Measuring Gender Bias in Digital Diplomacy

Keywords: gender bias, diplomacy, social media, politics

Extended Abstract

This paper examines online gender bias against women ambassadors across the globe. Today, diplomacy takes place not only behind closed doors but also online, with Twitter as a key social media platform. Despite increasing evidence that women pay a high price for participating in politics, often because of online hostility and harassment (Haakansson 2021, Gruzd et al 2021), little is known about the character and scope of gender bias against ambassadors on social media and thus the degree of gender discrimination against key foreign policy actors online. We develop a multidimensional approach to studying gender bias online, focusing on *visibility, negativity* in tone of tweets directed at diplomats and the use of *gendered language*. While there is relatively little academic work on gender bias in *digital* diplomacy, there is even less research that approaches the issue by going beyond Western countries or content in English.

The paper contributes to our understanding of gender bias and digital international politics in three ways. First, we present the first systematic and *global* analysis of how women diplomats are treated online across multiple continents. Second, we offer one of the first empirical studies of gender bias that is highly *multi-lingual* in its approach.

We test our hypotheses by constructing a unique dataset consisting of 1,960 ambassadors from 164 UN member states who have a Twitter account, their tweets as well as a multilingual set of 458,932 public tweet replies directed directly at them during January 31 to June 26 2021. We use XLM-RoBERTa (Conneau et al 2019) to infer the tone in the replies across all 65 languages.

Furthermore, we do not merely analyze words directed at ambassadors, but also investigate gender ladenness of the tweets. To this end, we employ the NRC VAD Lexicon (Mohammad, 2018), which to our knowledge is by far the largest and most reliable multilingual affect lexicon, and has been widely applied in linguistic studies. Using this approach, we examine the nature and scope *gendered language* by measuring the levels of dominance in the replies sent to the ambassadors before qualitatively analyzing the words most associated with replies sent to men and women ambassadors.

We find a clear gender bias in terms of *visibility*, measured as the number of retweets. Women receive on average 66.4% fewer retweets ($p<0.05$) than men. This is illustrated in the descriptive Figure 1 A. The difference in visibility through retweets is persistent even when examining ambassadors who themselves post many original tweets, retweets and replies as illustrated in Figure 1 D. Using a negative binomial variant of a Generalized Linear Mixed model (GLMM), we show that the difference in visibility is persistent even when controlling for sending countries (where the ambassadors are from) and receiving countries (where the ambassadors are assigned to). Using different model specifications for robustness, we also demonstrate that the gender bias in visibility remains even when controlling for the number of tweets uploaded by the ambassador and the global, diplomatic prestige of the receiving country. The latter is measured as the standardized indegree in a network of

countries (nodes), where a directed edge indicates that one country has sent an embassy to another country.

Contrary to our expectation, we find no evidence to support that women receive more *negativity* in public replies on Twitter than men. On a global level, Women ambassadors receive a 0.37 percent points lower proportion of negative replies than men (Figure 1 B). The difference is statistically significant ($p<0.05$), although not substantively large. The pattern remains the same when running linear regressions with country-level fixed effects, controlling for the log number of tweets uploaded by the ambassadors as well as the prestige of country that the ambassador has been assigned to. The difference is no longer statistically significant when controlling for the proportion of negative tweets posted by the ambassadors themselves. The overall pattern remains in the validation step, where we use incivility detection algorithm developed by Theocharis et al. (2020) to analyze the replies in English.

When examining the levels of *gendered language*, we find that women receive more dominant language in replies than their male colleagues (Figure 1 C). The average Dominance score in replies sent to women is 4% higher when comparing to men. A more qualitative analysis suggests that the words frequently sent to women are more associated with gender-stereotypical topics (e.g. “flowers”) than words sent to men.

Overall, our findings indicate that gender bias towards women ambassadors is more indirect and “hidden” than one may think. When examining the public replies alone, it may appear as if users predominantly interact with women ambassadors through a relatively positive tone. While the overall pattern in the language is gendered, the difference is not as extreme as one would expect outside of digital diplomacy. We demonstrate that women ambassadors themselves are much less visible than men in the same field. The difference is crucial and arguably as important as the more direct forms of gender bias through negative or gendered replies. The ambassadors’ main task on Twitter is to promote their country by engaging online audiences. Visibility is one of the main power resources on social media and a prerequisite for carrying out digital diplomacy in the first place. The lack of visibility may be a barrier that disadvantages women compared to their male colleagues.

References

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

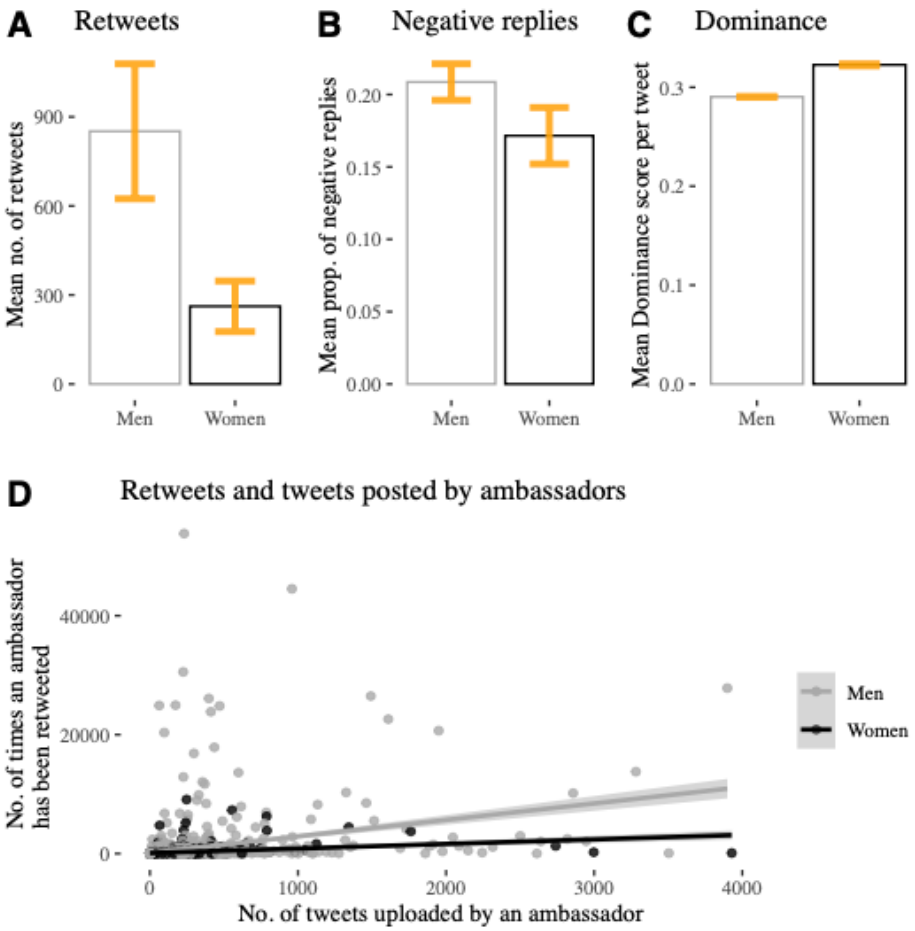
Haakansson, S. (2021). Do women pay a higher price for power? Gender bias in political violence in Sweden. *The Journal of Politics* 83 (2), 515–531.

Kumar, P., A. Gruzd, and P. Mai (2021). Mapping out Violence Against Women of Influence on Twitter Using the Cyber–Lifestyle Routine Activity Theory. *American Behavioral Scientist 65* (5), 689–711.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the As- sociation for Computational Linguistics* (Volume 1: Long Papers), Melbourne, Australia.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *Sage Open*, 10(2).

Figure 1: Retweets, negative replies and dominance received by ambassadors



All four figures show descriptive means with 95% confidence intervals and without controls. Figure **C** shows mean Dominance score *per tweet*, while the remaining figures show mean number of retweets and proportion of negative replies *per ambassador*.