A Leader and a Lady? A Computational Approach to Detection of Political Gender Stereotypes in Facebook User Comments

Appendix A

Table 1A

Stereotypical female- and male-linked traits and political traits used in the present study

Trait group	Traits from previous literature	Traits added by the present study
Female traits		
cognitive	artistic, creative, imaginative, intuitive	artsy, innovative, inventive, instinctive
personality	affectionate, gentle, sensitive, sympathetic, feminine, caring, motherly, compassionate, warm, loving, complaining, fussy, gullible, nagging, servile, spineless, whiny, emotional, talkative	soft, empathetic, considerate, girly, thoughtful, nice, bitching, bitchin, bitchy, whining, whiney, whinny, naive, picky, cranky, bugging, gutless, moany, dramatic, chatty
physical	beautiful, cute, gorgeous, pretty	stunning, adorable
Male traits		
cognitive	analytical	logical
personality	adventurous, aggressive, competitive, daring, driven, active, masculine, strong, tough, leader, arrogant, boastful, cynical, dictatorial, egoistical, greedy, hostile, unprincipled	agressive, aggresive, forceful, busy, cocky, condescending, pompous, hard, manly, narcissistic, prideful, tyrannical, oppressive, selfish, unscrupulous, determined, ambitious
physical	burly, muscular, rugged, strong, athletic	sturdy, tough, fit, hulking, buff, muscly
Political traits		
competence	intelligent, hardworking, knowledgeable, educated	informed, knowledgable, competent, smart, clever, diligent, industrious
empathy	compassionate	sympathetic, empathetic, considerate
integrity	decent, honest, moral	nice, truthful, virtuous, good, honorable
leadership	inspiring, leader, competitive, ambitious, determined, driven, active, aggressive, charismatic, arrogant, powerful	inspirational, personable, charming, strong, busy, agressive, aggresive, forceful, cocky, condescending, pompous

Note. The table's second column includes traits from prior research. We used a master list of traits from Schneider and Bos (2014) (trait checklist, Appendix A, p. 265). The authors produced this list as a result of a pretest to their experiment and a review of traits used in previous studies including Diekman and Eagly (2000), Eagly and Karau (2002), Funk (1999), Heilman, et al. (1995), Huddy and Terkildsen (1993) and Kinder (1986). We excluded traits represented by multiple words (e.g., trait "Really cares about people like me"), as these categories fit better for surveys and experiments, but they we do not represent the colloquialism of social media conversations. We ensured that their synonymous adjectives were included (e.g., “caring”).

The trait groups were extended using a validated procedure of selecting most similar words from the pre-trained GloVe Twitter-200 (Pennington, Socher, & Manning, 2014) word embeddings (the table’s third column).

Appendix B

Methods

Data Pre-Processing

Replacement of Names. To correct misspelled names, e.g., *Hilary* or *Killary*, we employed word embeddings trained on the same corpus to consider top 15 words that are most similar to the names of the politicians. After their manual validation, references to politicians that otherwise could have been overlooked were included. Ambiguous last names were only replaced with the identifiers if they were preceded by the first names of the politicians, e.g., *Ron Kind* but not *Kind Ron*. In the remaining cases, politicians' names were replaced with identifiers if their first name was followed by their last name. For the prominent politicians, the unique IDs also replaced isolated mentions of either their first or last name, e.g., *Donald* interpreted as a reference to *Donald Trump*.

Word Embeddings

Hyperparameter Choice. The following parameters of the model to train word embeddings were used to yield meaningful results. The size of context around the target words was adjusted to 5, which is the default window size of word2vec. Although the embeddings size in word2vec is set to 300, to reduce memory consumption, dimensionality of word embeddings in this study was adjusted to 200. For the training of the vector space the sub-sampling threshold for frequent words of 10^{-5} was used following the recommendation in Mikolov et al. (2013) and Mikolov, Sutskever et al. (2013) to reduce the effect of the most frequent words on the word vectors. To manage words of rare occurrence, words with total frequency less than 10 were ignored by the model. The default initial learning rate of 0.025 decreased linearly to 0.0007 during the training process. Since the higher number of negative samples results in a better estimation (Levy, Goldberg, & Dagan, 2015), it was set to 15.

Training and Evaluation. The trained word vectors were subject to intrinsic evaluation, where human judgements on word relations were used to compare semantic similarity and analogical reasoning with word embeddings. The correlation of the human ratings from the Word-Similarity 353 (WS353, Finkelstein et al., 2002) with the cosine similarities of the word vectors of the respective pairs yielded a Pearson's $r = 0.56$, $p < 0.001$ and a Spearman's $r_s = 0.57$, $p < 0.001$. The results are comparable to the evaluations of Levy et al. (2015) and Lai et al.(2015). Further, SimLex-999 (Hill et al., 2014), comprised of 999 word pairs assessing their semantic similarity was conducted. The correlation resulted in a Pearson's $r = 0.28$, $p < 0.001$ and a Spearman's $r_s = 0.27$, $p < 0.001$. For comparison, the best model in Levy et al. (2015) had a Spearman's correlation coefficient of $r_s = 0.438$. Finally, the analogy test from Mikolov et al. (2013) was employed to compute performance of our model on semantic and syntactic questions of the test. The evaluation demonstrated a relatively low performance on the task, 40% against 66% achieved by the authors of word2vec (Mikolov, Sutskever et al., 2013). Rogers, Drozd, and Li (2017) note that many word vector models score under 30% on analogy tests, suggesting that this type of evaluation may not be ideal for testing assessing quality of word embeddings. We argue that the results of the mainstream evaluation approaches should be interpreted with caution. Firstly, it is problematic to objectively evaluate the results of an unsupervised machine learning task (in this case, neural word embeddings). Secondly, the size and diversity of a data set influence the performance of the model on these tasks. For instance, in the case of this study, the data set has a relatively narrow subject of discussion, i.e. politics, which may lack some words from the evaluation sets in the vocabulary (e.g., no word describing geographical locations that are common in the analogy test). We rely on the detailed review of the problems associated with word similarity evaluations published by Faruqui, Tsvetkov, Rastogi, and Dyer (2016).

Association Strength. To calculate the association strength between the word vectors, cosine similarity was employed. As the analysis included normalized word vectors, denominator in the cosine similarity calculation is omitted. Also referred to as bias indicator.

$$bias\ indicator = v_m \cdot v_i$$

In this equation, v_i is the average vector of a trait group, v_m is a vector of the politician's name.

References

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. Personality and Social Psychology Bulletin, 26(10), 1171–1188.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. Psychological Review, 109(3), 573–598.

Faruqui , M., Tsvetkov , Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. arXiv:1605.02276 [cs.CL]. https://arxiv.org/abs/1605.02276

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2002). Placing search in context: The concept re-visited. *ACM Transactions on Information Systems,* 20 (1), 116–131.

Funk, C. L. (1999). Bringing the candidate into models of candidate evaluation. Journal of Politics, 61(3), 700–720.

Heilman, M. E., Block, C. J., & Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers? Journal of Social Behavior and Personality, 10(4), 237–252.

Hill, F., Reichart, R., & Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation*. arXiv:1408.3456 [cs].*

Huddy, L., & Terkildsen, N. (1993). Gender stereotypes and the perception of male and female candidates. American Journal of Political Science, 37(1), 119–147.

Kinder, D. (1986). Presidential character revisited. In R. R. Lau & D. O. Sears (Eds.), Political cognition (pp. 233–255). Hillsdale, NJ: Lawrence Erlbaum

Lai, S., Liu, K., Xu, L., & Zhao, J. (2015). How to generate a good word embedding? *CoRR*, *abs/1507.05523*

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. https://doi.org/10.1162/tacl_a_00134

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs.CL].* https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, *2*, 3111–3119. https://arxiv.org/abs/1310.4546

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543. doi:10.3115/v1/D14-1162

Rogers, A., Drozd, A. & Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. In Proc. of the 6th Joint Conf. on Lexical and Computational Semantics, 135 –148.

Schneider, M. C., & Bos, A. L. (2014). Measuring stereotypes of female politicians. *Political Psychology,* 35(2), 245–266. doi:10.1111/pops.12040