Exploring Differences in Semantics in the Language of Online Communities

Jina Suh, Emil Azadian, and Galen Weld

February 1, 2019

Introduction

Word embedding is a natural language processing (NLP) technique that maps words (or phrases, subwords, or even characters) to vectors. Availability of toolkits and pre-trained embeddings(e.g., word2vec [15], GloVe [17]) have popularized the use of the embeddings in many downstream tasks such as text sentiment classification, part-of-speech tagging, and name-entity recognition. One of the main strengths of word embeddings is that it captures semantic relationships between words well. For example, given the vectors for "king", "man" and "woman", the result of "king - man + woman $\approx$ queen" [14].

Word embeddings are also known to capture associations that may reflect stereotypes present in the training corpus [2, 3] to produce analogies such as "computer programmer - man + woman = homemaker." Such undesirable associations can be shown to propagate to downstream tasks[1]. Easily accessible pre-trained embeddings are used already in many existing tasks, such as sentiment analysis; 75% to 85% of the 219 sentiment analysis systems reviewed by Kiritchenko and Mohammad consistently predicted sentences related to one gender differently from those to the other gender [10] Some have proposed methods to "fix" the word embeddings [2] while others have argued against it [21]. It seems more likely that taking off-the-shelf word embeddings or irresponsibly relying on embedding techniques without inspection or mitigation could lead to human harm.

While these stereotypical associations might be undesired in some contexts, they could yield an excellent tool for inspecting existing semantic associations within online communities, since words often have different connotations in different contexts, depending on the community using them. For example, Gab is an online social network that is known to attract alt-right users and conspiracy theorists [24]. An analysis of comments on two social media platforms, Gab and Reddit, found dramatically different words to be used in conjunction with the word 'oven,' displayed in Figure 1. Echo chamber of different kinds also exist within certain Reddit subgroups. Since these groups attract distinct user bases, opinions, and topics, the language and use of words within these groups might significantly differ from that of Wikipedia.

Figure 1: Word clouds produced from comments on Gab, on the left, and Reddit, on the right. [1]

Problem Statement

Word embeddings can be used to capture how these same words are used differently across different online communities. Our goal for this project is to use word embedding models trained on different corpora to compare word usage and to implement a visualization tool for evaluating the results in order to answer our research question:

How does language usage vary across online communities, especially when we look at words in potential bias-categories?

[1] https://gist.github.com/rspeer/ef750e7e407e04894cb3b78a82d66aed

Related Work

Unwanted Associations in Word Embeddings

Some work has been done on methods for detecting, and potentially reducing, learned bias in word embeddings. Caliskan et al. show that bias in embeddings mirrors the human biases present in the corpora the embeddings are trained upon [3]. Bolukbasi et al. investigate female/male stereotypes in the word2vec embedding trained on the Google News corpus, and propose a debiasing framework [2]. Swinger et al. demonstrate a highly-unsupervised bias detection system on a number of publicly available embeddings [23].

Semantic Differences

Words have different meanings in different contexts or times, and word embeddings could be used to capture the dominant meaning of the word in the training corpus. Hamilton et al. [8] have looked at historical semantic evolution; by aligning word embeddings across different time periods, the authors were able to visualize the change paths in context of other words (see Figure 2). Rather than looking at how semantics evolve over time, our project will focus on how semantics of everyday words differ across online communities.

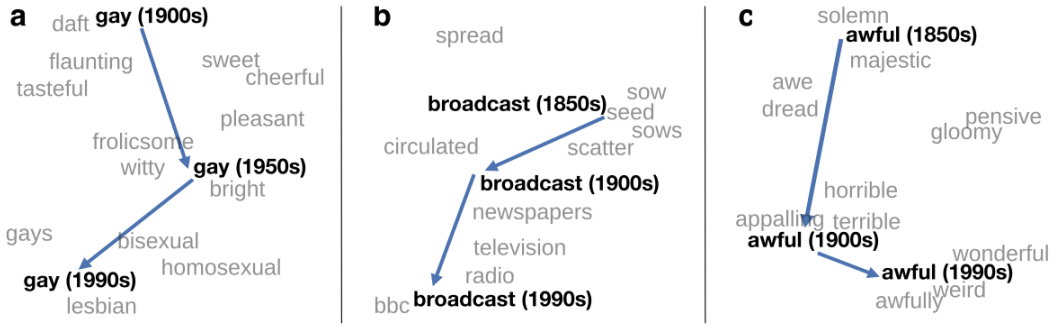

Figure 2: Changes in word meaning visualized by projecting historical word vectors into a 2-D space from [8]

Visualization and Interaction Support

Interactive visualizations and interfaces have been useful in analyzing word embedding models because the nature of the problems that word embeddings address are human-centric and benefit largely from domain expertise or human supervision in the analysis process [9]. There have been advances in algorithmic techniques for projecting high-dimensional embeddings into 2D space (e.g., tSNE [12]) with associated visualization support [22][2]. Some interactive tools are designed to facilitate exploring and inspecting word embeddings alone (e.g., [11, 19, 20]) while others support a related user task (e.g., interactive lexicon building [16]). Heimerl and Gleicher [9] summarize linguistic tasks that word embeddings employ and the role of visualization in support of those tasks. Our project will focus on defining and comparing concepts (e.g., gender, race) and projecting words along concept axes using radial visualization technique [4] across multiple word embeddings. This visualization is in turn inspired by existing tools for visualization gender bias in word embeddings [6, 13].

Datasets

For an insightful comparison of language differences across online communities, we must use a diverse set of corpora to train word embeddings on. We propose using the following datasets:

- A dataset of 16 million comments from the `/r/the_donald` subreddit, representing a far-right Reddit community [5]
- A dataset of comments from `/r/LateStageCapitalism`, representing a far-left Reddit community [7]
- A dataset of 34 million posts from the Gab social network, representing an alt-right community [**gab˙dataset**]
- A pretrained GloVe embedding, trained on a 2014 Wikipedia article corpus, representing a 'neutral' language set [18]; or word2vec embedding, trained on Google News dataset [14]

[2]http://projector.tensorflow.org/

Methods

Training

We will train the popular *word2vec continuous skip-gram model* for each of our different corpora, yielding several distinct word embeddings. Following a similar approach in [2], we define a bag of words (BOW) for each bias-category. These words will be chosen category-specific. For example, the bias-category ”Muslim” could have a possible BOW of {’Muslim’,’Hajj’,’Imam’,’Mosque’}.

Vocabulary

NOTE: This section might not be necessary.

Evaluation

We will take several steps in order to assess the quality of our trained embeddings. First, we will utilize a publicly available dataset such as WS353 or MEN to obtain a word similarity score for each WE. Given the task at hand however, this evaluation might not be representative of the quality of our word embeddings. This is due to the fact that our corpora are specifically selected for their supposibly highly biased use of words, which might lead to low similarity scores on standard evaluation sets, albeit being well-fit for the task at hand. We will therefore use a second, different approach for evaluating our networks, using an interactive visualization tool.

Interactive Visualization

Based on the work in [4], we will implement an interactive visualization tool that facilitates error discovery through semantic data exploration. This tool will be additionally used for the main goal of this project: Visualizing differences in bias-categories (BC) across different online communities (see Fig 3). The tool will work as follows: On a circle, three BC will be radially aligned, each represented by a predefined bag of words. Within the circle, words are placed based on differences in word similarity between the bias categories and the word itself. Hence the word will be closer to a BC that it is more similar to. Neutral words, or words that are equally similar to all BC, will end up in the middle. Shown side by side, each trained word embedding will have its own visualization. This allows for a direct comparison of where potentially biased words are contextually placed among different social-media communities.

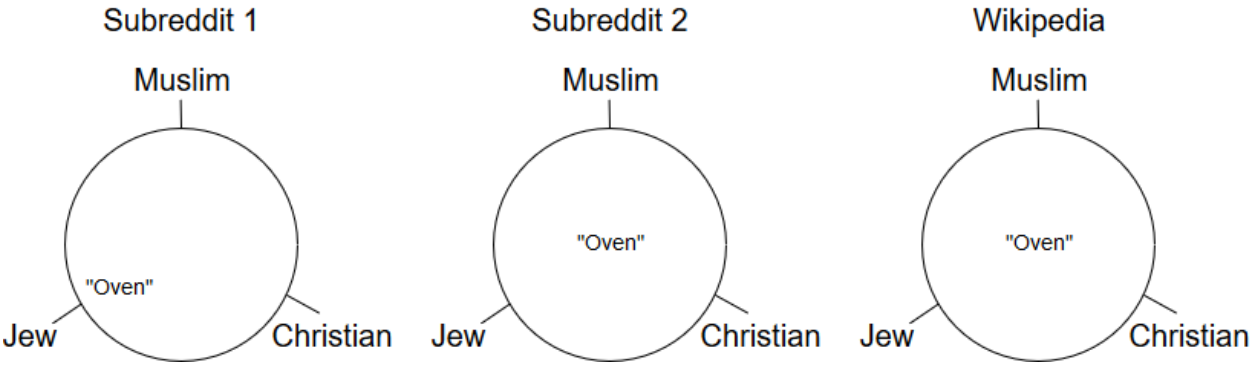

Figure 3: Sketch of a possible visualization. The chosen bias categories are ”Jew”,”Muslim” and ”Christian”, the word for comparison is ”Oven”. In the presumably right-wing dominated subreddit 1, ”Oven” is much more similar to Jew than in other corpora.

User Task

Using the visualization tool, users will be able to specify their own BC or choose from a predefined list we provide. Additionally, they will be able to search words of their interest that will then be highlighted in the circle.

Expected Outcome

We expect to deliver:

- Two or more word embeddings trained on different corpora
- Interactive visualization prototype that supports our said user task
- A set of words and bias-categories
- Demonstrations of different tendencies across online communities

References

[1] Jason Baumgartner. *Two Social Media Platforms' use of Oven*. Nov. 2018. URL: https://twitter.com/jasonbaumgartne/status/1059673359963828225.

[2] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *NIPS*. 2016.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[4] Nan-Chen Chen et al. "AnchorViz: Facilitating Classifier Error Discovery through Interactive Semantic Data Exploration". In: *23rd International Conference on Intelligent User Interfaces*. ACM. 2018, pp. 269–280.

[5] Claudia Flores-Saviaga, Brian C. Keegan, and Saiph Savage. "Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community". In: *CoRR* abs/1806.00429 (2018). arXiv: 1806.00429. URL: http://arxiv.org/abs/1806.00429.

[6] Tom Forth. *Gender Bias Calculator*. Dec. 2018. URL: https://www.tomforth.co.uk/genderbias/.

[7] *Google BigQuery 2015 Reddit Comments Dataset*. URL: https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2015_05?pli=1.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change". In: *arXiv preprint arXiv:1605.09096* (2016).

[9] Florian Heimerl and Michael Gleicher. "Interactive analysis of word vector embeddings". In: *Computer Graphics Forum*. Vol. 37. 3. Wiley Online Library. 2018, pp. 253–265.

[10] Svetlana Kiritchenko and Saif M Mohammad. "Examining gender and race bias in two hundred sentiment analysis systems". In: *arXiv preprint arXiv:1805.04508* (2018).

[11] Shusen Liu et al. "Visual exploration of semantic relationships in neural word embeddings". In: *IEEE transactions on visualization and computer graphics* 24.1 (2018), pp. 553–562.

[12] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[13] University of Maryland Institute for Advanced Computer Studies. *Gender Bias in Word Embeddings*. URL: http://wordbias.umiacs.umd.edu/.

[14] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations". In: *HLT-NAACL*. 2013.

[15] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[16] Deokgun Park et al. "ConceptVector: text visual analytics via interactive lexicon building using word embedding". In: *IEEE transactions on visualization and computer graphics* 24.1 (2018), pp. 361–370.

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

[18] *Pre-trained GloVe Word Embedding, Wikipedia 2014*. URL: https://nlp.stanford.edu/projects/glove/.

[19] Xin Rong. "word2vec parameter learning explained". In: *arXiv preprint arXiv:1411.2738* (2014).

[20] Xin Rong and Eytan Adar. "Visual tools for debugging neural language models". In: *Proceedings of ICML Workshop on Visualization for Deep Learning*. 2016.

[21] Tobias Schnabel et al. "Evaluation methods for unsupervised word embeddings". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 298–307.

[22] Daniel Smilkov et al. "Embedding Projector: Interactive Visualization and Interpretation of Embeddings". In: *CoRR* abs/1611.05469 (2016).

[23] Nathaniel Swinger et al. "What are the biases in my word embedding?" In: *arXiv preprint arXiv:1812.08769* (2018).

[24] Savvas Zannettou et al. "What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?" In: *arXiv preprint arXiv:1802.05287* (2018).