The Echoes of the 'I': Tracing Identity with Demographically Enhanced Word Embeddings

Ivan Smirnov
University of Technology Sydney
ivan.smirnov@uts.edu.au

Abstract

Identity is one of the most commonly studied constructs in social science. However, despite extensive theoretical work on identity, there remains a need for additional empirical data to validate and refine existing theories. This paper introduces a novel approach to studying identity by enhancing word embeddings with socio-demographic information. As a proof of concept, we demonstrate that our approach successfully reproduces and extends established findings regarding gendered self-views. Our methodology can be applied in a wide variety of settings, allowing researchers to tap into a vast pool of naturally occurring data, such as social media posts. Unlike similar methods already introduced in computer science, our approach allows for the study of differences between social groups. This could be particularly appealing to social scientists and may encourage the faster adoption of computational methods in the field.

1 Introduction

Identity is central and one of the most commonly studied constructs in the social sciences, shaping our understanding human behaviour, and society more generally (Leary and Tangney, 2003). While there is no universally accepted definition of identity, it generally refers to individual's self-perception that consists of self-ascribed personal traits, beliefs about themselves, as well as self-categorization into particular social groups and roles.

Research on identity spans disciplines from psychology to sociology, and from linguistics to political science offering rich theoretical insights into identity (Vignoles et al., 2011). However, measuring identity and related constructs remains challenging, which is why there is still a clear need for empirical studies that would allow to validate and refine existing theories (McLean and Syed, 2015). Established methods typically require the annotation of survey data by experts who have to be specially trained. Take, for instance, Loevinger theory of ego development (Loevinger, 1976) which is generally considered as one of the most empirically supported theories of personality development (Gilmore and Durkin, 2001). Traditionally, ego development is measured via the Washington University Sentence Completion Test (WUSCT) (Hy and Loevinger, 1996). That is a projective technique where participants are asked to complete sentence stems such as "What gets me in trouble..." or "A girl has a right to...". While WUSCT has been shown to be a reliable and valid method of measuring ego development (Gilmore and Durkin, 2001), its administration is resource-intensive and requires a specialized training for raters. At the same time, recent developments in computational methods suggest that social media data, at least at a macro level, can aid in assessing psychological constructs (Pellert et al., 2022). This could pave the way for alternatives to traditional survey-based assessments.

Computational approaches and natural language processing have been previously applied to study identity. In particular LIWC (Tausczik and Pennebaker, 2010) – a popular dictionary-based method – has been used to analyze responses to WUSCT (Lanning et al., 2018) or to identify salient identity in social media posts (Koschate et al., 2021). In our work, we propose using word embeddings as they allow capturing more complex semantic relationships in the text by considering the context in which words are used.

The common approach to using word embeddings in social science is to consider projections on semantic axes in word-vector space. It has been previously demonstrated that this technique could effectively recover human sentiments, judgments, and perceptions (An et al., 2018; Grand et al., 2022). This enabled computational social

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 112–118
July 5–10, 2020.

scientists to extract insights from large text corpora. The potential of this approach was most notably demonstrated in studies on stereotypes (Caliskan et al., 2017; Garg et al., 2018; Boutyline et al., 2023).

Typically, a word embedding model is trained on a specific corpus of interest. Then, the distance between target words and predefined reference poles, represented by opposing words or sets of words, is considered. This distance is interpreted as the semantic closeness between target words and reference poles, providing insights into underlying associations and relationships. More concretely, it has been shown that certain occupational terms, e.g. 'mechanic', are closer to words representing men ('man', 'boy', 'he', etc.), while other terms, e.g. 'nurse', are closer to words representing women ('woman', 'girl', 'she', etc.), indicating a gender bias (Garg et al., 2018). By training separate word embedding models on time-segmented historical texts, it has been further demonstrated that the changes in word distances over time reflect real-world changes in women's occupations.

Another study has found that words representing men are closer to words related to intelligence, while words representing women are closer to 'studying', reflecting a common stereotype in education: "boys are successful at school because they are smart and girls because they study a lot" (Boutyline et al., 2023). Training separate word embedding models on texts produced at different time points further showed that this stereotype emerged at specific point in time, consistent with sociological explanations of the phenomenon.

Simply computing word similarities in a given corpus is often not very informative. Therefore, researchers typically segment the corpus for comparative analysis. These segments might represent different time periods, as in the examples above, or the corpus could be split by other criteria, such as training distinct models on texts authored by Republicans versus Democrats (Rodriguez and Spirling, 2022). This approach, however, has a disadvantage as it reduces the amount of data available for training individual models, which could impair their performance. It also requires the alignment of resulting models in a common space, which could complicate the interpretation of the results (Hamilton et al., 2016).

In our work, we build upon these ideas by enhancing word embeddings with socio-demographic information and focusing on studying the self. More specifically, we replace every occurrence of the word 'I' in a large corpus of social media posts with $\mathrm{I}_{g,a}$ tokens, where g represents the gender of the post author and a their age. We then train a word embedding model on the altered corpus. Projecting the resulting enhanced vectors on semantic axes allows exploring identity as expressed in social media posts. By incorporating socio-demographic information into the I-tokens, it also becomes possible to compare different social groups without splitting the original corpus.

In the remainder of this paper, we provide a more detailed description of our method. We then characterise the obtained enhanced I-tokens and verify whether they meet the criteria for face validity. To further validate our approach, we check if it can reproduce established findings on gendered self-views. Next, we investigate the robustness of the results with respect to model specifications and corpus size. Finally, we discuss how our approach can be applied in different contexts and compare it with existing methods.

2 Methods

2.1 Data & Model

To train the model, we used data on 62,707,791 posts shared over a span of 5 years by 913,230 users on VK[1]–a popular social media platform predominantly used by Russian speakers. The process of collecting the corpus and filtering out fake profiles has been previously detailed in (Smirnov, 2017) and (Sivak and Smirnov, 2019). Unlike on many other social media platforms, age and gender are mandatory fields of a user profile on VK and are publicly available via its API. This allows us to construct $\mathrm{I}_{g,a}$ tokens for all posts in the dataset. While we use VK data for the results described in this paper, our approach could equally be applied to other data sources and to attributes beyond gender and age (see Discussion).

We normalized all adjectives and nouns in the corpus using pymorphy2, the state-of-the-art morphological analyzer and generator for Russian and Ukrainian languages (Korobov, 2015). This step is necessary because, in Russian, nouns and adjectives have distinct feminine and masculine forms. This makes words in feminine form artificially closer to $\mathrm{I}_{\mathrm{woman},*}$ tokens in vector space and words in masculine form closer to $\mathrm{I}_{\mathrm{man},*}$, preventing meaningful comparisons.

[1] https://vk.com

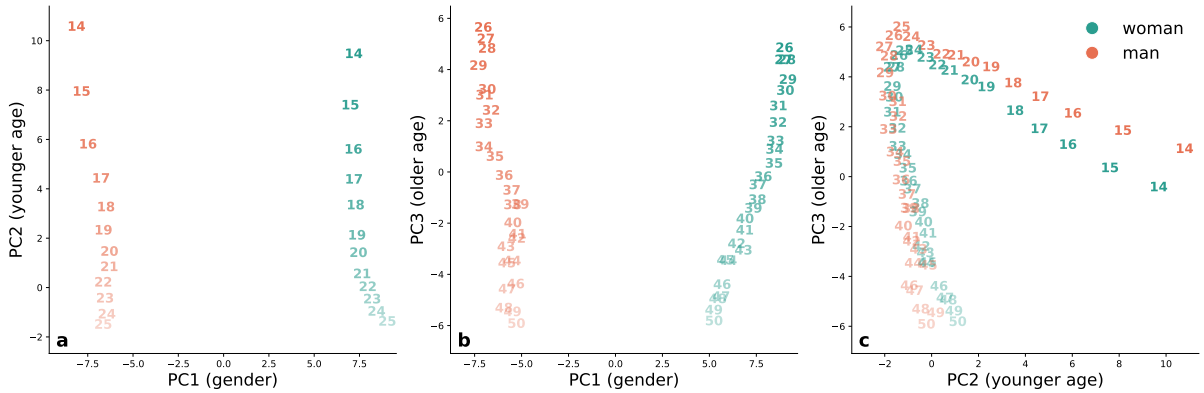

Figure 1: **The structure of enhanced I-token embeddings.** The first principal component extracted from embeddings of enhanced I-tokens corresponds to gender (a, b). Curiously, age is represented by two components: the second component corresponds to a younger age (a), while the third corresponds to an older age (b).

Next, we replaced all singular first-person pronouns used in posts with $\mathrm{I}_{g,a}$ tokens, where g and a correspond to the self-reported gender of an author and their self-reported age at the time of writing, e.g. 'I_woman_42', 'I_man_19'. We then trained a continuous bag-of-words model (Mikolov et al., 2013) with 100 dimensions over 10 epochs on this modified corpus. We report the main results for this specific model configuration; however, we also examine their sensitivity to model type (CBOW vs skip-gram), number of dimensions, number of epochs, and corpus size.

We examined the geometric structure of the obtained enhanced embeddings to ensure their face validity. Specifically, we expect $\mathrm{I}_{\mathrm{man},*}$ and $\mathrm{I}_{\mathrm{woman},*}$ to be clearly separated in vector space. We also expect that $\mathrm{I}_{g,a}$ tokens will be sequentially ordered by age, i.e., that $\mathrm{I}_{g,i}$ would be between $\mathrm{I}_{g,i-1}$ and $\mathrm{I}_{g,i+1}$.

Gendered self-views

To further validate our approach, we checked if it can reproduce existing findings on sex-trait stereotypes. Sex-trait stereotypes refer to the psychological characteristics or behavioral traits believed to be more prevalent in women than in men, or vice versa (Williams and Best, 1990). A common way to assess sex-trait stereotypes is to present participants with a series of adjectives and ask them to determine whether each adjective is more commonly associated with women or men. From such studies emerged a list of adjectives that participants consistently associate more with either women or men, whether they are describing others or themselves. Examples include 'affectionate' and 'sensitive' for women, and 'courageous' and 'ambitious' for men (the full list of adjectives used in this study is available in Table 1.1 of (Williams and Best, 1990)).

We translated this list into Russian and constructed a semantic axis (*gender stereotype axis*) by subtracting the average embedding for men-associated adjectives from the average embedding for women-associated adjectives. The original list consisted of 29 adjectives for women and 32 for men. This was reduced to 27 and 28 respectively, due to some English words having identical translations in Russian. While the original list was obtained by asking Euro-American college students, recent studies demonstrate that women and men consistently rate themselves higher on corresponding traits across 62 countries (Kosakowska-Berezecka et al., 2023). Thus, if our approach is valid, we expect the projections of $\mathrm{I}_{\mathrm{woman},*}$ on the *gender stereotype axis* to be positive, while projections of $\mathrm{I}_{\mathrm{man},*}$ to be negative.

3 Results

We found that the variation between $\mathrm{I}_{g,a}$ tokens is largely explained by gender and age variables. In particular, the first principal component extracted from these vectors corresponds to gender, clearly separating $\mathrm{I}_{\mathrm{woman},*}$ from $\mathrm{I}_{\mathrm{man},*}$ tokens (Figure 1a and 1b). The point-biserial correlation coefficient between gender and the first component is 0.986 ($P < 10^{-61}$).

We expected that the second principal component would correspond to age. However, the results are more nuanced: the second component corresponds to younger age (Figure 1a) with Spearman's $\rho = 0.965$, $P < 10^{-13}$, while the third component corresponds to older age (Figure 1b) with Spear-

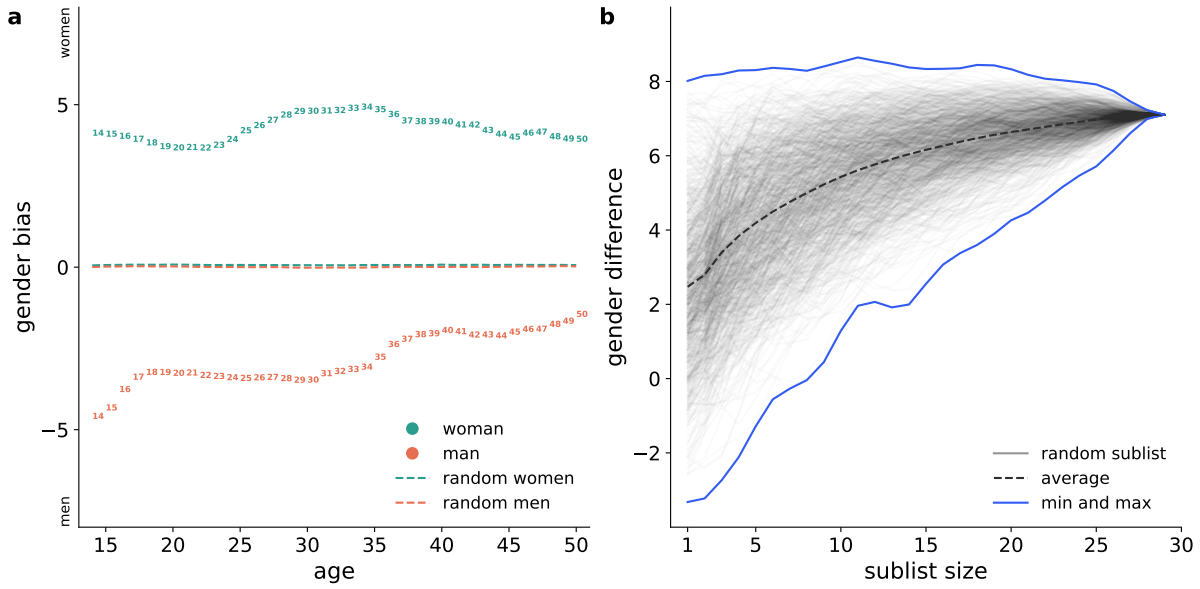

Figure 2: **Projection of enhanced I-tokens on *gender stereotype axis* reproduces established findings on gendered self-views.** $I_{woman,*}$ tokens are closer to women's pole of the axis, while $I_{man,*}$ tokens are closer to the men's pole (a). The gap between them narrows with age as $I_{man,*}$ tokens shift towards the center. The results are significant with $P < 10^{-3}$ and are robust with respect to the selection of adjectives, starting from a list size of around 10 (b). For visual clarity, a moving average with a window size of 3 is used.

man's $\rho = 0.928, P < 10^{-24}$. The interaction between these components and age is shown in Figure 1c. We hypothesise that this might be explained by graduation from university and the transition to working life, as the curve's turning point (25–26 years in Figure 1c) roughly matches the age when students typically complete their degrees in Russia.

Gendered self-views

If our approach is valid, we expect that projections of $I_{woman,*}$ on the *gender stereotype axis* will be positive, and projections of $I_{man,*}$ will be negative. This is indeed what we observe (see Figure 2a). We tested the significance of this result by randomly shuffling adjectives used to construct the *gender stereotype axis* and projecting the enhanced I-tokens on the resulting random axes. None of the biases computed for $1,000$ random axes were as strong as the one we observed, making our results significant with $P < 10^{-3}$.

We also tested the robustness of our results with respect to dictionary size, following the method suggested in (Spliethöver and Wachsmuth, 2021). To do this, we randomly selected k adjectives, with k varying from 1 to 28, from both the men-associated and women-associated lists. We then used these shorter lists to construct a gender axis and compute the bias, repeating the procedure $1,000$ times. This shows that the bias is consistently detected with a list size of around 10 adjectives for each gender (Figure 2b). Note that for this analysis we did not consider age separately, but computed the differences between the projections of aggregated I_{man} and I_{woman} on the *gender stereotype axis*, where the aggregated vectors represent averages over all ages.

Additionally, we were able to detect changes in the strength of this relationship over the years–a result that is difficult to capture in surveys, as they are typically conducted on samples of university students (Williams and Best, 1990; Kosakowska-Berezecka et al., 2023). This demonstrates the potential of our methodology not only to reproduce established findings but also to gain new insights that might be harder to obtain via traditional methods.

Robustness of the results

We evaluated how well the observed geometrical structure of enhanced I-tokens is preserved across different model specifications. We also checked for the robustness of the relationship between gender and gender-stereotypical adjectives. For this purpose, we computed point-biserial correlation coefficients between gender and the first principal component extracted from I-tokens, as well as between gender and the projection of I-tokens on the *gender stereotype axis* (see Figure 3).

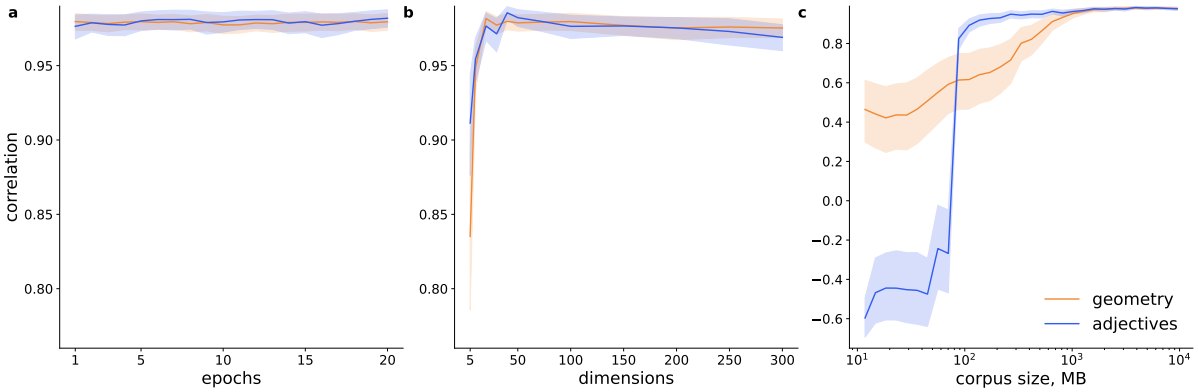

Figure 3: **Robustness of the results with respect to model specification.** We evaluated how much point-biserial correlations between gender and the first principal component extracted from enhanced I-tokens (orange), as well as between gender and the projection of I-tokens on the *gender stereotype axis* (blue), depend on model specification. We found that no further training is required beyond one epoch to reproduce the results (a). We also found that any reasonable number of dimensions can be used (b). Finally, we found that 100MB is a sufficient corpus size, but beyond that point, the performance drops for adjectives as they become too rare. The first principal component of enhanced I-tokens remains strongly associated with gender for all our experiments.

Although we trained the model for 10 epochs for the reported results, we found that further training offers little additional benefit and the main results can be reproduced after just one epoch (Figure 3a). We also found that any reasonable number of dimensions (50–300) could be used without compromising the model's performance (Figure 3b). The observed relationships are even more salient when vector sizes are between 50 and 100, which could be preferable due to the smaller model size. There was no substantial difference between the CBOW and skip-gram architectures.

We found that 100MB of data is sufficient to reproduce the results after training for one epoch (Figure 3c). For smaller datasets, performance drops for adjectives because they become too rare in the corpus. The first principal component of enhanced I-tokens is strongly correlated with gender in all our experiments. In practice, an even smaller corpus could be used. For example, the corpus of interest could be augmented by a neural one, such as a Wikipedia dump. This should result in better representations of rare words without affecting enhanced I-tokens, as they would only be present in the original corpus of interest.

4 Discussion

In this paper, we introduced a novel approach that leverages readily available data sources, such as social media, to study identity. Unlike traditional methods that rely on self-report surveys, our method allows for the study of identity in natural settings and on a larger scale. While we used data from VK, the same technique can be applied to other datasets as well. For example, self-reported gender and age have been extracted from posts on popular platforms such as Reddit and Twitter (Tigunova et al., 2020; Klein et al., 2022), making it possible to apply our method directly to these datasets. Attributes that can be used to construct enhanced I-tokens are not limited to gender and age. For instance, with datasets containing profession information on Reddit (Tigunova et al., 2020) or educational outcomes on VK (Smirnov, 2019), it becomes possible to study differences between various socio-economic groups. Moreover, this approach can be extended beyond social media data. Our experiments demonstrate that the corpus does not need to be exceptionally large for the method to be effective. Therefore, it could be applied to TV scripts to analyse the representation of different groups on television, building upon previous research in this area (Ramakrishna et al., 2015, 2017).

As a proof of concept, we applied our method to study gendered self-views. We found that the approach not only reproduces established results but also allows for new findings by covering a wider age range than is typically available in surveys. This method can similarly be applied to other phenomena using curated word lists. Alternatively, an open dictionary approach can be used to identify and examine words that are especially close to certain enhanced I-tokens in a corpus of interest.

The introduced method relies on natural language processing techniques that are admittedly no longer considered state-of-the-art. Since the introduction of word2vec (Mikolov et al., 2013), more advanced models have emerged, particularly fastText (Bojanowski et al., 2017), which operates at the character n-gram level and potentially offers superior embeddings for morphologically rich languages such as Russian. Later, contextual word embedding models were developed, most notably BERT (Devlin et al., 2019), which outperformed static models in a wide range of tasks. However, we believe that advances in machine learning outpace their adoption in social sciences, and there are still many opportunities for new insights to be obtained from using static continuous representations of words. While newer models have led to remarkable performance gains in machine learning applications, we believe that the higher interpretability and computational efficiency of simpler models might still make them preferable for analytical purposes and applications in social science.

The idea of using semantic projections traces its origins back to at least 2016 (Bolukbasi et al., 2016), when a gender axis was constructed to reveal biases in word embeddings. This methodology was later formally introduced in (An et al., 2018), re-introduced in (Mathew et al., 2020), and re-introduced again in (Grand et al., 2022). It was further extended to contextual word embeddings (Lucy et al., 2022; Engler et al., 2022). Despite these developments within the computer science literature, their adoption in social sciences has been relatively slow. One possible explanation is that these methods enable the identification of biases at an aggregated level of entire corpora, which, while interesting, has limited applications. In our paper, we build upon previous ideas and show how they can be extended to study differences between social groups. We believe this opens up many new possibilities that would be particularly appealing to social scientists.

Data and Code

The data and code used to obtain the main results of this paper are available at `https://github.com/ibsmirnov/echoes-of-i`.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. pages 2450–2461, Melbourne, Australia.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Andrei Boutyline, Alina Arseniev-Koehler, and Devin J Cornell. 2023. School, studying, and smarts: Gender stereotypes and education across 80 years of american print media, 1930–2009. *Social Forces*, page soac148.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, Minneapolis, Minnesota.

Jan Engler, Sandipan Sikdar, Marlene Lutz, and Markus Strohmaier. 2022. SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings. pages 4607–4619, Abu Dhabi, United Arab Emirates.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

John Manners Gilmore and Kevin Durkin. 2001. A critical review of the validity of ego development theory and its measurement. *Journal of personality assessment*, 77(3):541–567.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. pages 1489–1501, Berlin, Germany.

Le Xuan Hy and Jane Loevinger. 1996. *Measuring ego development*. Lawrence Erlbaum Associates, Inc.

Ari Z Klein, Arjun Magge, and Graciela Gonzalez-Hernandez. 2022. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *PloS one*, 17(1):e0262087.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Natasza Kosakowska-Berezecka, Jennifer K Bosson, Paweł Jurek, Tomasz Besta, Michał Olech, Joseph A Vandello, Michael Bender, Justine Dandy, Vera Hoorens, Inga Jasinskaja-Lahti, et al. 2023. Gendered self-views across 62 countries: A test of competing models. *Social Psychological and Personality Science*, 14(7):808–824.

Miriam Koschate, Elahe Naserian, Luke Dickens, Avelie Stuart, Alessandra Russo, and Mark Levine. 2021. Asia: Automated social identity assessment using linguistic style. *Behavior Research Methods*, 53:1762–1781.

Kevin Lanning, Rachel E Pauletti, Laura A King, and Dan P McAdams. 2018. Personality development through natural language. *Nature Human Behaviour*, 2(5):327–334.

Mark R Leary and June Price Tangney. 2003. The self as an organizing construct in the behavioral and social sciences. *Handbook of self and identity*, 15:3–14.

Jane Loevinger. 1976. *Ego development*. Jossey-Bass, San Francisco.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. pages 3477–3494, Abu Dhabi, United Arab Emirates.

Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *Proceedings of The Web Conference 2020*, pages 1548–1558.

Kate C McLean and Moin Syed. 2015. The field of identity development needs an identity: An introduction to the oxford handbook of identity development. *The Oxford handbook of identity development*, pages 1–10.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Max Pellert, Hannah Metzler, Michael Matzenberger, and David Garcia. 2022. Validating daily social media macroscopes of emotions. *Scientific reports*, 12(1):11236.

Anil Ramakrishna, Nikolaos Malandrakis, Elizabeth Staruk, and Shrikanth Narayanan. 2015. A quantitative analysis of gender differences in movies using psycholinguistic normatives. pages 1996–2001, Lisbon, Portugal.

Anil Ramakrishna, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. pages 1669–1678, Vancouver, Canada.

Pedro L Rodriguez and Arthur Spirling. 2022. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1):101–115.

Elizaveta Sivak and Ivan Smirnov. 2019. Parents mention sons more often than daughters on social media. *Proceedings of the National Academy of Sciences*, 116(6):2039–2041.

Ivan Smirnov. 2017. The digital flynn effect: Complexity of posts on social media increases over time. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pages 24–30. Springer.

Ivan Smirnov. 2019. Schools are segregated by educational outcomes in the digital space. *PloS one*, 14(5):e0217142.

Maximilian Spliethöver and Henning Wachsmuth. 2021. Bias silhouette analysis: Towards assessing the quality of bias metrics for word embedding models. In *IJCAI*, pages 552–559.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. RedDust: a large reusable dataset of Reddit user traits. pages 6118–6126, Marseille, France. European Language Resources Association.

Vivian L Vignoles, Seth J Schwartz, and Koen Luyckx. 2011. Introduction: Toward an integrative view of identity. In *Handbook of identity theory and research*, pages 1–27. Springer.

John E Williams and Deborah L Best. 1990. *Measuring sex stereotypes: A multination study, Rev*. Sage Publications, Inc.