Richer Countries and Richer Representations

Kaitlyn Zhou
Stanford University
katezhou@stanford.edu

Kawin Ethayarajh
Stanford University
kawin@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

We examine whether some countries are more richly represented in embedding space than others. We find that countries whose names occur with low frequency in training corpora are more likely to be tokenized into subwords, are less semantically distinct in embedding space, and are less likely to be correctly predicted: e.g., Ghana (the correct answer and in-vocabulary) is not predicted for, "The country producing the most cocoa is [MASK].". Although these performance discrepancies and representational harms are due to frequency, we find that frequency is highly correlated with a country's GDP; thus perpetuating historic power and wealth inequalities. We analyze the effectiveness of mitigation strategies; recommend that researchers report training word frequencies; and recommend future work for the community to define and design *representational guarantees*.

1 Introduction

How similar are the words "Brooklyn" and "Queens"? To a New Yorker, they evoke two very different places, cultures, and cuisines, but to a Seattleite, they are quite similar, both being boroughs of New York City.[1] Our perception of entities such as cities or countries is conditioned on our backgrounds. Here, we ask if language models are also susceptible to representational biases.

We suggest three criteria to characterize the quality of representations for particular entities or groups: consistency, distinctiveness, and recognizability. For consistency, are all entities of a certain type (such as all country names) represented with the same number of tokens in the lexicon? For distinctiveness, are entities of the same category seen as equally distinct in representational space? For recognizability, are models capable of generating

all entities of a certain type in response to questions? And are the differences between entities confounded across lines of historical inequity (like wealth of countries)?

Focusing on BERT (`bert-base-cased`[2]) representations ([Devlin et al., 2019](#)), we find that names of countries that appear less frequently in training data are less likely to be in-vocabulary, are less semantically distinct from other countries, and are less frequently predicted in the masked language modeling (MLM) task. Disappointingly, we find similar behavior in `bert-base-multilingual-cased` and `roberta-base`. We identify these differences as intrinsic representational harms where low frequency countries are more likely to be conflated with one another and their existence less recognized.

A more troubling result is that training data frequency is highly correlated with the gross domestic product of a country (GDP) (Pearson's $r = 0.82$). Our training data and thus the representation of entities through our language models encodes wealth and power disparities and perpetuates representational harms. Given these significant differences in representation, what could it look like to impose a minimum quality of representation for significant entities? We recommend that the community consider designing *representational guarantees* for language models.

In summary we: 1) reveal multiple ways in which the BERT representation of high GDP countries is systematically richer than that of low GDP countries; 2) study the effectiveness of potential mitigation efforts; and 3) propose the idea of *representational guarantees* as future work for the community.

---

[1]https://en.wikipedia.org/wiki/View_of_the_World_from_9th_Avenue.

[2]https://huggingface.co/bert-base-cased

2 Related Work

Language technologies have long been studied for potential intrinsic and extrinsic harms (Galliers and Jones, 1993). Known intrinsic harms include misrepresentation of gender (Bolukbasi et al., 2016), race (Abid et al., 2021), and ability (Hutchinson et al., 2020) — all types of representational harms (Barocas et al., 2017; Crawford, 2017; Blodgett et al., 2020). Other intrinsic harms include forms of erasure through under-representation of LGBTQ+ identity terms (Strengers et al., 2020; Oliva et al., 2021) and racial groups (Gehman et al., 2020). Extrinsic harms are often found in downstream tasks and include disparities in quality of service among user groups (Zhang et al., 2020) such as African-American users (Blodgett and O'Connor, 2017; Koenecke et al., 2020). However, low statistical power has also made it difficult to make conclusive claims about the presence or absence of bias (Ethayarajh, 2020).

Many of these representational harms have been linked to word frequency in static embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018; Bordia and Bowman, 2019; Ethayarajh et al., 2019b; van Loon et al., 2022). Low frequency words also differ geometrically from other words, with smaller inner products (Mimno and Thompson, 2017) and lower variance (Ethayarajh et al., 2019a). Recent work has also shown how frequency impacts *contextual* embeddings such as the under-estimation of cosine similarity among high-frequency words (Zhou et al., 2022) and the discrepancies in representations of personal names (Shwartz et al., 2020; Wolfe and Caliskan, 2021). Our work extends these lines of work via the examination of representational harms for country names.

3 Rich Countries have their own Tokens

Are poor and rich countries tokenized the same way? Here, we focus on BERT's tokenization process and measure the consistency (or rather inconsistency) in how names of countries are tokenized and then represented. Our GDP data is retrieved from the United Nations Statistics Division from 2019 [3][4].

Of the 159 single-word countries names from

Subwords	Freq	GDP (M)	Example
1 (n=134)	74,882	430,596	Uzbekistan
2 (n=32)	68,148	870,702	Comoros
3 (n=15)	8,711	34,896	Grenada
4+ (n=12)	4,309	14,980	Eswatini

Table 1: The average BERT training data frequency and GDP associated with the countries, binned by the number of subwords the country was tokenized into (e.g., "Grenada" was tokenized into three subwords.). Using OLS to predict number of subwords, frequency explains 38% of the variation in number of subwords.

the United Nations members list, 134 of them are in-vocabulary, — the remaining 25 are out-of-vocabulary (OOV). In WordPiece tokenization, OOV words are tokenized into in-vocabulary subwords (e.g. "Andorra" becomes "And" and "##orra", see table 1). Additionally, as a limitation of the unigram vocabulary, the 34 multi-word country names (e.g. "United States") are also OOV and represented as distinct tokens ("United" and "States".)[5] Each word of multi-word countries can also be OOV (e.g., Sao Tome and Principe is tokenized into 9 different subwords).

We used ordinary least squares regression to predict the number of subword tokens in each country name, using training data frequency of each country name as the feature (BERT training data estimated from the March 1st, 2020 Wikimedia Download and BookCorpus) (Zhu et al., 2015; Hartmann and dos Santos, 2018).[6]. We found that training data frequency explains 38% of the variance in number of subwords ($p < 0.01$), despite the confounder of multi-word countries being considered OOV (Table 2 in Appendix). This is likely due to the fact that the tokenizer builds its vocabulary based on fitting training likelihood. Given that the training data frequency of a country's name correlates strongly with its GDP (Pearson's $r = 0.82$), the countries that have the highest number of subwords are also the ones with the lowest GDP.

---

[3] https://unstats.un.org/unsd/snaama/Basic; GDP data was shown in USD.

[4] Code for this paper can be found at https://github.com/katezhou/country_distortions

[5] Acronyms of some high GDP countries e.g. US, USA, UK, UAE are in-vocabulary but are not included in this study to avoid introducing confounders concerning which to include and how to pool. This is likely a conservative bias since including them would have increased the effects we study.

[6] Additional tools used: https://github.com/IlyaSemenov/wikipedia-word-frequency; https://github.com/attardi/wikiextractor
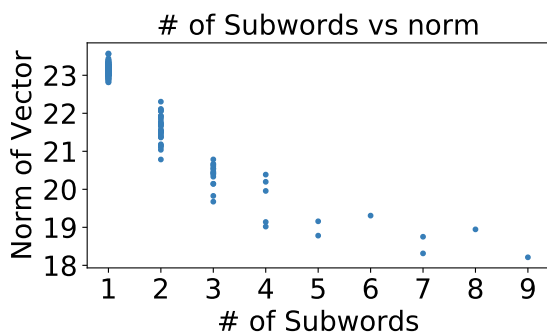
Figure 1: Average norm of embeddings in relation to the number of subwords of the embedding. OOV words are represented by averaging subwords. Pearson's correlation between average norm and number of subwords, $r = -0.92$.

3.1 Geometric Impact of being OOV

The methods used to represent OOV words can impact the geometry of their representations. For example, the representation of OOV words often have smaller norms (L^2 of the word embedding) than in-vocabulary words. This is because a common way to represent OOV words is to take the average of their subwords (Pilehvar and Camacho-Collados, 2019; Blevins and Zettlemoyer, 2020; Bommasani et al., 2020). However, since the average values for each dimension are near zero, the more subwords that are averaged, the smaller the norm of the vectors (Adi et al., 2017) (figure 1). As a result, the Pearson's correlation between the norm and number of subwords is $r = -0.92$. A common alternative to averaging subwords is to use the first subword to represent an OOV word. When this representation is used, the correlation between the norm and the number of subwords reverses and is slightly positive, Pearson's $r = 0.22$ (figure 4 in Appendix). One possible explanation could be that the first subword of OOV words are more likely to be stop words, which are known to have larger norms (Ethayarajh, 2019). The difference in geometry between OOV and in-vocabulary words exists in both representation methods. This could could result in impacts on tasks using embedding-based retrieval (e.g., nearest-neighbor LM) as low-frequency names will have additionally distinguishing geometric characteristics as a result of tokenization.

Inconsistency in tokenizing country names leads to inconsistency in geometric representation. A potential mitigation might be to have all country names be in-vocabulary with a dedicated token. This might not address all impacts of training data frequency imbalances, but would at least prevent additional geometric differences due to tokenization.

4 Richer Countries are Most Distinct in Embedding Space

Nations or entire regions are subject to bias. The African continent, for example, is often treated journalistically as a single homogeneous entity (Nothias, 2018), as if African countries are all substitutable for one another. We draw on this finding to ask whether historically disadvantaged countries are also conflated with one another in the embeddings of their names (i.e., seen as less distinct from each other than other countries).

We measure the semantic similarity between pairs of the 134 in-vocabulary countries by creating word embeddings for each name (done by averaging the last four hidden layers of BERT). We calculate the average in-group cosine similarity of country names as grouped by frequency (i.e., we take the countries in each decile of frequency and measure the average cosine similarity across the $\binom{14}{2}$ pairs). We repeat this ten times and find that the 10% least frequent names have an average in-group similarity of 0.610 compared to an average similarity of 0.582 for the 10% most frequent countries (mean $\delta = 0.028$, permutation test $p < 0.01$).

We then calculate the average semantic similarity of a country's embedding to all other countries to measure a country's *distinctiveness* (averaged over ten trials). Using OLS to predict average cosine similarity, frequency explains 8% of the variance (Table 4 in Appendix). For example, in our experiments, France had a cosine similarity of ≥ 0.7 with 21 other countries while Haiti shared a cosine similarity of ≥ 0.7 with 59 other countries. France's distinctiveness contrasts with Haiti's similarity with other countries. Using these embeddings and cosine similarities in a downstream task like IR (or MT, where embedding cosines are used in algorithms like BERTScore) would yield vastly different results despite using the same threshold. We visualize how average cosine similarity correlates with a country's GDP in Figure 2.

To ensure that the semantic similarity discrepancies are not simply a consequence of how these countries are written about in test examples, we run the same OLS experiment on an artificial dataset where names of countries appear in identical contexts. The results are consistent: frequency explains 9% of the variance in average cosine similar-
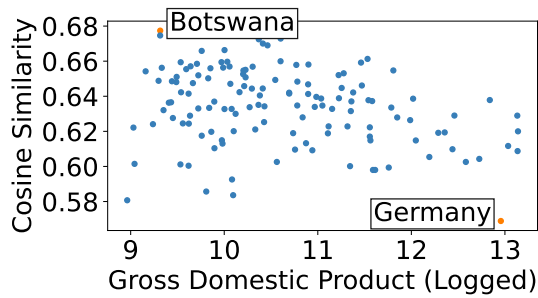
Figure 2: Average cosine similarity of a country to all other countries vs. its GDP. Pearson's $r = -0.28$ for average cosine similarity and GDP. Wealthier countries are more distinct in BERT embedding space.

Figure 3: Number of times a country was predicted in the MLM task versus its GDP (logged). Pearson's correlation between GDP (logged) and number of times a country is predicted, $R = 0.64$.

ity (Table 5 and Table 6 in Appendix). Independent of *how* these countries are being written about in any potential downstream task, NLP models like BERT embed country names in a way that results in higher semantic similarity for low frequency names, and hence low GDP countries. This is a representational harm: distinctiveness of a nation's name in NLP representations correlates with the nation's wealth, resulting in poorer countries being more likely to be conflated with one another.

4.1 Tokenization and Ssemantic Similarity

As discussed in the section above, tokenization results in geometric differences between OOV and in-vocabularly words, and here we examine tokenization's effect on cosine similarity. OOV country names have higher in-group similarity averages when averaging subpieces (0.668 vs 0.628; mean $\delta = 0.040$, permutation test $p < 0.01$). Given that OOV words have smaller norms and are closer to the centroid, this signals a concentration of low-frequency words. However, when using the first subpiece to represent OOV words, OOV country names have lower in-group similarity averages than in-group similarity among in-vocabulary words (0.589 vs 0.625; mean $\delta = 0.037$, permutation test $p < 0.001$) — we showed that these embeddings conversely have larger norms and could be more widely dispersed. The key takeaway here is that both methods show semantic differences between in-vocabulary and OOV words; again there are potential impacts in embedding-based downstream tasks.

5 Richer Countries are more Frequently Predicted

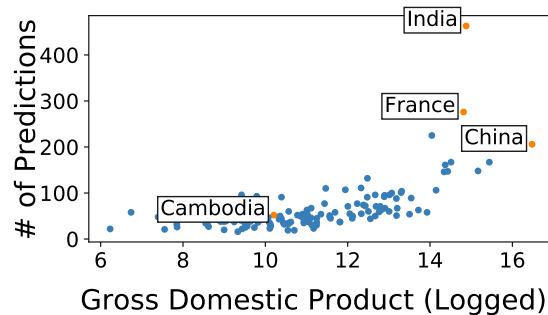The lack of recognition of people and groups (Strengers et al., 2020; Oliva et al., 2021; Gehman

et al., 2020) has often been cited as an representational harm. Here, we use the masked language modeling task as a proxy for many downstream tasks that need to be able to predict the name of a country (e.g., as the answer to a question, or in a summary) to measure whether countries are all minimally predicted or whether instead we see representational harms such as erasure.

We randomly sample sentences from Wikipedia that contain the name of a country, replace the name with a BERT mask token ([MASK]), and use the masked-out country name as the gold label. We use 100 examples of each of the 134 in-vocabulary country names.[7] The model will only predict in-vocabulary words, which again illustrates the impact of inconsistent tokenization: the very task BERT was trained on is unable to handle OOV country names without modification.

High frequency countries (75th percentile) have an average accuracy of 42% while low frequency (25th percentile) countries have an accuracy of 26% (table 7 in Appendix). How often a country is predicted is highly correlated with its training data frequency (Pearson's $r = 0.64$, figure 3). Of all the country names predicted, the 10 least predicted countries make up 2% of total guesses compared to 25% for the top 10 most predicted countries. BERT fails this task in drastic ways; predicting China, India, Brazil for, "The poorest country in the world is [MASK]." (Burundi and Somalia ranked as the poorest by GDP/capita). This reduced recognition of poor countries is a representational harm.

6 The Limitations of Mitigation Efforts

We analyze the effectiveness of two popular mitigation efforts.

---

[7]This is a significantly harder task than ordinary MLM as only names of countries are masked out, there could be multiple appropriate answers, and a limited context is given.

1. Could these frequency-based effects be mitigated with **additional training data**? We measure the performance of Multilingual BERT (BERT-ML) which includes Wikipedia articles from 104 other languages. We continue to find accuracy disparities between low and high GDP countries (15% vs 29%). Additional training data fails to mitigate these harms, most likely because the additional data continues to amplify existing imbalances (i.e., German, French, and Polish are the next three biggest languages of Wikipedia articles). Data augmentation as a mitigation technique is challenging as datasets could easily maintain existing or introduce new frequency imbalances. We also tested RoBERTa (trained on over 160GB more data) on this task with similar results (table 7 in Appendix).

2. Could fine-tuning or **continuing pre-training** mitigate these harms? We select 20 random countries (Appendix table 8); for each, we select 1,000 random sentences from Wikipedia that mentions the country's name. After four epochs of training, we then test on an evaluation dataset (100 examples/country) and we see a 13% increase in performance on our selected countries and a 3% decrease for other countries (table 9 and figure 5 in Appendix). Our subset of interested countries originally made up 17% of all guesses, but this rate more than doubles in our tuned model. This pre-training mitigation method shows promise but has trade-offs in performance, requires practitioners to be aware of inequalities, and have access to enough training samples to continue pre-training effectively.

7   Discussion and Conclusion

We find significant disparities in the quality of representation of country names and show how these differences result in representational harms that perpetuate existing wealth and power inequalities. We make two recommendations on paths forward. We recommend the release of training word frequencies to increase transparency and isolate current representational harms (Gebru et al., 2021; Mitchell et al., 2019; Bender and Friedman, 2018; Ethayarajh and Jurafsky, 2020). Practitioners who use these models in their systems and research should have access to the topic and entity distribution of our models given the potential for frequency-related harms. We also recommend the community consider designing *representational guarantees* for significant entities to mitigate these

downstream harms. Our work illustrates the potential harms that arise when entities such as country names do not have representational guarantees. We encourage the community to consider the following questions: How can we ensure entities such as names of country be in-vocabulary? How can we guarantee a minimum distinctness and ensure recognition of historically disadvantaged groups? Such guarantees will likely be difficult to design and will require expertise from multiple domains but mitigating these representational harms is an important task that we cannot ignore.

Acknowledgements

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. *In Proceedings of SIGCIS, Philadelphia, PA*.

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

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017.

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.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota.

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

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China.

Kawin Ethayarajh. 2020. Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2914–2919.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

J.R. Galliers and K.S. Jones. 1993. *Evaluating Natural Language Processing Systems*. Computer Laboratory Cambridge: Technical report. University of Cambridge, Computer Laboratory.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online.

Nathan Hartmann and Leandro Borges dos Santos. 2018. NILC at CWI 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 335–340, New Orleans, Louisiana.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Allison Koenecke, Andrew Joo Hun Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Toussaint Nothias. 2018. How western journalists actually write about africa. *Journalism Studies*, 19(8):1138–1159.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25(2):700–732.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning 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 1267–1273, Minneapolis, Minnesota.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Austin van Loon, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. 2022. Regional negative bias in word embeddings predicts racial animus– but only via name frequency. *arXiv preprint arXiv:2201.08451*.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 110–120, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Appendix

A.1 Appendix for section 3

To illustrate the skew of BERT's vocabulary on a larger dataset, we repeat this experiment for names of cities across the world.[8] Cities, similar to country names are also information-rich entities representing peoples and places. Filtering for populous (>100,000) single-word cities, 50% of North American cities and 25% of European cities are in-vocabulary compared to less than 6% of city names from Asia, Africa, and Central and South America (Table 3). Contrast Nigeria — world's fourth largest population of English speakers — with the United Kingdom. Lagos is the only city in-vocabulary out of a possible 89 populous cities, compared to the United Kingdom where 53 out of its 64 populous cities are in vocabulary.

A.2 Appendix for section 4

A.3 Appendix for section 5

---

[8]The city and population data was created by MaxMind, available from http://www.maxmind.com/.

	coef	std err	t	P> |t|	[0.025	0.975]
Dep. Variable:	# of subwords		**R-squared:**		0.384	
Model:	OLS		**Adj. R-squared:**		0.380	
Method:	Least Squares		**F-statistic:**		118.8	
Date:	Mon, 15 Nov 2021		**Prob (F-statistic):**		7.95e-22	
Time:	16:41:27		**Log-Likelihood:**		-272.30	
No. Observations:	193		**AIC:**		548.6	
Df Residuals:	191		**BIC:**		555.1	
Df Model:	1					

	coef	std err	t	P> |t|	[0.025	0.975]
Constant	7.1702	0.515	13.912	0.000	6.154	8.187
freq_logged	-0.5502	0.050	-10.901	0.000	-0.650	-0.451

Omnibus:	90.415	**Durbin-Watson:**	2.177
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	339.064
Skew:	1.900	**Prob(JB):**	2.36e-74
Kurtosis:	8.265	**Cond. No.**	74.0

Table 2: OLS Regression Results: Using training data frequency (logged) to predict number of subwords as tokenized by BERT. Training word frequency explains 38% of the variance in number of subpieces as tokenized by BERT.

Region	#	In-Vocab	Population
Africa	372	5%	23,296,502
Americas	594	7%	9,335,510
Asia	1,466	4%	34,046,146
Europe	372	25%	11,278,318
N America	49	50%	1,834,020
Oceania	20	75%	1,316,862

Table 3: Average number of populous (>100,000) cities that are in BERT-base-cased's vocabulary.

Dep. Variable:	Average Cosine Similarity		**R-squared:**		0.084	
Model:	OLS		**Adj. R-squared:**		0.077	
Method:	Least Squares		**F-statistic:**		12.13	
Date:	Sat, 13 Nov 2021		**Prob (F-statistic):**		0.000675	
Time:	11:05:47		**Log-Likelihood:**		330.05	
No. Observations:	134		**AIC:**		-656.1	
Df Residuals:	132		**BIC:**		-650.3	
Df Model:	1					

	coef	std err	t	P> |t|	[0.025	0.975]
Constant	0.6973	0.019	37.114	0.000	0.660	0.735
freq_logged	-0.0061	0.002	-3.482	0.001	-0.010	-0.003

Omnibus:	8.539	**Durbin-Watson:**	1.998
Prob(Omnibus):	0.014	**Jarque-Bera (JB):**	8.470
Skew:	-0.604	**Prob(JB):**	0.0145
Kurtosis:	3.236	**Cond. No.**	113.

Table 4: OLS Regression Results: Using training data frequency (logged) to predict the average cosine similarity between a country compared to all other countries (in-vocabulary countries only). Frequency explains 8% of the variance. The more frequent the country, the lower average cosine similarity — indicating its distinctness from all other countries.

Dep. Variable:	Average Cosine Similarity		**R-squared:**		0.086	
Model:	OLS		**Adj. R-squared:**		0.079	
Method:	Least Squares		**F-statistic:**		12.47	
Date:	Mon, 15 Nov 2021		**Prob (F-statistic):**		0.000570	
Time:	16:46:42		**Log-Likelihood:**		440.48	
No. Observations:	134		**AIC:**		-877.0	
Df Residuals:	132		**BIC:**		-871.2	
Df Model:	1					

	coef	std err	t	P> |t|	[0.025	0.975]
Constant	0.8458	0.008	102.629	0.000	0.830	0.862
freq_logged	-0.0027	0.001	-3.531	0.001	-0.004	-0.001

Omnibus:	6.442	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.040	**Jarque-Bera (JB):**	3.258
Skew:	-0.112	**Prob(JB):**	0.196
Kurtosis:	2.270	**Cond. No.**	113.

Table 5: OLS Regression Results: Using training data frequency (logged) to predict the average cosine similarity between a country compared to all other countries (in-vocabulary countries only) when names of countries appear in identical contexts. Frequency explains 9% of the variance. The more frequent the country, the lower average cosine similarity — indicating its distinctness from all other countries.

Sentence
I am from COUNTRY.
I live in COUNTRY.
I hope this January I will get to travel to COUNTRY.
I am interesting in traveling to COUNTRY.
My friend is from COUNTRY.
COUNTRY is well known for its history.
COUNTRY has a diverse culture and a fascinating history.
COUNTRY has been involved in a number of historical events.
COUNTRY is developing its economic sector rapidly.
COUNTRY fought in a number of wars.
Today my history teacher taught us about COUNTRY and its history.
The geography of COUNTRY is fascinating.
A number of scientists from COUNTRY have gained fame for their work.
Living in COUNTRY definitely has its advantages and disadvantages.
The government of COUNTRY is facing criticism.
I never thought to visit COUNTRY until my neighbor told me about it.
The news says that COUNTRY is going through some severe climate change.
The athlete from COUNTRY has just won the Olympic medal.
The actress was born in COUNTRY and immigrated as a kid.
A number of fossils has been found in COUNTRY where scientists least expected.

Table 6: List of artificial sentences used in section 4 to measure cosine similarity of country names in identical contexts.

GDP Quartile	BERT Base	BERT ML	RoB-ERTa	Tuned
1	26%	15%	14%	**28%**
2	26%	18%	22%	**30%**
3	**32%**	22%	26%	30%
4	**42%**	29%	39%	38%

Table 7: Performance on the MLM task as binned by gold label's GDP (by quartiles) from section 5. We see that performance is best on the high GDP countries and that our tuned model is able to perform better on the lower GDP countries which are in our set of interested countries (Table 8).

Country	GDP (millions)	Freq
India	2,891,582	505,003
Iran	603,779	160,699
Poland	595,862	197,231
Egypt	317,359	101,897
Qatar	183,466	22,766
Angola	85,000	19,839
Myanmar	76,784	23,840
Uzbekistan	57,921	13,244
Serbia	51,475	57,957
Uganda	32,609	33,860
Cambodia	27,097	22,095
Iceland	24,188	31,763
Senegal	23,664	14,672
Syria	20,379	53,987
Jamaica	15,830	40,257
Madagascar	14,104	23,512
Bahamas	13,578	13,882
Guinea	12,354	54,010
Chad	11,271	38,650
Barbados	5,209	13,829

Table 8: List of twenty randomly selected countries from section 6. Additional data for each of these countries was used to continue pre-training the model — resulting in an increased in performance on this subset of countries.

	BERT-Base	Tuned
Accuracy		
Interested Countries	31.75%	44.55%
Other Countries	31.56%	28.89%
% of predictions		
Interested Countries	17.62%	44.44%
Other Countries	82.37%	55.56%

Table 9: Average accuracy MLM task of BERT-Base and our tuned model for our random set interested countries and all other countries. Average % of predictions out of all countries predicted (non-countries words are not in denominator).
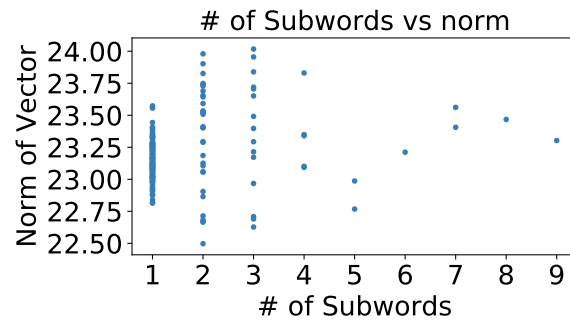
Figure 4: Average norm of embeddings in relation to the number of subpieces of the embedding. OOV words are represented by the first subword of the country name. Pearson's correlation between average norm and number of subwords, $R = 0.22$.
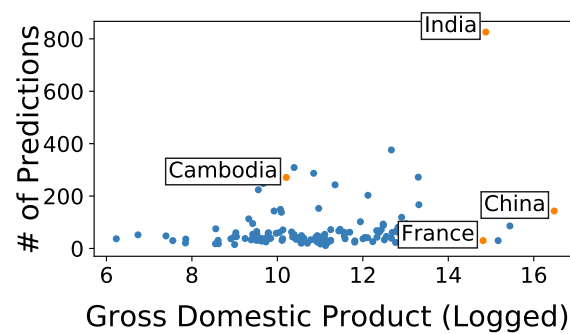

Figure 5: Number of times a country was predicted in the MLM task versus its GDP (logged) in our tuned model. Pearson's correlation between GDP (logged) and number of times a country is predicted for our tuned mode, $R = 0.22$.