Methodology

To analyze congressional responses to a series of harassment reports, researchers first obtained all 44,792 Facebook posts created between Oct. 1 and Dec. 30, 2017, from each of the 491 voting members of the U.S. Congress who use official Facebook pages for outreach to the public and who posted at least once during the study period.

An algorithm was used to identify posts related to sexual misconduct. The algorithm used information based on a small number of human decisions – in this case, whether or not a post discussed sexual misconduct – in order to classify the content of the entire set of Facebook posts. This procedure ultimately identified 551 posts as posts related to sexual misconduct. Overall, these posts represented 1.2% of all posts in the data, though 44% of lawmakers mentioned sexual harassment at least once on their official Facebook accounts.

In order to determine whether or not each post in the sample contained a discussion of sexual misconduct, two researchers each manually classified a sample of 50 posts, including randomly selected posts, as well as posts that contained one of several keywords related to sexual misconduct. They reached perfect agreement across all 50 posts. After determining that the findings of both researchers were consistent with each other, an additional sample of 500 posts were classified. These 500 human-classified posts were used as the basis for creating an algorithm which would classify the remaining posts.[1] The algorithm was then applied to the full set of all remaining posts, after adjusting for the oversampling of particular words.

An analysis of a small number of posts that had been classified by the researchers found that the model was able to detect the discussion of sexual misconduct in posts with a very low rate of either false positives or false negatives.[2]

Assessing the level of online engagement that a post received required taking into account the fact that some members have more followers than others and might be more prone to receiving more likes and comments regardless of the content of their posts. Directly comparing members of Congress who average thousands of likes on each post with those who average dozens of likes is like comparing apples and oranges. No matter what the member with a high level of engagement posted about, those posts would all appear to inspire vastly more engagement than posts created by the member who averages only a small number of likes and comments.

[1] The specific algorithm researchers used was Extreme Gradient Boosting, implemented by the XGBoost module in Python. The algorithm used features generated via Term Frequency-Inverse Document Frequency (TFIDF). Each row was augmented using word embeddings generated by a Word2Vec model trained on Google News.
[2] Using 5-fold cross validation, recall was 0.95 and precision was 0.97.

To create an "apples to apples" comparison, researchers first normalized the number of likes and comments that every post received, to help account for a small number of posts with extremely high levels of likes and comments. This helps reveal the underlying association between the *content* of the posts and engagement, without letting posts by the most popular members of Congress bias the analysis. And, because the engagement numbers between members differ by orders of magnitude, researchers transformed the numbers using the base 10 logarithm – a common mathematical transformation. They then estimated statistical models – a tool that uses information from all the individual posts in the study in concert – to examine the relationship of interest. These models estimate how the number of likes and comments that post received online relates to attributes of each post – for example, whether or not it included discussion of sexual misconduct.

Specifically, researchers examined the number of likes or comments a post received with respect to the gender of the legislator who created the post, a variable that captured whether or not the post discussed sexual misconduct, and the party of the legislator who created the post. The models also included a separate baseline for each member of Congress, to help account for their different levels of engagement.

Posts that discuss sexual misconduct receive more likes and comments than other posts

	Log_{10} (Likes)	Log_{10} (Comments)
Sexual misconduct	0.224	0.186
	(0.026)	(0.032)
Male legislator	0.005	-0.035
	(0.053)	(0.058)
Sexual misconduct * Male legislator	-0.045	0.037
	(0.035)	(0.043)
Republican	-0.057	0.342
	(0.042)	(0.046)
Constant	1.922	1.257
	(0.046)	(0.051)
Random effects for each member	X	X
N	44,792	44,792

Note: The models are ordinary least squares multiple regressions with random effects for each member.