Interpretable Word Embeddings via Informative Priors

Miriam Hurtado Bodell[*]
Linköping University
miriam.hurtado.bodell@liu.se

Martin Arvidsson[*]
Linköping University
martin.arvidsson@liu.se

Måns Magnusson
Aalto University
mans.magnusson@aalto.fi

Abstract

Word embeddings have demonstrated strong performance on NLP tasks. However, lack of interpretability and the unsupervised nature of word embeddings have limited their use within computational social science and digital humanities. We propose the use of informative priors to create interpretable and domain-informed dimensions for probabilistic word embeddings. Experimental results show that sensible priors can capture latent semantic concepts better than or on-par with the current state of the art, while retaining the simplicity and generalizability of using priors.

1 Introduction

Increased availability of large digitized corpora and significant developments in natural language processing (NLP) have sparked a growing interest within computational social science and digital humanities (CSSDH) to use computational methods for textual data (Laver et al., 2003; Grimmer, 2010; DiMaggio et al., 2013; Jockers and Mimno, 2013; Tsur et al., 2015). Word embeddings, a family of unsupervised methods for representing words as dense vectors (Mikolov et al., 2013b; Pennington et al., 2014), are one such development. Although word embeddings have demonstrated strong performance on NLP tasks (Mikolov et al., 2013a,c), they have yet to gain widespread attention within CSSDH.

We believe two key limitations can help explain the lack of applications within CSSDH. First, since the dimensions of the word embeddings are largely uninterpretable, it is not clear how to disentangle why words are similar. Substantive interpretability is key for CSSDH research, and thus, the lack thereof is a major limitation. Second, off-the-shelf word embedding models generally lack a channel through which substantive research questions can be incorporated.

To improve interpretability, previous research suggests using sparsity constraints (Murphy et al., 2012; Sun et al., 2016; Faruqui et al., 2015) and rotation techniques (Park et al., 2017; Rothe and Schütze, 2016; Dufter and Schütze, 2019). Other work considers dimension-specific constraints to remove gender-bias, and, as a by-product, improve interpretability (Zhao et al., 2018). Within CSSDH, previous work derives interpretable dimensions via post-processing in the form of antonym-pair vector algebra (Kozlowski et al., 2018; Garg et al., 2018) and ideal-point anchoring of antonym word-pairs (Lauretig, 2019). Recent formulations of word embeddings as probabilistic models (Vilnis and McCallum, 2015; Rudolph et al., 2016; Barkan, 2017; Havrylov et al., 2018) enable the incorporation of domain knowledge through priors. In this paper, we add to the literature on interpretable word embeddings, proposing a novel use of *informative priors* to create pre-defined interpretable dimensions – thus leveraging the expressiveness and generalizeability of the probabilistic framework.

2 Informative Priors in Word Embeddings

The central idea of this paper is to use informative priors to restrict the degree to which different words can inhabit different dimensions, such that one or more dimensions become interpretable and connected to one's research interest. Specifically, we place informative priors on word types that we expect to discriminate on a particular dimension, e.g. *man-woman* for a gender dimension.

Let $\mathcal{V}_+$ and $\mathcal{V}_-$ be the set of *anchor word types* that informative priors is placed on, e.g. $\mathcal{V}_+ = \{man\}$ and $\mathcal{V}_- = \{woman\}$. Also, let $\mathcal{V}_\pm = \mathcal{V}_+ \cup \mathcal{V}_-$, and let $\overline{\mathcal{V}}_\pm$ be the word types without any informative priors, i.e. the complement set.

Given a corpora with vocabulary size V, we represent each token $x_i \in \{0, 1\}^V$ as a one-hot vector

---

[*] Equal contribution

6323

with a single nonzero entry at v. Here, $v \in \mathcal{V}$ represents the word type at position i in the text. Following Rudolph et al. (2016), we model each individual entry $x_{iv} \in \{0, 1\}$ of the one-hot vectors conditional on its context $\mathbf{x}_{c_i}$, i.e the tokens surrounding x_i, where c_i denotes the positions belonging to the context.

Each word type is associated with an *embedding vector*, $\rho_v \in \mathbb{R}^K$, which governs the distribution of x_i, and a *context vector*, $\alpha_v \in \mathbb{R}^K$, governing the distributions of the tokens for which x_i is part of $\mathbf{x}_{c_i}$. The conditional probability of x_{iv} is modelled as a linear combination of the embedding and context vectors, i.e.

$$p(x_{iv}|\mathbf{x}_{c_i}) \quad \sim \quad \text{Bernoulli}(\eta_{iv}), \qquad (1)$$

where

$$\eta_{iv} \quad = \quad \text{logit}^{-1}[\rho_v^\mathsf{T}(\textstyle\sum_{j \in c_i} \sum_{v'} \alpha_{jv'} x_{jv'})],$$

with logit serving as the link function. Rudolph et al. (2016) place a zero-centered Gaussian prior with variance σ on the embedding and context vectors. Letting $\theta_v = \{\rho_v, \alpha_v\}$, this translates to

$$\theta_v \sim \mathcal{N}(0, \sigma). \qquad (2)$$

In addition, since evolution of semantic concepts are of special interest in CSSDH (Tahmasebi et al., 2015), we also consider a dynamic word embedding model to capture temporal dynamics in the dimension of interest. We follow the specification in Rudolph and Blei (2017) which extends Eq. (1) by associating each token with a time slice t, and fit separate $\rho_v^{(t)} \in \mathbb{R}^K$ for each t. Thus, $\theta_v^{(t)} = \{\rho_v^{(t)}, \alpha_v\}$. To share statistical strength between time-points, a Gaussian random walk prior is placed on $\rho_v^{(t)}$, i.e.

$$\rho_v^{(t)} \sim \mathcal{N}(\rho_v^{(t-1)}, \sigma_d I). \qquad (3)$$

Where $\sigma_d = \frac{\sigma}{100}$, as in Rudolph and Blei (2017), determines the smoothness of the trajectories. This shows how, in contrast to the state of the art (Kozlowski et al., 2018; Garg et al., 2018; Lauretig, 2019; Zhao et al., 2018), informative priors allow easy integration with other, more complex, probabilistic models.

2.1 The Standard Basis Prior

In the following sections, we introduce a number of prior specifications that differ in how they restrict the degree to which words can occupy different dimensions. Letting K represent the dimension that we want to make interpretable, and

dimensions $1 : K - 1$ be standard word embedding dimensions not subject to interpretation, we define our first prior specification, the *Standard Basis Prior*, as

$$\begin{aligned} \theta_{\overline{\mathcal{V}}_\pm}^{1:K} &\sim \mathcal{N}(0, \sigma) & \theta_{\mathcal{V}_+}^{K} &\sim \mathcal{N}(1, \gamma) \\ \theta_{\mathcal{V}_\pm}^{1:K-1} &\sim \mathcal{N}(0, \omega) & \theta_{\mathcal{V}_-}^{K} &\sim \mathcal{N}(-1, \gamma), \end{aligned} \qquad (4)$$

where $\theta_{\mathcal{V}_+}^{K}$ and $\theta_{\mathcal{V}_-}^{K}$ are priors on the dimension of interest (K) of ρ_v and α_v for word types in $\mathcal{V}_+$ and $\mathcal{V}_-$ respectively, and where $\theta_{\mathcal{V}_\pm}^{1:K-1}$ is the shared prior for all anchor word types on dimensions $1 : K - 1$. Finally, $\theta_{\overline{\mathcal{V}}_\pm}^{1:K}$ is the standard prior from Eq. (2), placed on all dimensions for all non-anchor word types. Hyperparameters ω and γ are shared for $v \in \mathcal{V}_\pm$, controlling the strength of the prior. As $\gamma, \omega \to 0$, we force $v \in \mathcal{V}_\pm$ to essentially become a standard basis, defined by the word types in the prior. Consequently, the dot product for these vectors will be 0 for all dimensions except K, and thus the effect of the anchored word types on the rest of the vocabulary will exclusively depend on K.

However, this implies that, as $\gamma, \omega \to 0$, word types within $\mathcal{V}_-$ and $\mathcal{V}_+$ obtain exactly the same word embedding. For example, with $\mathcal{V}_+ = \{brother, king\}$, *brother* is treated as semantically identical to *king*. To address this issue, we consider increasing ω, allowing the anchor word types to exist more freely in the first $K - 1$ dimensions while remaining close to -1 and 1 in the Kth dimension. This permits the use of both $brother$ and $king$ as prior anchors without assuming that they are exactly the same word. We henceforth refer to a standard basis prior with $\omega = 10^{-6}$ as *strict* and $\omega = 1$ as *weak*.

2.2 The Truncated Prior

A limitation of the prior specifications introduced thus far is the implicit assumptions that (i) anchor word types discriminate equally on the dimension of interest, and (ii) that we know their exact location in this dimension. To address this, we consider the *Truncated Prior* that does not assume a basis for the anchored word types, but only that they live on the positive and negative real line as

$$\begin{aligned} \theta_{\overline{\mathcal{V}}_\pm}^{1:K} &\sim \mathcal{N}(0, \sigma) & \theta_{\mathcal{V}_+}^{K} &\sim \mathcal{N}^+(0, \gamma) \\ \theta_{\mathcal{V}_\pm}^{1:K-1} &\sim \mathcal{N}(0, \omega) & \theta_{\mathcal{V}_-}^{K} &\sim \mathcal{N}^-(0, \gamma). \end{aligned} \qquad (5)$$

Where $\mathcal{N}^+(0, 1)$ and $\mathcal{N}^-(0, 1)$ is the positive and negative truncated normal distribution, truncated at 0, respectively.

2.3 Using Neutral Words

So far, we have only considered word types known to discriminate on the dimension of interest as prior anchors. However, domain knowledge might also inform us about *neutral* words in this dimension (e.g. stop words). We thus consider placing informative priors on a third set of prior anchors $\mathcal{V}_*$ containing neutral word types as

$$
\begin{aligned}
\theta_{\mathcal{V}_*}^{1:K-1} &\sim \mathcal{N}(0, \sigma) \\
\theta_{\mathcal{V}_*}^{K} &\sim \mathcal{N}(0, \psi),
\end{aligned} \tag{6}
$$

where ψ is the variance for dimension K. By guiding neutral words close to 0 in the dimension of interest, explanatory power that otherwise might have been attributed to word types in $\mathcal{V}_*$ will instead be allocated to other words in $\mathbf{x}_{c_i}$.

3 Experiments

Our main empirical concern is how well the proposed priors can capture meaningful latent dimensions. We summarize our empirical questions as follows:

1. Which prior specification can best capture predefined dimensions?

2. How does the best prior specification compare with the state of the art in CSSDH (Garg et al., 2018; Kozlowski et al., 2018) (henceforth referred to as SOTA)?

We consider two semantic dimensions, *gender*, which is explored in SOTA, and *sentiment*, a dimension proven difficult to capture in standard word embedding models (Tang et al., 2014). We follow SOTA when choosing prior anchors for gender, while using the AFINN dictionary (Nielsen, 2011) to find prior anchors for sentiment. To evaluate the effect of "few" vs. "many" prior anchors, we run experiments using between 2 and 276 words depending on the dimension of interest and the dataset at hand. All prior word types used in the experiments can be found in Sec. B in the supplementary material. We follow Rudolph et al. (2016), obtaining maximum a posteriori estimates of the parameters using TensorFlow (Abadi et al., 2015) with the Adam optimizer (Kingma and Ba, 2015) and negative sampling [1]. We set the size of the embeddings $K = 100$, use a context window size of 8 and $\sigma = 1$ throughout all experiments.

---

[1] Code is available at: https://github.com/martin-arvidsson/InterpretableWordEmbeddings

We examine the proposed priors using three commonly sized English corpora for textual analysis within CSSDH: the top 100 list of books in Project Gutenberg (2019), a sample from Twitter (Go et al., 2009) and the U.S. Congress Speeches 1981-2016 (Gentzkow et al., 2018). The number of tokens ranges from 1.8M to 40M after preprocessing (see Sec. A in the supplementary material for details). The various origins, sizes and contents of these datasets work as a check of the effect of the priors in different types of corpora.

To measure the extent to which the inferred dimension reflects the semantic concept of interest, we consider how well-placed a number of pre-specified hold-out word types (not a part of $\mathcal{V}_\pm$) are in this particular dimension. Specifically, accuracy is computed as the fraction of hold-out words that are placed on the correct side of 0 on the dimension of interest in the embedding and context vectors. For the gender dimension, hold-out word types include the 200 most common names of the last century (Social Security Administration, 2019) and gendered words, while for sentiment, a sample of the AFINN dictionary is used (see Sec. C in the supplementary material). The number of hold-out word types ranges between 213 and 275, since not all exist in each corpus.

The experimental configurations are compared to SOTA, which we implement by (i) fitting the probabilistic word embedding model in Eq. (1) and (2) without informative priors, and (ii) deriving the interpretable dimension post-hoc by subtracting normalized embedding vectors of antonym-pairs, e.g. $\rho_{gender} = (\rho_{man} - \rho_{woman}) + (\rho_{he} - \rho_{she})$. Comparing the sign of the cosine similarity between hold-out words and the created vector allows us to contrast the accuracy of our method with that of SOTA. To get a measure of uncertainty we calculate binomial confidence intervals using the normal approximation.

4 Results

We begin by comparing the strict and weak *Standard Basis Priors*, with $\gamma = 10^{-6}$, varying the number of anchor words used to inform the dimension of interest. The general pattern in Fig. 1 shows how (i) increasing the number of anchor words improves the accuracy, and that (ii) this improvement is much greater for the weak than for the strict *Standard Basis Prior*. We explain this
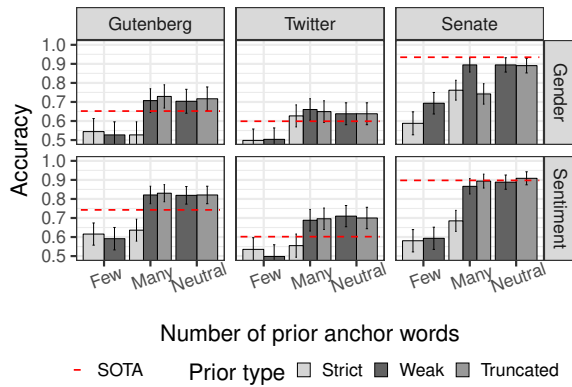
Figure 1: Comparison of prior specifications per dataset and dimension

difference by the nature of the strict prior; it forces all anchor words to have exactly the same meaning – which clearly is untrue. These results speak against methods that transform the vector space based on strict standard basis vectors, as in Lauretig (2019). Further noticeable is that the average accuracy is greater for the Senate corpus, which, using subsampling, we find is driven by corpora size.

Relaxing the location assumptions with the *Truncated Prior* ($\gamma = 1000$) and placing informative priors on neutral words ($\psi = 0.01$) show varying degrees of improvements. However, due to the limited number of hold-out test words, large improvements are required for significant differences – which only is observed for the sentiment dimension in the Senate corpus.

Fig. 1 further shows how our proposed approach generally performs better or on-par with SOTA. The one exception is the gender dimension in the Senate corpus. Follow-up analysis show that this difference is driven by misclassification of a cluster of ambiguous names, e.g. *Madison* (founding father) and *Jordan* (country). These names become correctly classified when using the full vectors, suggesting that gender has not been completely isolated in the Senate corpus (see Gonen and Goldberg (2019) for an in-depth discussion on issues of concept-isolation in word embeddings).

4.1 Case Study of Semantic Change

Using the *Truncated Prior* with neutral words, we leverage the dynamic embedding model described in Eq. (3) to explore temporal patterns in the Senate corpus. We set $\sigma_d = 0.05$ to allow for the identification of abrupt temporal changes.

The left panel of Fig. 2 displays sentiment-trajectories for two words experiencing drastic three-year consecutive changes in the sentiment dimension; *September* between 1999 and 2001 and *Oklahoma* between 1993 and 1995, capturing two terror events: the attacks on the World Trade Center on September 11th 2001, and the Oklahoma City bombing in 1995. The change precedes the event due to the smoothness of the prior. In the years that follow, the words gradually regain their previous sentiment – reflecting a decline in their association with terror.

Second, we test and find support for the gender-occupation results in Kozlowski et al. (2018), i.e. that occupations' temporal positioning in the gender dimension correlates well with the proportion of men and women within those fields. The right panel of Fig. 2 displays the gender-dimension trajectories for the words *Nurse* and *Laywer*, occupations with high shares of women and men, respectively (Kozlowski et al., 2018).

In sum, these examples showcase how interpretability can be gained using informative priors. By isolating prespecified concepts, e.g. sentiment, into one dimension (K), one can infer word type-concept associations such as *September-sentiment* – for all non-anchor word types – allowing for more meaningful within and between word type comparisons.
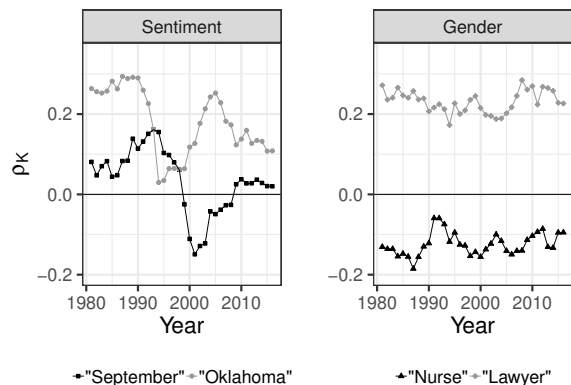

Figure 2: Gender and sentiment trajectories in the Senate corpus.

5 Conclusion

In this paper, we show how informative priors provide a simple and straightforward alternative for constructing interpretable dimensions for probabilistic word embeddings – allowing CSSDH researchers to explore how words relate to each

other in prespecified dimensions of interest, e.g. gender and sentiment.

Our results demonstrate that the biggest gains in interpretability are obtained by (i) using many prior words and (ii) increasing the degree to which they can live outside the predefined dimension. The weak *Standard Basis Prior* and the *Truncated Prior* overall capture the predefined dimensions with similar accuracy. Aligned with previous research, the experimental results indicate some issues with incomplete isolation of concepts which we believe could be addressed in future work by placing informative priors on multiple dimension to capture more complex concepts, and move away from simplistic antonym-driven dimension definitions.

Finally, while being flexible and easily extended to other probabilistic word embedding models, our approach performs better or on par with SOTA.

Acknowledgement

References

Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2015. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org*, 1(2).

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

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.

Philipp Dufter and Hinrich Schütze. 2019. Analytical methods for interpretable ultradense word embeddings. *arXiv preprint arXiv:1904.08654*.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.

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.

M. Gentzkow, J. M. Shapiro, and M. Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. http://data.stanford.edu/congress_text. Accessed: 2019-04-17.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Serhii Havrylov, Ivan Titov, et al. 2018. Embedding words as distributions with a bayesian skipgram model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1775–1789.

Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Adam Lauretig. 2019. Identification, interpretability, and Bayesian word embeddings. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 7–17, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. *Proceedings of COLING 2012*, pages 1933–1950.

F. Å. Nielsen. 2011. AFINN. http://localhost/pubdb/p.php?6010.

Sungjoon Park, JinYeong Bak, and Alice Oh. 2017. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 401–411.

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

Project Gutenberg. 2019. http://www.gutenberg.org. Accessed: 2019-04-24.

Sascha Rothe and Hinrich Schütze. 2016. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–517.

Maja Rudolph and David Blei. 2017. Dynamic bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.

Social Security Administration. 2019. Top names over the last 100 years. https://www.ssa.gov/oact/babynames/decades/century.html. Accessed: 2019-05-02.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Sparse word embeddings using l1 regularized online learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2915–2921. AAAI Press.

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik D Johansson, Richard Johansson, Mikael Kågebäck, et al. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4):169–187.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1629–1638.

Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.