Supplemental Materials (Online Appendix)

Politicians in the Line of Fire: Incivility and the Treatment of Women on Social Media

Additional Information on Data Collection

We retrieved messages from the Twitter platform using the public streaming API during a period of one month for each country. For Canada, data collection took place from 7:00 AM to 3:00 AM, Eastern time, using a script launched automatically every day. The period of collection ranges from April 24 to May 26, 2017. For the US Senators, the platform was streamed in real-time between May 27 and July 5, 2017, during the same hours each day. In total, we collected 551,373 tweets addressed to Canadian politicians, and 5.6 million targeted at US Senators. There were no interruptions of service during that period. The Twitter streaming API is limited to 1% of the total quantity of statuses posted on the site at any given point (for detailed discussions, see Morstatter et al. 2013; Morstatter, Pfeffer, and Liu 2014; Joseph, Landwehr, and Carley 2014). Since we relied upon very specific search filters—the handles used by each politician—we generally remain well below the rate limits. This means that our data represent not a sample of tweets during that period, but virtually all the messages matching our search criteria. More specifically, Twitter reports the number of statuses that could not be retrieved when exceeding the rate limit. In total, 1,952 messages were not retrieved in Canada due to rate limits (about 0.4% of the total corpus size), and 81,322 for the United States (about 1.4% of the total). In other words, we were able to collect roughly 99% of all messages meeting our criteria. Finally, note that some politicians had more than one Twitter account, in which case we used the one associated with their official function.

The statuses were processed to extract the displayed text using custom scripts. We considered statuses with at least three tokens (words or punctuation marks). We removed external links (URLs) from these messages, and after associating them to the politician they target, we removed all handles from the text. We removed all duplicate texts and purged the corpus from shared messages (retweets). Hence, our data collection is restricted to unique messages addressed directly to politicians. Moreover, we restricted the data to messages sent to a unique politician (that is, we exclude messages addressed to more than one recipients from our sample of politicians). Finally, we exclude a few tweets sent by the politicians themselves, to restrict our focus on the general public. The curated datasets contain 170,114 and 2.1 million tweets, respectively for Canada and the USA. We coded the gender and other attributes of politicians in our sample using their official biographies. Our measure of politician visibility is a variable measuring the count of followers on the site. This information was extracted from the website using the REST API between June 8 and June 10, 2017.

Defining Uncivil Tweets

Our training data annotated by human coders was described in the text, but we provide additional information here. Workers were provided with specific guidelines to identify uncivil tweets based on the six criteria mentioned in the text and discussed below. We also included specific examples and advice to interpret these criteria. FigureEight (formerly CrowdFlower) uses test questions—questions for which we provided the ground truth—to create a trust score for each coder. The platform's algorithm then computes a confidence in each judgment as the proportion choosing the majority category weighted by the individual trust scores. The average confidence is 93.0% for the American sample, and 94.3% for the Canadian sample.

When devising instructions, we defined as uncivil those messages containing explicit forms of offensive language that human coders can readily detect—namely swear words, vulgarities, and direct insults (for instance, "idiot", "stupid")—as well as forms of incivility along the lines of those used in Papacharissi (2004)—namely threats, personal attacks directed at one's private life, and attacks toward groups (hate speech). This choice differs from a body of literature in political science that adopts broad definitions of incivility for the study of elite discourse—including negative campaign advertisements or an adversarial tone during televized debates. For example, in their experiment on televized incivility, Mutz and Reeves (2005) organized a mock debate between politicians, with subjects exposed to either a civil or an uncivil version of the same exchange. The uncivil tone was characterized with phrasings such as "You're really missing the point" (Mutz and Reeves 2005, 199), which represent mild violations of social norms yet were sufficient to affect the subjects' levels of political trust. Brooks and Geer (2007, 5) adopt a slightly different definition, identifying incivility in terms of discursive behaviours resorting to "animosity and derision" and the addition of "inflammatory comments that add little in the way of substance to the discussion." These conceptions of what constitutes civility have merits for studying elites, but they establish a high bar when analyzing political debates among members of the public on social media, where transgressions tend to be more extreme and more common. For example, it would be unlikely to witness a politician using profanity in public statements, yet such forms of incivility are part of the linguistic register in social media.

By establishing a higher threshold for what counts as incivility, we allow for the adversarial tone and heated exchanges to be expected in online debates. Actual examples from our corpus may help to illustrate the implications of our definition. The following two examples contain direct insults:

> *I bet your sick & twisted mind gets off on it. I know ppl like you; chip on shoulder, rejected by the opposite sex. You have a"loser" aura.*
>
> *How about you put a sock in it and go away!!!! You and your pant suit sisters need to ride off into the sunset. You are a b***h.* [*expletive blurred*]

In both cases, the nature of the message goes beyond the expression of political opinions during a heated exchange. Since they comprise one or more elements of our above definition (direct insults, personal attacks), we view them as uncivil. On the other hand, we consider the following example to fall within the boundaries of civility:

> *You have no idea what rights and freedoms even mean to Canadians. Youre out of touch.*

This tweet expresses a forceful criticism of a politician's character, and the statement rests on subjective assumptions. Unlike the two previous examples, however, the message does not contain offensive language or attacks referring to someone's private life. Notice that such a comment could be considered uncivil using Mutz and Reeves (2005)'s definition, if it were used in the context of a debate between political candidates. But it exemplifies a common type of criticism on social media. Conflating statements of that nature with the previous two would seriously boost our estimates about the prevalence of incivility, and in the process we would risk overlooking the severity of the more abusive comments. We prefer to rely on a more conservative approach.

Description of Machine Learning Models

Our models use three types of linguistic features as predictors for the category of a tweet. First, we make use of the 2,000 unigrams and bigrams (sequences of one and two words) most predictive of the class of a tweet in the annotated sample, based on chi-square values. Occurrences for these 2,000 expressions are converted into numerical values using a term-frequency/inverse document frequency (TF-IDF) weighting scheme, which gives additional importance to less common utterances. Prior to this step, we lemmatized the training sample (that is, we reduced each noun and verb to its root form) and removed English stop words, user handles, as well as mentions of the names of politicians in our main sample. These last steps avoid the reliance on clues too closely related to the recipient of the tweets when predicting their category. A few tweets with no textual content left after these steps were removed from the sample.

Second, we devise an indicator measuring the semantic similarity of a tweet with respect to a reference list of insults and swear words. This reference list is a filter for inappropriate content on the web, namely the *swearjar* JavaScript library.[1] We use a dataset of word embeddings—

[1] The list contains 247 common swear words and vulgarities for which we can compute similarity metrics.

the numerical coefficients of neural network models predicting word co-occurrences in large collections of texts (Mikolov et al. 2013; Pennington, Socher, and Manning 2014)—to compute the cosine similarity between any new lemma and those contained in the reference list. Specifically, we rely on publicly released word embeddings, pre-trained on a corpus of 27 billion tokens from the Twitter platform, fitted using the GloVe algorithm (Pennington, Socher, and Manning 2014). Our indicator is the maximum cosine similarity with the reference list of abusive words, for each tweet: the higher this maximum value, the more likely a tweet contains an offensive word.[2]

Third, we measure the sentiment of each tweet as a numerical value. We rely upon the *vader* library for Python (Gilbert and Hutto 2014), which was designed for social media data. The library computes a compound score ranging from -1 to 1 representing the emotional polarity of a document, from negative to positive. Since uncivil messages are more likely to be negative in tone, we expect sentiment to be a relevant predictor, even though this feature would be insufficient by itself.

Our objective is to fit a model that can both predict the incivility of individual tweets and the aggregate proportions of uncivil tweets accurately in the full corpus. To find the most suitable model, we compared the performance of classifiers commonly used for the analysis of text documents: support vector machines (SVM), decision trees, and logistic regressions. Our most accurate model is a SVM classifier fitted using 50 bootstrap aggregating (bagging) replications (Breiman 1996). Put simply, bagging consists of running the predictions multiple times after randomly resampling the training examples, and choosing the class (civil/uncivil) predicted the most often by the models. This method reduces concerns about overfitting (Bauer and Kohavi 1999). We also rely on a bagging estimator that accounts for the imbalance between the classes using random undersampling of the majority category.[3]

Table A1 reports accuracy statistics for our models, comparing SVMs with and without the bagging algorithm. Following conventions in the field of machine learning, we evaluate each model by first separating the sample into training and testing sets, to emulate the accuracy in

[2] This indicator accounts for obfuscation spellings and neologisms commonly used as insults on social media.

[3] We fit all models using the *sklearn* and *imblearn* libraries for Python. Our models will be made available to researchers upon publication.

the prediction of unseen documents. The statistics in Table A1 are averaged over 10 replications, using stratified 10-fold cross-validation (i.e. randomly splitting the sample into 10 parts and repeating the training and prediction stages 10 times using a different testing sample each time). The first two statistics evaluate the accuracy of individual class predictions in the testing sample: the percent correctly predicted and the area under the receiver operating characteristic curve (AUROC). As can be seen, the models using balanced bagging correctly predict the class of a tweet for close to 90% of cases in the American sample, and about 92% of cases in the Canadian sample. The distribution of tweets across the two classes being unbalanced, the AUROC statistic represents a more reliable metric since it assesses the capacity of each model to avoid both false positives and false negatives (the closer to 1, the better the model). Once again, the bagging estimators outperform the standard models. Finally, the proportion error is the absolute difference in the aggregate proportions of tweets in each class, that is, the difference between the percentage of tweets deemed to be uncivil by human coders and the percentage predicted to be uncivil by the model. The lower the error, the more accurate the prediction. We compare this last statistic to the one computed using Hopkins and King (2010)'s estimator (*ReadMe*), which we fit on the first part of a random 50/50 split of the annotated sample, and evaluate on the other part.[4] This model is not designed to predict individual documents, so the first two accuracy metrics cannot be computed. However, the *ReadMe* estimator tends to be more accurate at fitting proportions. As a result, it represents a useful benchmark to assess our models. Our final models generate proportions close to those achieved by this estimator.

Additional Results

Table A2 compares the baseline rates of uncivil tweets reported in the main text, along with additional word frequencies based on popular lexicons. These are frequencies by 1,000 words of expressions contained in the *swearjar* lexicon introduced earlier, and in two categories from the 2015 dictionaries of the popular psycholinguistic software LIWC (Tausczik and Pennebaker

[4] We use random subsets of 20 words and 300 repetitions. We fitted the model using the same 2,000 unigrams and bigrams as for the other classifiers. For information on these parameters, see (Hopkins and King 2010).

Table A1: Accuracy Results

Sample	Model	Accuracy	AUROC	Proportion Error
USA	SVM	87.05%	0.711	0.031
	SVM (Balanced Bagging)	89.27%	0.763	0.024
	ReadMe			0.020
Canada	SVM	90.65%	0.704	0.023
	SVM (Balanced Bagging)	91.68%	0.766	0.010
	ReadMe			0.007

Accuracy statistics are computed using stratified 10-fold cross-validation. We report average statistics over the 10 folds. The accuracy is the percent correctly predicted in the testing sets. AUROC stands for the area under the receiver operating characteristic (ROC) curve. We use Platt's method to retrieve the probability of a positive outcome with SVMs (Platt 1999). The proportion error is the absolute difference between the predicted and the true proportions of civil tweets.

2010), namely swear words and negative words. As is the case for the predicted proportions of uncivil tweets, the additional frequencies suggest that swear terms and negative words are used more frequently in messages sent to male politicians than in messages sent to female politicians. Again, these comparisons ignore the differences in status between politicians, which are relevant to derive substantively meaningful conclusions. Since men tend to be overrepresented among visible politicians, a multivariate analysis taking into account this confounder is justified.

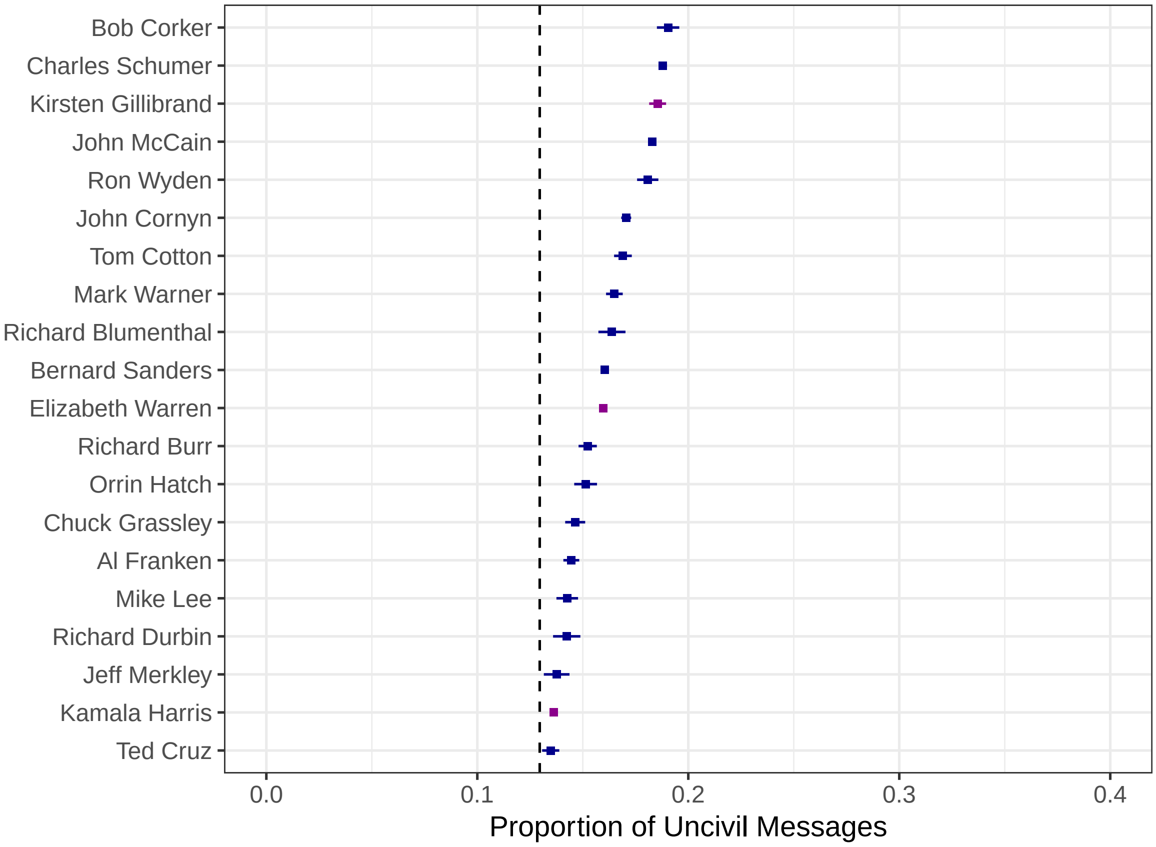
Table A2: Inferring the Level of Incivility by Gender

		United States			Canada		
	Method/Lexicon	Women	Men	Total	Women	Men	Total
Fitted Proportions	Classifier	12.95%	14.54%	14.13%	8.55%	11.66%	10.69%
Frequencies by 1,000 Words	Swear Jar	6.67	7.44	7.25	3.97	7.44	6.32
	LIWC Swear Words	11.35	12.63	12.31	7.01	11.98	10.38
	LIWC Negative Words	71.74	73.48	73.05	49.29	61.23	57.38
Corpus Size		530,663	1,545,175	2,075,838	53,195	116,919	170,114

Proportions predicted with a balanced bagging model using 50 replications of support vector machine estimators.

Figure A1 replicates the figure presented in the main text for Canada, and shows the 20 US Senators most often targeted by uncivil messages, restricting to those having received at least 10,000 tweets. As can be seen, the primary targets tend to occupy important positions in the upper house. For instance, the Democratic minority leader Chuck Schumer ranks in second position, and Senators with a large follower count on Twitter such as John McCain and Bernard Sanders also feature in the Top 20. There are few women in the Senate to begin with, and there are even fewer of them enjoying a high status. But for the women who do have visibility, for instance New York Senator Kristen Gillibrand and Elizabeth Warren, uncivil messages are frequent.

Figure A1: US Senators Most Targeted by Uncivil Messages



The vertical line indicates the average proportion of uncivil messages received by Senators with at least 10,000 messages addressed to them in the corpus. We use a color code to distinguish between female and male politicians.

Aggregate Empirical Models

To assess the robustness of the multivariate results presented in the main text, we replicated the analysis by aggregating the count of uncivil tweets received by each politician. This considerably reduces the sample size (to 195 politicians in Canada, and 100 Senators in the United States). The aggregate dependent variable also accumulates prediction errors, and as a result this transformation may inflate standard errors. Nonetheless, we show that the main finding is replicated with an aggregate dataset, for both countries. Moreover, we replicate the results using the counts of swear words contained in the tweets sent to each politician as the dependent variables. The counts are based on the LIWC dictionary and the swearjar list mentioned above. We show that the relationship emphasized in the main text is supported when using these alternative dependent variables. These replications suggest that the interaction of gender and visibility is not simply an artifact of the methodology used to predict uncivil tweets. For all models, we rely on quasi-Poisson regressions accounting for overdispersion.[5] All models use an offset of the log of the total number of tweets received to account for exposure.

To begin, Tables A3 and A4 report models with only two covariates and an interaction term, using each of the three different aggregated count variables as the outcome. As can be seen, the interaction between the female gender and the measure of visibility remains positive and statistically significant, as was the case in the main models. Again, this suggests that women politicians are more likely to become targets of uncivil messages, but conditional on gaining visibility. Without a high level of visibility, however, men are more likely to face incivility. As was the case for the main models, the results appear more robust, in terms of statistical level of confidence, when considering the sample of Canadian politicians.

Tables A5 and A6 report the output of count models including control variables. We account for party affiliation and visible minority status, as well as a variable relevant for each country. For Canada, we include a binary variable accounting for the level of government (which equals 1 for the federal level). For the United States, we include instead the seniority of a Senator in logged number of years. As can be seen, the main finding holds after accounting for these control variables.

[5] On the properties of quasi-Poisson regressions, see Ver Hoef and Boveng (2007).

Table A3: Aggregate Models of Incivility (Canada)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-2.078^{***}	-2.998^{***}	-2.614^{***}
	(0.500)	(0.569)	(0.713)
Log Follower Count	0.150^{***}	0.189^{***}	0.205^{***}
	(0.010)	(0.011)	(0.013)
Gender × Log Follower Count	0.195^{***}	0.270^{***}	0.233^{***}
	(0.043)	(0.049)	(0.062)
Intercept	-4.248^{***}	-5.079^{***}	-5.790^{***}
	(0.149)	(0.155)	(0.196)
Observations	195	195	195

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table A4: Aggregate Models of Incivility (United States)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-1.757^{**}	-1.790^{**}	-1.836^{**}
	(0.546)	(0.561)	(0.683)
Log Follower Count	0.069^{***}	0.056^{***}	0.053^{**}
	(0.014)	(0.014)	(0.017)
Gender × Log Follower Count	0.120^{**}	0.122^{**}	0.125^{*}
	(0.040)	(0.041)	(0.050)
Intercept	-2.828^{***}	-3.097^{***}	-3.580^{***}
	(0.181)	(0.184)	(0.224)
Observations	100	100	100

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table A5: Aggregate Models of Incivility, with Controls (Canada)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-2.440^{***}	-4.930^{***}	-4.009^{***}
	(0.535)	(0.752)	(0.977)
Log Follower Count	0.184^{***}	0.179^{***}	0.207^{***}
	(0.015)	(0.016)	(0.022)
Gender × Log Follower Count	0.240^{***}	0.461^{***}	0.373^{***}
	(0.048)	(0.067)	(0.087)
Visible Minority	0.497^{***}	0.134	0.226
	(0.076)	(0.091)	(0.122)
Party (Liberal = 1)	-0.155^{*}	-0.453^{***}	-0.351^{**}
	(0.078)	(0.088)	(0.121)
Federal Level	0.063	0.416^{***}	0.311^{**}
	(0.072)	(0.087)	(0.117)
Intercept	-4.674^{***}	-4.900^{***}	-5.777^{***}
	(0.178)	(0.196)	(0.270)
Observations	195	195	195

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.

*p<0.05; **p<0.01; ***p<0.001

Table A6: Aggregate Models of Incivility, with Controls (United States)

	Dependent variable:		
	Uncivil Tweets	LIWC Swear Words	Swearjar Words
Gender (Female = 1)	-2.170^{***}	-1.882^{**}	-2.136^{**}
	(0.551)	(0.608)	(0.753)
Log Follower Count	0.058^{***}	0.055^{***}	0.046^{*}
	(0.014)	(0.015)	(0.018)
Gender × Log Follower Count	0.150^{***}	0.124^{**}	0.145^{*}
	(0.041)	(0.046)	(0.057)
Visible Minority	0.194^{*}	0.136	0.179
	(0.089)	(0.098)	(0.120)
Party (Democrat = 1)	0.122^{*}	0.112^{*}	0.062
	(0.048)	(0.053)	(0.066)
Log Seniority	0.118^{***}	0.049	0.067
	(0.031)	(0.034)	(0.042)
Intercept	-3.006^{***}	-3.221^{***}	-3.671^{***}
	(0.174)	(0.189)	(0.233)
Observations	100	100	100

Notes: Quasi-Poisson regression models using the count of uncivil tweets (model 1), or the count of lexicon words based on the resource indicated in the column headers (models 2 and 3). Each model includes an offset for exposure, using the log of the total number of tweets received during the period.

*p<0.05; **p<0.01; ***p<0.001

References

Bauer, Eric, and Ron Kohavi. 1999. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." *Machine learning* 36(1-2): 105–139.

Breiman, Leo. 1996. "Bagging predictors." *Machine learning* 24(2): 123–140.

Brooks, Deborah Jordan, and John G. Geer. 2007. "Beyond Negativity: The Effects of Incivility on the Electorate." *American Journal of Political Science* 51(January): 1–16.

Gilbert, C.J., and Eric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14).* pp. 216–225.

Hopkins, Daniel J, and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1): 229–247.

Joseph, Kenneth, Peter M Landwehr, and Kathleen M Carley. 2014. Two 1% s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's Streaming API. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction.* Springer pp. 75–83.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR.*

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu. 2014. When Is it Biased? Assessing the Representativeness of Twitter's Streaming API. In *Proceedings of the 23rd International Conference on World Wide Web.* ACM pp. 555–556.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.*

Mutz, Diana C., and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *The American Political Science Review* 99(1): 1–15.

Papacharissi, Zizi. 2004. “Democracy online: civility, politeness, and the democratic potential of online political discussion groups.” *New Media & Society* 6(April): 259–283.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Platt, John C. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*. MIT Press pp. 61–74.

Tausczik, Yla R., and James W. Pennebaker. 2010. “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods.” *Journal of Language and Social Psychology* 29(1): 24–54.

Ver Hoef, Jay M., and Peter L. Boveng. 2007. “Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?” *Ecology* 88(11): 2766–2772.