Quantifying Gender Bias Over Time Using Dynamic Word Embeddings

Aodong Li
UC Irvine
aodongl1@uci.edu

Robert Bamler
UC Irvine
rbamler@uci.edu

Stephan Mandt
UC Irvine
mandt@uci.edu

Abstract

Dynamic word embeddings [1] are a powerful tool to measure the evolution of word semantics over time, but have not been exploited to-date due to a lack of available software implementations. In this work, we utilized them to quantify gender bias over time. By identifying a *gender direction*, we find that certain words dramatically change their orientation along this direction in the 1960s, which could be attributed to the Women's Rights Movement in these years. We specifically demonstrate shifts of gender bias over time in three corpora, proving the versatility of dynamic word embeddings as a tool for the social sciences and humanities.

Introduction. Gender bias in word embeddings may harm their application by contributing to a feedback loop [2] that perpetuates prejudice. Drawing the connection with stereotypes in social science, measuring the temporal changes of gender bias is a meaningful task [3]. Garg et al. [3] described the bias over time resorting to other gender-neutral concepts, e.g., occupations and adjectives. These analyses are built on temporally independent word embeddings aligned by transformations [4], which suffers from either coarse temporal scale or noisy perturbations [5].

Dynamic word embeddings [1], by imposing a temporal prior on the low-dimensional space, is a powerful approach to quantifying *smooth* evolutions of text corpora. However this model has not been well investigated in practice. In this work, we use it to investigate temporal changes of gender bias, drawing an analysis by Bolukbasi et al. [2].

Dynamic word embeddings. Word2vec [6, 7, 8] is a set of unsupervised learning algorithms to cluster words with similar usage as measured on the contexts in which they appear. Among these methods, the Skip-gram model [6] learns word embeddings by maximizing the log-likelihood of a coocurrence matrix n^+ and a negative sampling matrix n^-,

$$\log p(n^{\pm}|U,V) = \sum_{i,j=1}^{L} \left(n_{ij}^{+} \log \sigma(u_i^\top v_j) + n_{ij}^{-} \log \sigma(-u_i^\top v_j) \right). \tag{1}$$

Here, L is the vocabulary size and $U \equiv (u_i)_{i=1}^L$ and $V \equiv (v_j)_{j=1}^L$ are collections of embedding vectors for words i and contexts j, respectively. Barkan [9] proposed a Bayesian variant of word2vec by imposing Gaussian priors. Further using a Gaussian process prior for the time evolution of embeddings, Bamler and Mandt [1] developed a probabilisitc temporal variant termed Dynamic Word Embeddings. Adding a time dimension $t \in \{1, \dots, T\}$ to $n^{\pm}$, U, and V, the joint distribution is

$$p(n^{\pm}, U, V) = \left(p(U_1)p(V_1) \prod_{t=2}^{T} p(U_t|U_{t-1})p(V_t|V_{t-1}) \right) \left(\prod_{t=1}^{T} p(n_t^{\pm}|U_t, V_t) \right), \tag{2}$$

where the first brackets contain priors and $p(n_t^{\pm}|U_t, V_t)$ is the likelihood for timestep t (see Eq. 1).

The Dynamic Word Embeddings model, Eq. 2, embeds all words from all time steps in a joint embedding space. This makes it possible to draw smooth trajectories of word embeddings over time.

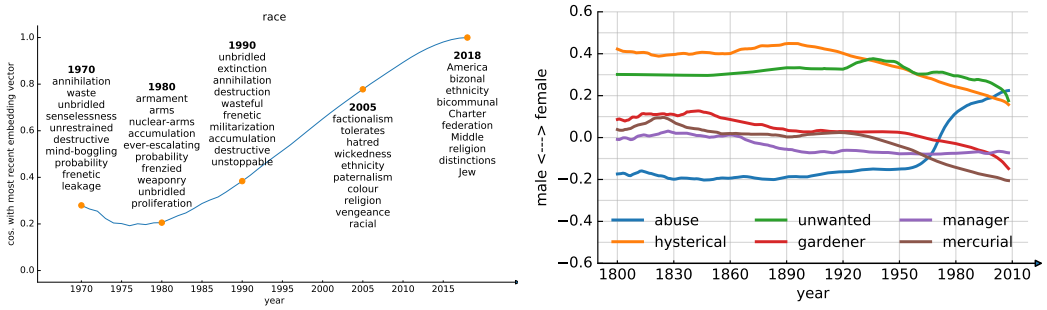

Figure 1: (left) Trajectory of the word 'race' from UN general debates corpus. The word's usage during the arms race of the cold war era differed from its contemporary focus on ethnicity. (right) Gender bias of word samples measured by the most recent gender direction from Google books.

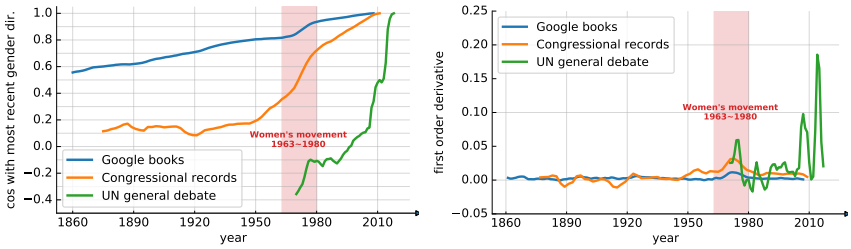

Figure 2: (left) Trajectories of gender direction for our three corpora. (right) Estimated derivative (central difference) of the trajectories. All corpora consistently yield more shifts during the Women's Rights Movement (red region).

For example, we can plot the temporal trajectory of the word 'race' by drawing the cosine distance $u_t^\top u_T / (||u_t|| \, ||u_T||)$ between its embedding u_t at different times $t \in \{1, \ldots, T\}$ and the most recent one, u_T. This leads to the smooth trajectory in Figure 1 (left). The figure also shows the 10 nearest neighboring words at 10-year intervals, which show that the dominant meaning shifts over time. In general, a change of the trajectory correspond to a change in word meaning.

To infer such posteriors, we use black-box variational inference [10, 11], an efficient approximate Bayesian inference method. In this work, we use the Skip-gram smoothing algorithm with augmented symmetry breaking optimization [12]. For further details we refer readers to [1].

Datasets. In the experiments, we explore three temporal corpora available online: Google books [13], Congressional records [14], and United Nations General Debates [15]. Google books is the largest corpus, with about 100M tokens per year from 1800 to 2008; Congressional records is smaller, with 13M to 52M tokens per two-year period from 1875 to 2011; the UN General Debates corpus has about 250k to 450k tokens per year from 1970 to 2018.

Gender bias. We follow the definition of a gender direction in [2] using the collection of gender pairs from [3]. For each gender pair (such as 'woman : man' or 'she : he'), we take the difference vector between the word embedding (posterior mode) for the male and the female word. The gender direction is the first principal component of a PCA over all these difference vectors. This definition results in a gender direction for each time step, allowing us to study both the evolution of gender bias of specific words over time as well as the evolution of the notion of gender itself.

Figure 1 (right) shows an example of our results. We systematically selected six words for demonstration from auto-generated lists of 'most biased words of all time' and 'most bias-shifted words from 1820 to 2008'. The figure shows that most words undergo a shift during the Women's Rights Movement (1963-1980). We further analyze the evolution of the gender concept itself by plotting the trajectory (Figure 2 (left)) in the similar fashion with Figure 1 (left). We found that all three corpora have a pronounced shift during Women's Rights Movement. The first order derivative of those trajectories (Figure 2 (right)) further confirms this observation.

References

[1] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org, 2017.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

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

[4] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, 2016.

[5] Syrielle Montariol and Alexandre Allauzen. Empirical study of diachronic word embeddings for scarce data. *arXiv preprint arXiv:1909.01863*, 2019.

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

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[9] Oren Barkan. Bayesian neural word embedding. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] Robert Bamler and Stephan Mandt. Improving optimization for models with continuous symmetry breaking. In *International Conference on Machine Learning*, pages 423–432, 2018.

[13] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

[14] Matthew Gentzkow, JM Shapiro, and Matt Taddy. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. In *URL: https://data. stanford. edu/congress text*, 2018.

[15] Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. United Nations General Debate Corpus, 2017.