CERIAS

The Center for Education and Research in Information Assurance and Security

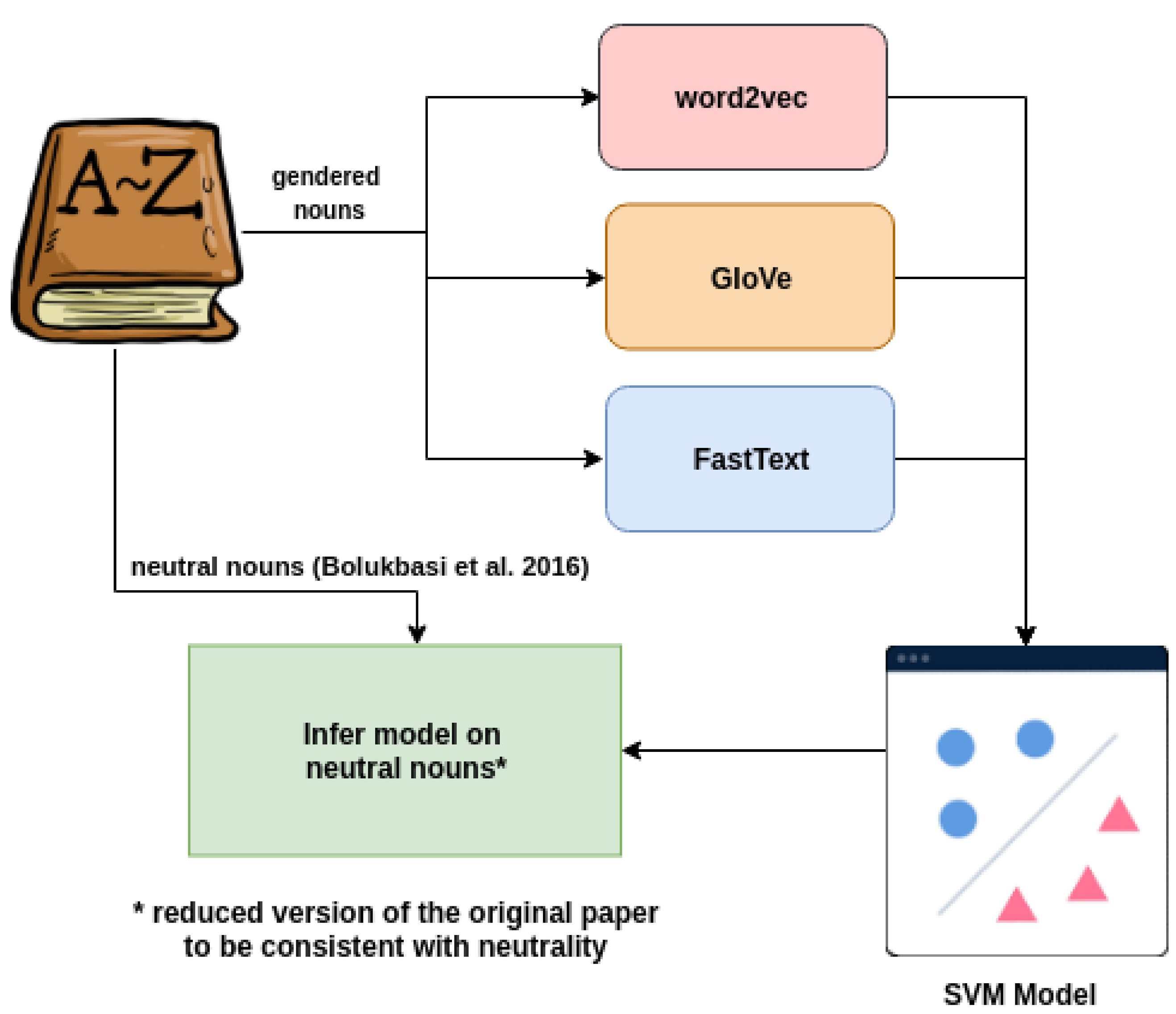
Towards trustworthy NLP systems: detecting bias in popular models

John Phan (phan13@purdue.edu), Kanishka Misra (kmisra@purdue.edu),
Dr. Julia Rayz (jtaylor1@purdue.edu)

Abstract

Recent media coverage of AI applications have shown a growing concern and distrust in the creation of artificial intelligent machines. This work takes a closer look at biases that may be present in word embedding technology. Trustworthy information from these NLP systems is crucial to the security of the applications that utilize word embedding technology. In this project, we aim to study the gender bias implicitly present within three forms of popular pretrained word embeddings – word2vec, GloVe, and FastText. Previous attempts at identifying and eliminating bias have shown to be ineffective, as described in Gonen & Goldberg (2019). SVM is used to determine the two most salient features from the vector space for each set of pretrained word embedding. The classification algorithm is then tasked to predict potential gender of neutral words. Results of our method is consistent with previous work in identifying gender bias within word embeddings using other means, such as from Bolukbasi et al (2016).

Methodology



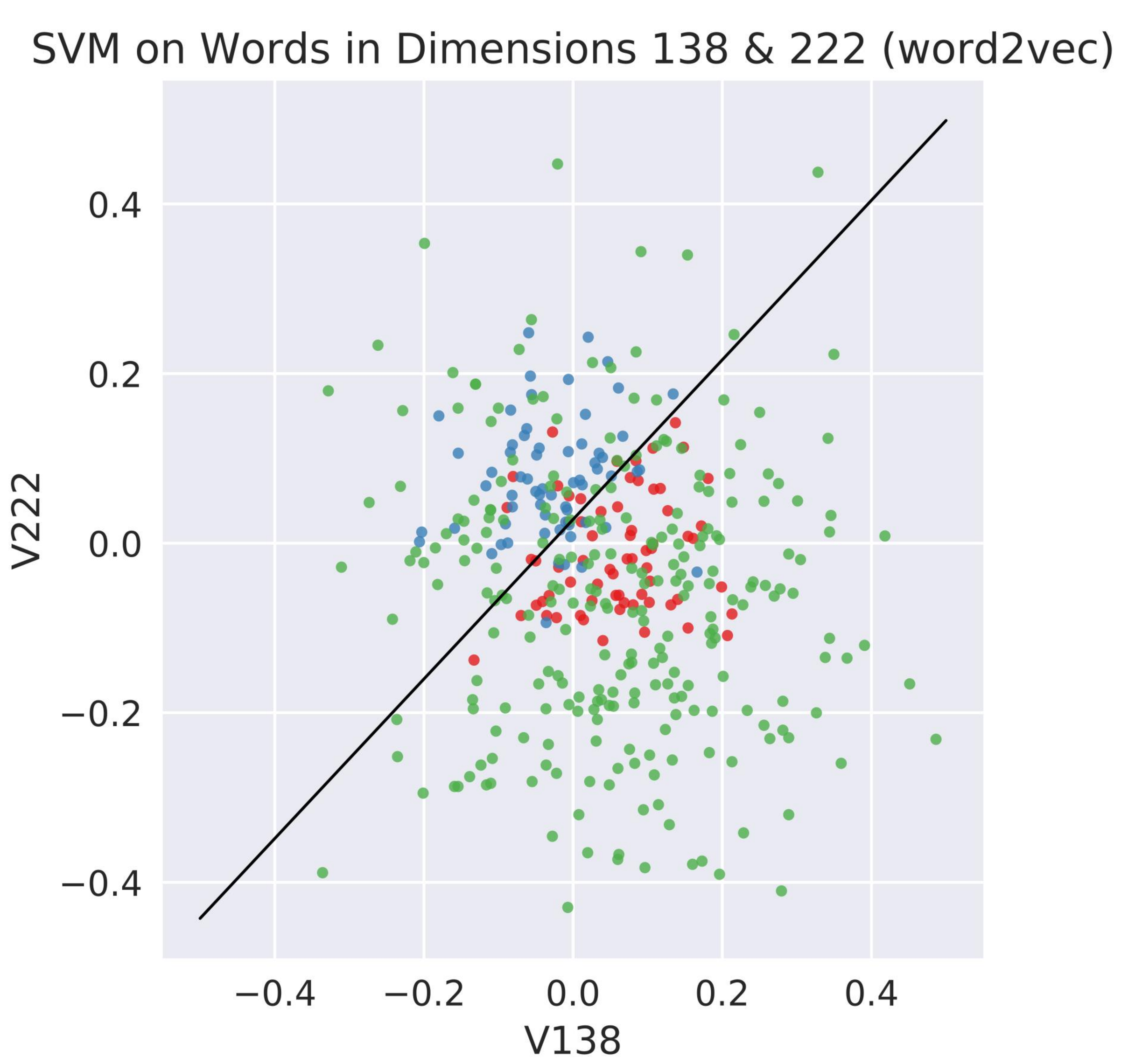
Summary

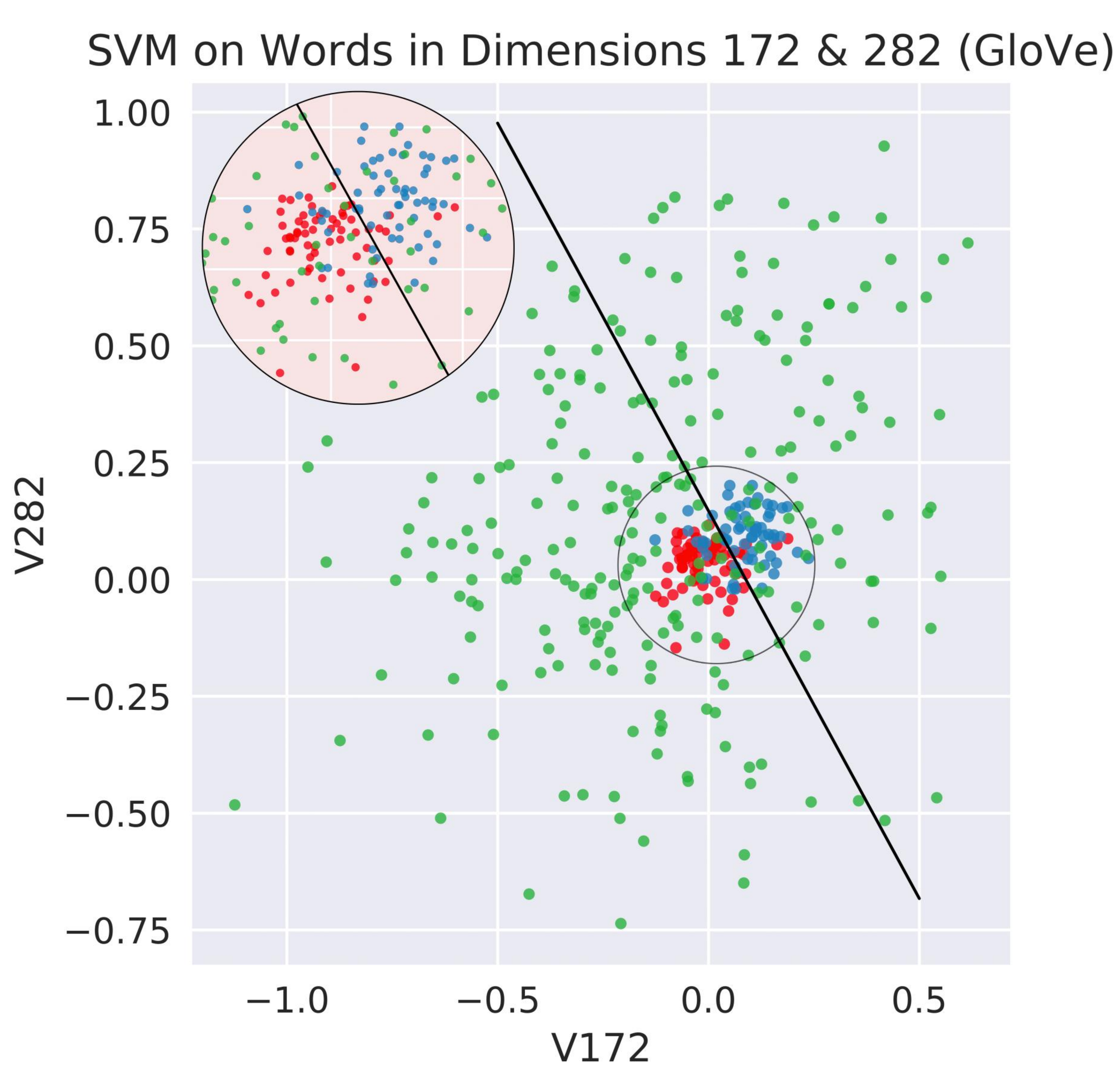
Table 1: SVM Classification Summary of Neutral Words ($n = 250$)

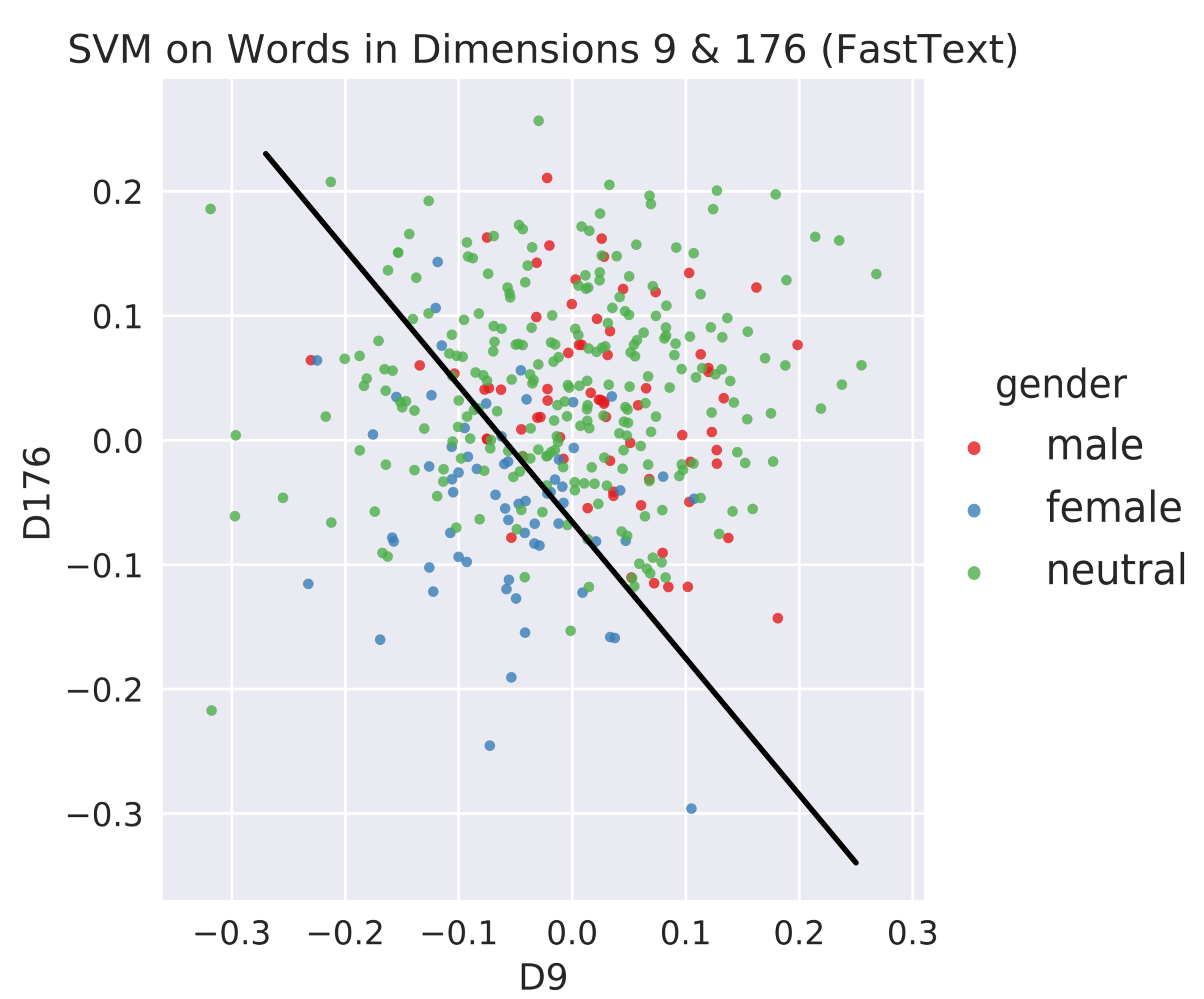
Embedding	As Male %	As Female %
word2vec	75.20%	24.8%
GloVe	64.40%	35.6%
FastText	79.60%	20.4%

- All models performed very differently
- Distribution of male classified neutral nouns is higher than female
- Common & dangerous stereotypes are captured (doctor, programmer, and politican were classified as male across all models)
- Interesting predictions among neutral words - 'person' classified as male across all three models, 'president' classified as female (word2vec)

Results







References

1. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., & Kalai, A.T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS.*
2. Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. Retrieved from http://arxiv.org/abs/1903.03862
3. Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, London: Longman (1968).
4. Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
5. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In NIPS
6. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In EMNLP.
7. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. TACL 5:135–146

Acknowledgements

We would like to express our sincere thanks and gratitude towards the Center for Science of Information – without their aid this project would not have been possible. We would also like to show our appreciation to the members of AKRaNLU for their constant support throughout the term of this project.

PURDUE UNIVERSITY Discovery Park | PURDUE UNIVERSITY | Polytechnic Institute | Center for Science of Information NSF Science and Technology Center | AKRaNLU | CERIAS