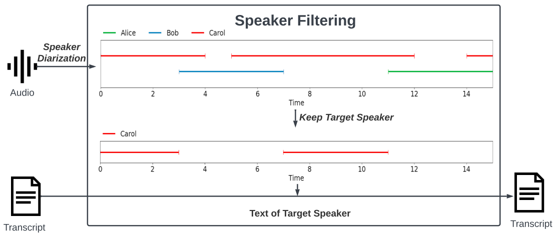


Figure 8: Representation of transcript preprocessing using speaker filtering.

bel. In other words, it answers the question 'who spoke when' (Park et al., 2022). We use an off-the-shelf speaker diarization pipeline provided by the pyannotate.audio library[13] (Bredin, 2023) from the Huggingface Hub.[14] The pipeline contains a pretrained end-to-end neural speaker diarization model (Plaquet and Bredin, 2023), reporting significant improvements in overlapping speech and robust cross-domain performance.

The steps for performing speaker filtering on the user-uploaded transcripts are the following:

1. Downloading the YouTube video's audio file using the yt-dlp library and reducing the file's quality for faster preprocessing. The frame rate is decreased to 22050 Hz, the sample width is set to 2 bytes, and the audio channel is converted to mono.

2. Applying the speaker diarization pipeline to identify and obtain a group of speaker segments.

3. If multiple speakers are found, perform a *speaker disambiguation task* to identify the target speaker (Figure 9; details below).

4. Trimming target speaker segments that intersect with background speaker's segments.

5. Removing transcript segments that do not overlap with the target speaker's segments based on start time and duration. When no speaker is identified, or the set of speaker segments is empty, we discard the entire transcript.

Since no prior audio samples are obtained from the target speakers and due to the unsupervised nature of speaker diarization, we incorporate an additional step to identify the target speaker. We frame this as a *speaker disambiguation task* (Figure 9), determining the target speakers from a group of background speakers identified in the audio file. We propose an algorithm that leverages a collection of videos from the same YouTube channel as the video being processed to identify the target speaker. The algorithm initially compiles a list of videos from the channel sorted in descending view count and then by duration (rounded to the nearest 1000 views and 15 minutes, respectively), excluding music videos and movie trailers. The assumption is that longer, more popular videos are more likely to feature the YouTube channel owner who is being identified. It then iterates over the sorted list of videos, each time concatenating a quality-reduced audio file with the audio from the previously processed videos and performing speaker diarization. The iteration steps continue until a speaker representing 70% of the speaking time is identified, indicating our target speaker.

Speaker filtering offers advantages and limitations. This step filters out non-target speakers and background noises like musical cues or annotations not transcribing spoken words, such as silent videos with a story narrated using user-uploaded transcripts. A limitation to consider is the iden-

[13] github.com/pyannote/pyannote-audio
[14] huggingface.co/pyannote/speaker-diarization-3.1

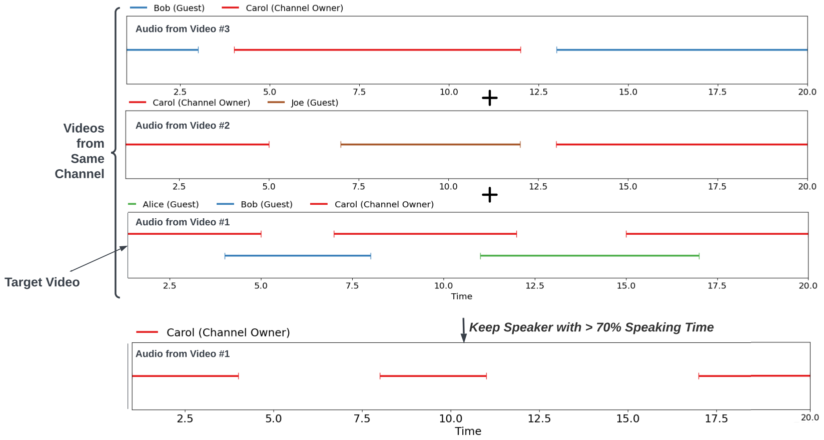


Figure 9: Illustration of the speaker disambiguation task employed in the speaker filtering preprocessing step.

tification of the target speaker. Considering the curation of YouTube videos from a channel, the algorithm may identify the target speaker as a different commonly occurring speaker that is not the actual candidate YouTuber from the target or control group.

A.3 Corpus Statistics

Reduction in corpus size. To assess the impact of the corpus pipeline on the final size of TRANSCRIPT, we provide statistics (Table 5) on the number of users, transcripts, and tokens retained during each stage of the corpus creation pipeline. Initially, sampling from Wikidata resulted in $9,900$ potential YouTubers. After collecting YouTube transcripts, the number of users decreased by 96.9% to 306 potential candidates, with $7,999$ transcripts with a total of 18.9 million tokens. The collection from YouTube significantly reduces the number of authors, since the focus is on authors who include user-uploaded transcripts in their YouTube videos. During preprocessing (Section 3.1), the discarding of music videos and movie trailers, and the application of speaker filtering, led to a 28% token reduction. The final TRANSCRIPT (Table 1) contains 66% of the initially collected data from YouTube.

A.4 Topic Modelling Details

Data and preprocessing. Using the TRANSCRIPT corpus containing $6,172$ transcripts from 168 individuals, we represent each transcript as a continuous string of text.

Similarly to Devinney et al.'s (2020) subset of preprocessing steps, we tokenize, lemmatize, and POS tag each document with spaCy (Honnibal and Montani, 2017) using the en_core_web_md model. We remove all punctuations, non-alphabetical words, two-character words, and words included in spaCy's list of English stopwords. Using the POS information from the tagger, we keep only nouns, verbs, adjectives, and adverbs. We maintain only the lowercase lemmatised list representations of each document and exclude all high-frequency terms that appear in more than 80% documents. We test various thresholds and empirically determined that excluding the top 20% of words achieves the best results. Subsequently, we then use the gensim library[15] to convert the documents into a bag-of-words (BoW) format.

Implementation details. We use the LDA implementation available in the gensim library. We set the number of topics to $k = 15$ and the chunk size to 2500, the number of iterations to 500, and the number of passes to 50. All other hyperparameter settings are kept to their default values. We do not perform any evaluation steps during the model's training.

For a more interpretable understanding of a topic's meaning, we measure the *relevance* of a term to a topic (Sievert and Shirley, 2014). The relevance of a term w described in Equation 1 is measured by calculating a weighted average of the term's probability within a topic t with the ratio of a term's probability within a topic to its marginal probability across the corpus. We set the weight parameter value to be $\lambda = 0.6$.

[15] radimrehurek.com/gensim/

	Users	Transcripts	Tokens (T)	% T Kept
Wikidata Collection	9,874	N/A	N/A	N/A
YouTube Collection	306	7,999	18,877,552	100%
Filter Music and Movie Trailers	176	6,359	16,958,559	90%
Speaker Filtering	172	6,246	13,599,334	72%
Language Filtering	170	6,195	13,563,152	72%
Rule-based Text Filtering	168	6,172	12,451,290	66%
TRANSCRIPT	168	6,172	12,451,290	66%

Table 5: Reduction of users, transcripts, and tokens (T) at each step of the corpus creation pipeline, from the initial Wikidata and YouTube collection to the final TRANSCRIPT corpus.

$$relevance(w|t) = \lambda * p(w|t) + (1-\lambda) * \frac{p(w|t)}{p(w)} \quad (1)$$

Relevant Words per Topic. In Table 6 we provide the 30 most relevant terms for each topic of the 15 topics obtained from our topic modelling analysis.

B Representation Probing

B.1 Analyzed PLMs

For all experiments in the paper, we use the models described in Table 7.

B.2 Probing Methods

Traditional probing classifiers. The input dimension is set to match the PLM's output dimension, which varies based on the model size: 384 for xsmall, 512 for small, 768 for the base, and 1024 for large models. The hidden dimension is set to 100, and the output dimension is either 2 or 5, depending on the classification setup. Using PyTorch default parameters, we train the probing classifiers using a batch size 32 and a learning rate of 1×10^{-3} with the Adam optimiser (Kingma and Ba, 2015). Early stopping is applied based on validation loss, with a patience of 5 epochs. The learning rate is halved if the validation loss does not improve for a single epoch. We follow the same approach when implementing MDL, using the same network architecture and training procedure as described for the probing classifiers.

MLD probing. We use the *online coding* method to estimate MDL. First, the dataset $D = (x_i, y_i)_{i=1}^{N}$ is divided into S time steps $1 = t_0 < t_1 < ... < t_S < N$.[16] Then, a traditional probing classifier is trained on the samples $(1, ..., t_i)$ and used to predict the output label for the next $(t_i+1, ..., t_{i+1})$. This process continues until the entire dataset is processed. MDL is calculated as the sum of the cross-entropy loss of the classifier over the data for each time step and the uniform encoding of the first block:

$$\begin{aligned} \text{MDL} = L^{\text{online}}(y_{1:n}|x_{1:n}) &= t_1 \log_2 K \\ &- \sum_{i=1}^{S-1} \log_2 p_{\theta_i}(y_{t_i+1:t_{i+1}}|x_{t_i+1:t_{i+1}}) \end{aligned} \quad (2)$$

C Encoding of Gender Identity Information

C.1 Results

C.1.1 Pair-wise Frozen Representations

The pairwise classification results on frozen representations across all models and labels are presented in Table 8.

C.1.2 Pair-wise Author-controlled Frozen Representations

The pairwise classification results on author-controlled frozen representations across all models and labels are presented in Table 9.

C.1.3 Pair-wise Fine-tuning

The pairwise classification results on fine-tuned models across all labels are presented in Table 10.

C.2 Error Analysis

The number and types of errors across all three experimental settings are shown in Figure 10 for

[16] Following Voita and Titov (2020), we use time steps that correspond to 0.1%, 0.2%, 0.4%, 0.8%, 1.6%, 3.2%, 6.25%, 12.5%, 25%, 50%, and 100% of the dataset.

Topic Label	Relevant Terms
Food	*eat, paint, squishy, cake, ice, taste, cream, cook, pizza, little, food, cheese, chocolate, egg, flavor, chicken, butter, banana, bread, add, milk, sugar, good, cider, oil, chip, strawberry, sauce, sandwich, candy*
Sports/Games	*right, win, dude, play, man, ball, guy, run, wait, throw, team, good, kill, cut, time, let, game, come, point, hit, mean, jump, try, beat, bad, shoot, catch, second, die, kid*
Other	*little, look, good, eye, nerf, house, let, light, like, close, brush, right, bit, need, use, sleep, actually, plant, want, nice, perfect, ahead, room, face, color, asmr, ear, come, sure, open*
Life Experiences	*feel, thing, time, want, people, day, life, love, kind, like, lot, year, wear, friend, tell, hair, actually, work, good, talk, start, ask, body, try, find, way, look, come, happen, little*
Anime	*anime, season, character, episode, series, slime, show, genre, fan, animation, watch, robot, fantasy, adaptation, fight, franchise, ending, animate, japanese, world, new, villain, hype, girl, sonic, superhero, scene, titan, foul, time*
Business/Finance	*fucking, money, company, fuck, dollar, pay, product, shit, sell, buy, industry, price, cash, sale, market, business, corporation, cost, executive, spend, bullshit, service, worth, year, rich, economy, literally, profit, nft, investment*
Content Creation	*video, description, link, click, sound, language, channel, word, use, deaf, sign, caption, book, app, hear, content, number, read, different, learn, creator, hearing, sponsor, write, month, page, watch, website, voice, today*
Government/Politics	*people, country, government, law, state, work, year, system, community, business, vote, support, need, say, customer, help, disabled, public, world, vaccine, political, military, bill, job, legal, war, report, power, claim, police*
Health/Medicine	*doctor, pain, baby, pregnancy, symptom, pregnant, medical, risk, birth, patient, period, blood, hospital, abortion, diagnosis, medication, pill, chronic, disorder, diagnose, body, test, cause, illness, surgery, joint, health, uterus, infection, cycle*
Video games/Accessibility	*game, player, accessibility, play, setting, controller, feature, developer, release, gamer, mode, option, accessible, videogame, gameplay, console, disabled, support, gaming, button, offer, microtransaction, level, loot, design, control, audio, menu, launch, ability*
Entertainment	*guy, video, song, music, love, thank, want, like, laugh, girl, say, watch, look, right, literally, funny, favorite, kid, send, mean, draw, sorry, post, picture, good, bitch, cool, tweet, hope, channel*
Travel/Tourism	*look, nice, wee, guy, hotel, right, good, thank, kinda, food, stuff, eat, place, maybe, dunno, walk, area, come, street, way, bit, beach, drink, shop, brother, bloody, local, try, big, man*
Vlogs	*let, sharer, awesome, ready, look, car, crazy, check, cool, right, share, grace, epic, gun, pond, balloon, super, thing, box, come, vlog, comment, work, way, water, giant, big, button, wait, goodness*
Movies/Books	*movie, film, character, story, book, musical, world, audience, end, death, version, human, way, author, history, scene, write, bad, original, die, thing, evil, plot, narrative, culture, alien, kill, time, kind, novel*
LGBTQIA+	*woman, people, gender, man, gay, sex, person, tran, queer, lesbian, lot, relationship, trans, transgender, conversation, say, identity, white, sexual, black, thing, feel, talk, pronoun, male, binary, understand, identify, right, way*

Table 6: Top 30 relevant terms for each of the corresponding 15 topics.

DeBERTa-large, which achieved the highest probing scores.

Model	Layers	Attention Heads	Hidden size	URL https://huggingface.co
roberta-base	12	12	768	/FacebookAI/roberta-base
roberta-large	24	16	1024	/FacebookAI/roberta-large
deberta-base	12	12	768	/microsoft/deberta-base
deberta-large	24	16	1024	/microsoft/deberta-large
deberta-v3-xsmall	12	6	384	/microsoft/deberta-v3-xsmall
deberta-v3-small	6	12	768	/microsoft/deberta-v3-small
deberta-v3-base	12	12	768	/microsoft/deberta-v3-base
deberta-v3-large	24	16	1024	/microsoft/deberta-v3-large

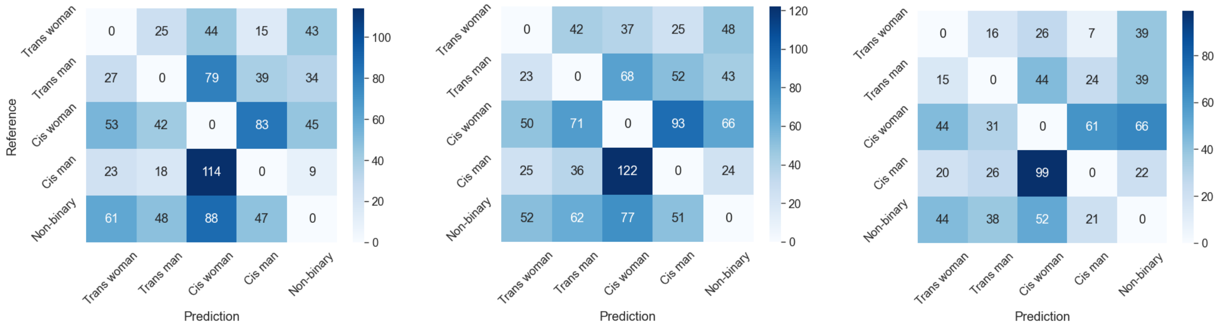
Table 7: PLMs utilised in our experiments, with architecture details and Huggingface model repository links.

		RoBERTa				DeBERTa				DeBERTa V3							
		Base		Large		Base		Large		XSmall		Small		Base		Large	
		F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL
Trans woman	Trans woman	-		-		-		-		-		-		-		-	
	Trans man	.90	171	.92	161	.92	171	.92	144	.88	212	.89	184	.90	178	.91	167
	Cis woman	.88	158	.89	143	.88	144	.89	130	.85	190	.87	164	.86	163	.88	152
	Cis man	.92	201	.93	190	.93	188	.94	172	.91	238	.92	203	.93	207	.93	197
	Non-binary	.86	222	.87	212	.87	211	.88	188	.83	276	.85	241	.84	241	.87	228
Trans man	Trans woman	.90	171	.92	161	.92	171	.92	144	.88	212	.89	184	.90	178	.91	167
	Trans man	-		-		-		-		-		-		-		-	
	Cis woman	.84	199	.85	189	.85	194	.87	169	.77	272	.82	223	.81	228	.83	214
	Cis man	.89	256	.90	244	.90	253	.91	227	.82	309	.86	269	.87	281	.86	260
	Non-binary	.82	238	.86	224	.83	235	.87	216	.79	289	.82	254	.82	255	.84	231
Cis woman	Trans woman	.88	158	.89	143	.88	144	.89	130	.85	190	.87	164	.86	163	.88	152
	Trans man	.84	199	.85	189	.85	194	.87	169	.77	272	.82	223	.81	228	.83	214
	Cis woman	-		-		-		-		-		-		-		-	
	Cis man	.81	308	.81	298	.80	297	.81	283	.74	356	.77	321	.77	323	.78	324
	Non-binary	.80	202	.81	198	.81	200	.83	180	.75	276	.79	234	.79	235	.79	226
Cis man	Trans woman	.92	201	.93	190	.93	188	.94	172	.91	238	.92	203	.93	207	.93	197
	Trans man	.89	256	.90	244	.90	253	.91	227	.82	309	.86	269	.87	281	.86	260
	Cis woman	.81	308	.81	298	.80	297	.81	283	.74	356	.77	321	.77	323	.78	324
	Cis man	-		-		-		-		-		-		-		-	
	Non-binary	.87	269	.88	260	.88	269	.88	251	.83	331	.86	300	.86	299	.87	285
Non-binary	Trans woman	.86	222	.87	212	.87	211	.88	188	.83	276	.85	241	.84	241	.87	228
	Trans man	.82	238	.86	224	.83	235	.87	216	.79	289	.82	254	.82	255	.84	231
	Cis woman	.80	202	.81	198	.81	200	.83	180	.75	276	.79	234	.79	235	.79	226
	Cis man	.87	269	.88	260	.88	269	.88	251	.83	331	.86	300	.86	299	.87	285
	Non-binary	-		-		-		-		-		-		-		-	

Table 8: Results from pairwise probing of gender identities using frozen representations from our set of PLMs. We report F1 and MDL scores of the probing classifier and MDL probe. Rows marked with '-' indicate unary classifications and are thus omitted.

		RoBERTa				DeBERTa				DeBERTa V3							
		Base		Large		Base		Large		XSmall		Small		Base		Large	
		F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL	F_1	MDL
Trans woman	Trans woman	-		-		-		-		-		-		-		-	
	Trans man	.81	195	.83	189	.83	196	.85	166	.80	234	.81	209	.84	199	.84	193
	Cis woman	.87	221	.86	218	.87	218	.88	203	.83	249	.85	228	.85	231	.86	217
	Cis man	.90	170	.90	169	.91	166	.92	156	.87	213	.89	187	.90	187	.90	175
	Non-binary	.77	252	.77	235	.78	244	.79	215	.71	303	.75	273	.75	270	.77	263
Trans man	Trans woman	.81	195	.83	189	.83	196	.85	166	.80	234	.81	209	.84	199	.84	193
	Trans man	-		-		-		-		-		-		-		-	
	Cis woman	.81	287	.81	286	.81	281	.82	256	.78	325	.79	310	.79	303	.81	289
	Cis man	.86	214	.87	212	.87	216	.88	192	.83	280	.84	250	.86	243	.86	237
	Non-binary	.80	260	.81	261	.80	263	.83	240	.76	301	.80	275	.79	280	.80	260
Cis woman	Trans woman	.87	221	.86	218	.87	218	.88	203	.83	249	.85	228	.85	231	.86	217
	Trans man	.81	287	.81	286	.81	281	.82	256	.78	325	.79	310	.79	303	.81	289
	Cis woman	-		-		-		-		-		-		-		-	
	Cis man	.74	332	.76	321	.76	324	.77	309	.71	366	.74	334	.73	339	.77	336
	Non-binary	.80	296	.80	287	.82	293	.83	270	.78	338	.81	313	.80	315	.80	312
Cis man	Trans woman	.90	170	.90	169	.91	166	.92	156	.87	213	.89	187	.90	187	.90	175
	Trans man	.86	214	.87	212	.87	216	.88	192	.83	280	.84	250	.86	243	.86	237
	Cis woman	.74	332	.76	321	.76	324	.77	309	.71	366	.74	334	.73	339	.77	336
	Cis man	-		-		-		-		-		-		-		-	
	Non-binary	.86	234	.85	230	.88	225	.89	209	.82	288	.85	252	.85	261	.86	242
Non-binary	Trans woman	.77	252	.77	235	.78	244	.79	215	.71	303	.75	273	.75	270	.77	263
	Trans man	.80	260	.81	261	.80	263	.83	240	.76	301	.80	275	.79	280	.80	260
	Cis woman	.80	296	.80	287	.82	293	.83	270	.78	338	.81	313	.80	315	.80	312
	Cis man	.86	234	.85	230	.88	225	.89	209	.82	288	.85	252	.85	261	.86	242
	Non-binary	-		-		-		-		-		-		-		-	

Table 9: Results from pairwise probing of gender identities using topic-controlled frozen representations from our set of PLMs. We report F1 and MDL scores of the probing classifier and MDL probe. Rows marked with '-' indicate unary classifications and are thus omitted.



(a) Frozen representations (b) Author-controlled representations (c) Fine-tuned representations

Figure 10: Number of errors (y-axis) from the 5-way probing of representations from DeBERTa-large.

		RoBERTa		DeBERTa		DeBERTa V3			
		Base	Large	Base	Large	XSmall	Small	Base	Large
		F$_1$	F$_1$	F$_1$	F$_1$	F$_1$	F$_1$	F$_1$	F$_1$
Trans woman	Trans woman	-	-	-	-	-	-	-	-
	Trans man	.89	.91	.90	.89	.81	.89	.90	.88
	Cis woman	.84	.91	.86	.90	.85	.85	.84	.91
	Cis man	.91	.95	.91	.95	.92	.91	.92	.95
	Non-binary	.85	.91	.84	.91	.84	.82	.83	.90
Trans man	Trans woman	.89	.91	.90	.89	.81	.89	.90	.88
	Trans man	-	-	-	-	-	-	-	-
	Cis woman	.85	.90	.86	.90	.81	.83	.84	.90
	Cis man	.90	.92	.91	.93	.90	.89	.90	.92
	Non-binary	.88	.89	.87	.91	.86	.87	.88	.90
Cis woman	Trans woman	.84	.91	.86	.90	.85	.85	.84	.91
	Trans man	.85	.90	.86	.90	.81	.83	.84	.90
	Cis woman	-	-	-	-	-	-	-	-
	Cis man	.79	.84	.77	.83	.76	.76	.77	.85
	Non-binary	.79	.86	.80	.87	.77	.78	.78	.86
Cis man	Trans woman	.91	.95	.91	.95	.92	.91	.92	.95
	Trans man	.90	.92	.91	.93	.90	.89	.90	.92
	Cis woman	.79	.84	.77	.83	.76	.76	.77	.85
	Cis man	-	-	-	-	-	-	-	-
	Non-binary	.87	.91	.86	.92	.86	.87	.88	.92
Non-binary	Trans woman	.85	.91	.84	.91	.84	.82	.83	.90
	Trans man	.88	.89	.87	.91	.86	.87	.88	.90
	Cis woman	.79	.86	.80	.87	.77	.78	.78	.86
	Cis man	.87	.91	.86	.92	.86	.87	.88	.92
	Non-binary	-	-	-	-	-	-	-	-

Table 10: F1 results across our set of fine-tuned PLMs on pairwise prediction of gender identities in TRANSPROB. We include redundant rows of inverted class pairings for easier intra-class comparisons. Rows marked with '-' indicate unary classifications and are thus omitted.