# Creating economically relevant measures of sexist sentiment from Twitter data

An exploration of machine learning for economic research

Mckay Jensen*†

April 2019

## 1 Introduction

Large amounts of textual data are continuously being generated by users of social media websites; these data contain information that could potentially be useful for predicting economic trends. However, it is not obvious which characteristics of such data would have predictive power – most online content is uninformative "noise" as textual data is essentially an extremely high-dimensional feature set with most features having little predictive power. Machine learning presents a viable way to identify patterns (useful features) in data sets with such complicated features. In particular, a number of machine learning models have been developed to analyze the sentiment of text posted on social media. This project is a demonstration of how such models can be used to quantify sentiment and explore connections between online sentiment and real-world economic conditions. In particular, the question I investigate here is whether data from Twitter can be used to create a measure of sexism that correlates with gender economic inequality in U.S. states. Using a recurrent neural network and a set of approximately 4 million Twitter status updates (tweets), I was able to create a measure of sexist sentiment with significant variation from state to state; this sexism measure is correlated with various indicators of women's economic well-being. I end this analysis with a discussion of possible extensions to this project and a general exposition of the utility of machine learning

*The code for this project is available at `https://github.com/quevivasbien/twitter-sexism`.

†This project was sponsored by the University of Utah Office of Undergraduate Research. Thanks to Günseli Berik (University of Utah) for general guidance.

for research in economics.

## 2 Background

Using text as data is relatively new for economic research, having been facilitated relatively recently by advances in machine learning that make it easier to reduce the high-dimensional features in written language to a level where they can be integrated into quantitative economic analysis. Nevertheless, some researchers have already begun exploring the possibilities of such methods and have suggested some guidelines (see Gentzkow, Kelly, & Taddy [5]). In line with those recommendations and the existing (and substantive) literature on machine learning for text analysis, the approach I take here is to use deep learning to generate predictions on the sexist content of vectorized text (see "Machine learning"). Those predictions can then be compiled to create aggregate sexism scores that can be plugged into econometric models for further analysis.

There has lately been a great deal of interest in the use of computational tools for identifying online sexism, in part fomented by Alice Wu's analysis of the Economics Job Market Rumors forum [16]. In line with this interest, I chose to begin my investigation by focusing on sexist language on social media. However, it should be noted that my analysis is substantively different from Wu's work: whereas Wu used a lasso-logistic model to determine what types of words were most strongly associated with males and females, my goal is to employ machine learning to identify sexist *tone* in social media posts and identify geographical and temporal patterns in the resulting data. Additionally, while Wu's work was presented as evidence of sexist attitudes toward women within the economics research community, my intent to explore whether sexist language is indicative of broad economic circumstances for women in the United States.

Various events in the socio-economic sphere (e.g. elections, holidays) have been shown to have very measurable effects on sentiment scores as calculated from social media data [3]. Broad economic conditions also have a measurable effect on social media sentiment – for example, Lansdall-Welfare, Lampos, and Christianini [13] showed that cuts in government spending corresponded to an increase in negative sentiment in Twitter data. These research efforts have shown that sentiment data can be used to retroactively identify economically significant events. This project is the first stage of an effort to verify the utility of machine learning-generated

sentiment data as an economic indicator and begin investigating the causal and predictive power of such indicators. Although the results presented here are mainly correlative, I present a basic plan for further investigation that would build upon my present findings (see "What's next").

Because much of the past research on machine learning classification of sexist language has focused on data from Twitter, I chose to focus my analysis on sexist language on Twitter.

## 3 Machine learning

To build a machine learning classifier capable of recognizing sexism in tweets, I used a data set from Waseem and Hovy [15] containing 12215 tweets, each labeled as sexist or not sexist (2949 were labeled as sexist). The data set also included labels for racism, although for simplicity I chose to focus only on sexism. The labels were assigned manually, according to the following criteria:

> A tweet is offensive if it
>
> 1. uses a sexist or racial slur.
> 2. attacks a minority.
> 3. seeks to silence a minority.
> 4. criticizes a minority (without a well founded argument).
> 5. promotes, but does not directly use, hate speech or violent crime.
> 6. criticizes a minority and uses a straw man argument.
> 7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
> 8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"
> 9. defends xenophobia or sexism.
> 10. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

As shown by Badjatiya, Gupta, et al. [1], a long short-term memory (LSTM) neural network with a trainable word embedding works well for this particular classification task. The specific model architecture I found to work best is the following:

| Layer | Details |
|---|---|
| Embedding | 25-dimensional, constrained to 40 words max |
| Dropout | 25% dropout |
| LSTM | 50 outputs, ELU activation function |
| Dropout | 50% dropout |
| Fully-connected (dense) | 1 output, sigmoid activation function |

The weights for the embedding layer were initialized with GloVe vectors pre-trained with text from Twitter[1] and were allowed to be refined during the training process. Dropout layers were added to minimize the degree to which the model over-fits to the data it is trained on. The final layer outputs a value between 0 and 1 that can be interpreted as the likelihood that a given tweet has the same sort of sexist speech elements found in the training set (see *Appendix 1: Interpretation of machine learning output*).

To evaluate the model's performance, I ran a 10-fold cross-validation on the training data and computed the precision, recall, and F1 scores over each fold given a cutoff value of 0.5 (i.e. tweets are assumed to be sexist if the neural network outputs a value over 0.5 and not sexist otherwise). When averaged over all folds, the scores are considerably better than those achieved by Waseem and Hovy and comparable to the best outcomes of more sophisticated attempts by Badjatiya et al. and Pitsilis, Ramampiaro and Langseth [1] [11], as shown in Table 1.

Table 1: Performance comparison for sexism classifiers

| | Precision | Recall | F1 |
|---|---|---|---|
| Waseem & Hovy | 0.7290 | 0.7774 | 0.7391 |
| Badjatiya et al. | 0.9300 | 0.9300 | 0.9300 |
| Pitsilis et al. | 0.9305 | 0.9334 | 0.9320 |
| My classifier | 0.8877 | 0.8901 | 0.8880 |

It should be noted that the other classifiers in Table 1 were trained using both sexist and racist tweets, whereas my classifier was trained only for sexist tweets. For this reason, the performances are not directly comparable, although they do suggest that my classifier is nearly optimized given the available technology and the limitations of the data set (including racist tweets should improve the classifier's performance since the Waseem & Hovy data set includes more racist tweets than

[1] `https://nlp.stanford.edu/projects/glove/`

sexist tweets). The scores also change depending on the cutoff value used; for example, increasing the cutoff will increase precision but reduce recall, while the F1 score should stay about the same. In practice, I simply used the raw output from the classifier, rather than using any cutoff value, since I was interested in the likelihood of a given tweet being sexist; however, I used a cutoff here for purposes of comparison with previous work with this same dataset.

## 4 Data collection

I made use of Twitter's publicly available search API[2] to collect a sizeable data set of tweets originating within the United States. All tweets were from 2016 (to match the tweets in the training set as well as the data on economic outcomes used later), and no more than 100 tweets were collected for any single user. Since I wanted to explore differences by state, only users for which I could identify their location were used. The general process for collecting data can be summarized as

1. Collect random user IDs from the real-time public stream, filtering for users based in the United States.
2. For each user ID:
   (a) Download up to 100 tweets from that users timeline, with filters so only tweets from 2016 are retrieved.
   (b) Clean text (remove unwanted characters, vectorize based on GloVe embedding, and save location, user ID, and date).
   (c) Add user to master list to prevent duplicates.

The total number of tweets collected was 4,282,103, with at least 20,000 tweets for each state.

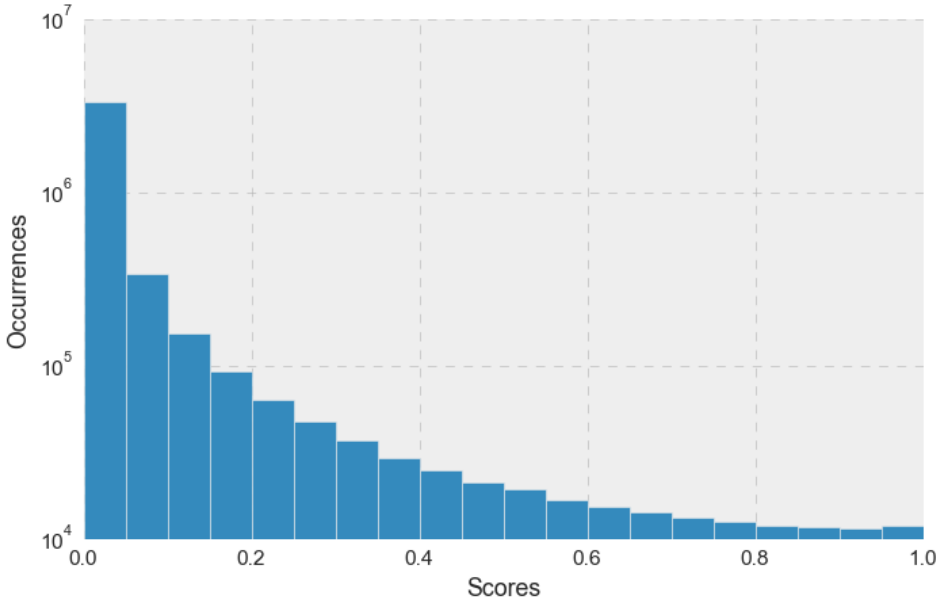
## 5 Data analysis

Using the pre-trained machine learning classifier, each tweet was assigned a "sexism score," which was simply the value output by the final layer of the neural network. Recall that these values fall between 0 and 1 and are estimates of the likelihood that

[2] `https://developer.twitter.com/`

a given tweet has the same sort of sexist speech elements found in the training set. Tweets with higher scores, therefore, are more likely to be sexist in nature. The distribution of scores for all tweets collected is shown in the histogram in Figure 1. Notice that most tweets have low scores (note the log scale on the y-axis).

Figure 1: Distribution of sexism scores, all tweets



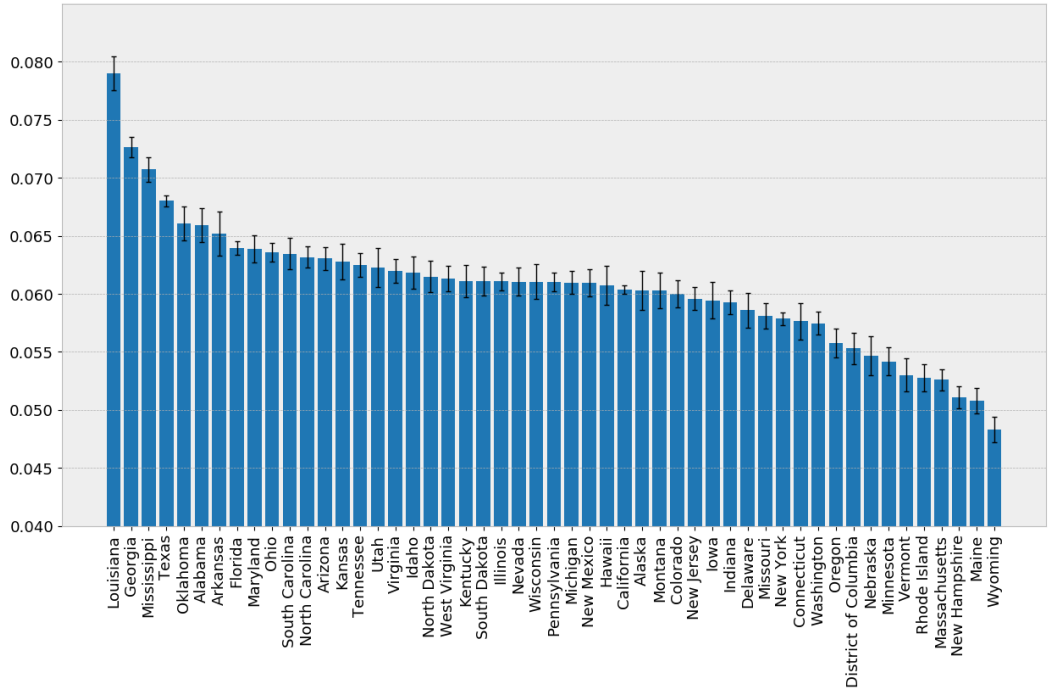

I was interested in creating some measure to allow for comparisons in the amount of sexist language between states. A natural idea is to simply compute the score for a given state as the average score for all tweets from that state. This score can be interpreted as an estimate of the probability that a randomly selected tweet from that state is sexist, since, for example if we have $n$ tweets $t_1, \ldots, t_n$ from a state $S$, then we can approximate the probability of a given tweet $t$ from state $S$ being sexist as

$$P(t \text{ is sexist} | t \in S) \approx \frac{1}{n} \sum_{i=1}^{n} P(t_i \text{ is sexist}).$$

When scores are calculated for each state in this way, the scores for each state are summarized in Figure 2.

Although most states seem to cluster around a score of about 0.06, there are clearly some states with higher rates of sexist language: Louisiana in particular

Figure 2: Sexism scores for all states. Vertical bars are 95% confidence intervals.



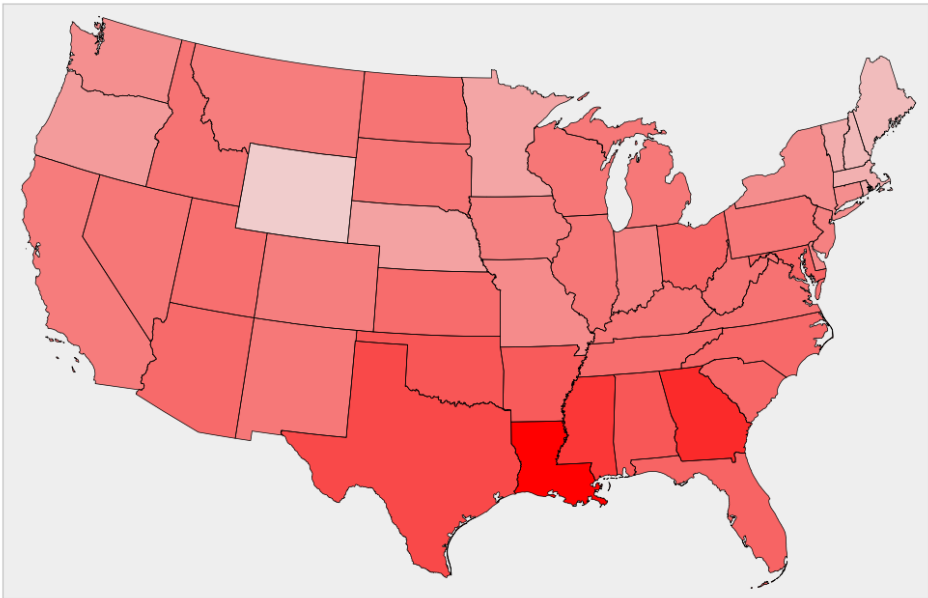

stands out, with a score more than 50% higher than some states near the end of the spectrum. Exact scores, standard errors, and sample sizes can be found in *Appendix 2: Sexism scores*.

When the relative scores are displayed on a map of the U.S., we can see that states in the southeastern U.S. tend to have the highest sexism scores; there isn't any other obvious geographical trend in the scores. This is displayed in Figure 3. A basic linear model indicates that, all else equal, states in the U.S. Census Bureau's southern region[3] can be expected to have sexism scores 0.007 higher than other states ($p \approx 0$). This is quite significant considering that the baseline score (mean score for states not in the south) is 0.058 – scores for southern states are, on average, about 12% higher. Survey data on attitudes toward women has indicated that, at least among the college-age population, attitudes toward women are somewhat more

[3]These states include Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia. See `https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf`.

conservative in southern states compared to the rest of the country [14]. There is also some evidence suggesting that prejudiced attitudes toward women are more prevalent in Christian and Republican-leaning populations [2]. The geographical trend in the scores presented here suggest that these attitudes are reflected in online content.

Figure 3: Heatmap of sexism scores. Darker shades of red correspond to higher Twitter sexism scores.



## 6 Relationship with women's economic status

The Twitter sexism scores are informative by themselves. But in evaluating these scores' usefulness as an economic measure, it is enlightening to see how they relate to other data about women's economic status by state.

Using data from the Institute for Women's Policy Research (IWPR),[4] I determined the correlations between the state-by-state Twitter sexism scores and various measures of women's economic status. The results of this correlative analysis are listed in Table 2.

[4]Data from the IWPR's 2015 report is available at `https://statusofwomendata.org/explore-the-data/download-the-data/`. For this analysis, I used data from the 2018 report, although that data is not available online.

Table 2: Correlations between Twitter sexism scores and gender-based economic variables.

| **Measure** | **Correlation** | **p-value** |
|---|---|---|
| Median Annual Earnings Full-Time, Year-Round for Employed Women | -0.4337586 | 0.001473 |
| Earnings Ratio between Full-Time, Year-Round Employed Women and Men | -0.3151199 | 0.0243 |
| Percent of Women in the Labor Force | -0.453408 | 0.0008337 |
| Percent of Employed Women, Managerial or Professional Occupations | -0.3737292 | 0.006904 |
| Percent of Women with Four or More Years of College (2000) | -0.4406056 | 0.001213 |
| Percent of Businesses that are Women-Owned (2012) | 0.4388755 | 0.001275 |
| Percent of Women Living above Poverty | -0.4021554 | 0.00344 |
| Percent of Women Aged 18-64 with Health Insurance | -0.5514594 | 2.74E-05 |

The Twitter sexism scores are significantly correlated with all the measures of women's economic well-being examined (all shown in Table 2). Note that this does not establish a causal relationship; a more complicated analysis would be necessary to establish causality in either direction, and it seems more plausible in any case that the correlations seen here are due to underlying societal conditions that give rise to both sexist sentiment (as expressed on social media) and the measured economic conditions for women. However, these correlations are still important since they indicate that (1) women living in states with poorer economic outcomes for women may face not only economic challenges but also increased social challenges in the form of sexist sentiment from those around them, and (2) in the absence of other data, data like the sexism scores compiled here can serve as a rough indicator of women's well-being (one that doesn't require expensive collection of survey data, only access to the internet and a bit of computational power).

# 7 What's next

This analysis is encouraging, since it suggests that patterns of sentiment that can be detected via machine learning are significantly associated with economic variables. This motivates further investigation into the use of such sentiment variables as economic indicators. Further steps to pursue are:

1. Expand the analysis here to other sentiment variables and types of economic outcomes (beyond sexist language and women's economic status).

2. Explore causal relationships between sentiment variables and economic variables (beyond merely correlative analyses). These relationships can be relevant both in macroeconomic time series models and in various microeconomic applications.

## 7.1 Other variables of interest

The narrow focus on sexist language and women's economic status was sufficient for a preliminary analysis. However, based on previous work in this area (e.g., Bollen et al., 2011 [3]), it is reasonable to assume that other sentiment factors should be relevant for more general economic analysis. A proposed expanded list of sentiment variables to focus on is presented in Table 3.

Table 3: Other sentiment variables

| **Class** | **Specific sentiments** |
|---|---|
| General sentiment | Positive/negative affect |
| Tone | Anger, anxiety, optimism |
| Interpersonal tension | Racism, sexism |

The main restriction here is the availability of training data for calibration of machine learning classifiers. Various data sets already exist (see Kouloumpis et al., 2011 [7]), but it may be necessary to create new training data for some applications.

## 7.2 Predictive analysis in time series data

One way to explore the predictive power of this type of sentiment variable is as part of a vector autoregression (VAR) model with other economic variables. In contrast to the geographical analysis presented above, this would represent a way to study the temporal interactions of sentiment with other economic variables.

To compile sentiment time series, I (or other researchers) would build and train neural networks to classify various types of sentiment, split the Twitter data into discrete time periods (e.g. quarterly), and create aggregate sentiment scores for each period using the classifications from the neural networks. If Twitter data is collected over a significant span of time (with the maximum being the thirteen years since the site's founding in 2006) then it may be possible to identify statistically significant interdependencies between sentiment measures and external economic variables. It should then be possible to employ forecast error decompositions as a measure of the impact of the sentiment variables on those economic variables. Depending on the stationarity of the data, Granger causality tests may also provide some insights. This approach may be able to untangle a bit the arrow of causality, although, due to technical limitations on Twitter's API (probably meant to protect users' privacy), it can be a bit difficult to scrape the type of time-batched data needed here from the site.

### 7.3 Possible applications for microeconomic research

The time series analysis proposed above is applicable mostly to macroeconomic trends. However, data generated via machine learning certainly has some useful microeconomic applications as well. For example, Gentzkow, Shapiro, and Taddy used machine learning-inspired methods to measure polarization in congressional speech [6]. When deployed on content from social media, tools like those used by Gentzkow et al. or the ones I have used here could be used to measure differences in ideology in the general population and study how they interact with economic decisions and outcomes, as well as how ideology propagates through social networks. Such tools should also be generally useful for predicting group identity and may therefore lead to interesting investigation in group behavior and group outcomes.

Machine learning can also be used to measure well-being based on social media activity. Braithwaite et al. showed that Twitter users at high suicidal risk can be accurately differentiated from the general population using machine learning [4]. I was involved with the same research team in testing the usefulness of this approach for out-of-sample predictions and for identification of Twitter users at risk for other mental illnesses (e.g., eating disorders). Such tools are invaluable for economic research as they allow for easily scalable insights into the mental health of populations, serving as a valuable indicator of welfare for economic research. This approach has several important advantages:

- Using social media data allows access to large segments of the population, with different demographics favoring different platforms.[5]
- Social media sentiment data is quick and inexpensive to collect once the base model has been developed. This allows researchers to study samples of potentially millions of social media users without the need for expensive surveys.
- Social media data allows insights into fine-grained aspects of people's well-being: models can be adapted to analyze different sentiments, and sentiment for individual users and groups of users can be tracked over time.

In projects I have participated in (see, for example, Michelman et al. [9]), I have used similar methods to the ones described here to link historical records together and track economic outcomes for individuals over their lifespans.

Overall, it is clear that machine learning is a valuable tool for microeconomic analysis. In particular, the abundance of social media data makes it an attractive source of information that can serve as inputs for future economic research.

# 8 Conclusion

Machine learning is a powerful tool for identifying trends in sentiment in online data. The analysis presented here shows that measures of sexist sentiment derived from machine learning on Twitter data are complementary to existing data on sexist attitudes and are strongly correlated with measures of women's economic standing.

Further research is needed to explore causal connections and the degree to which various other measures of sentiment (beyond the sexist sentiment focused on here) can be used to draw conclusions and make predictions about conditions in the economy. The basic approach used here can serve as a foundation for various applications of machine learning in both macro- and microeconomic research.

[5] See `https://pewrsr.ch/2VxJuJ3` for the results of a Pew Research Center survey on social media usage among the U.S. population.

## Appendix 1: Interpretation of machine learning output

The machine learning model used for this project is "trained" by minimizing a binary cross-entropy loss function, defined in this case as

$$-\sum_{i=1}^{n} \left[ y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

where $n$ is the size of the training set, $y_i$ is the true value of tweet $i$ from the training set (equal to 1 if that tweet is sexist and 0 if that observation is not sexist), and $\hat{y}_i$ is the predicted value output by the neural network (ranging between 0 and 1).

Intuitively, we can think of this as a measure of how closely the empirical distribution (the distribution actually observed in the data) matches the distribution predicted by the model. It turns out that minimizing the binary cross-entropy is equivalent to finding a maximum likelihood estimator for the weights and other parameters in the neural network:

If we define $y(x)$ as the empirical probability distribution function, and $\hat{y}(x; \theta)$ as the probability distribution function of the neural network, then the likelihood function for observations $X_1, \ldots, X_n$ is

$$L(\theta) = \prod_{i=1}^{n} \hat{y}(X_i; \theta)^{y(X_i)} \left(1 - \hat{y}(X_i; \theta)\right)^{(1 - y(X_i))}$$

and the log likelihood function is

$$l(\theta) = \sum_{i=1}^{n} \left[ y(X_i) \log \hat{y}(X_i; \theta) + (1 - y(X_i)) \log(1 - \hat{y}(X_i; \theta)) \right].$$

This is just the negative of the binary cross-entropy. Hence, the $\hat{\theta}$ that minimizes the binary cross-entropy is a maximum likelihood estimator for the model parameters $\theta$. That is, the neural network with optimally-set parameters is the 'closest' among the class of neural networks with the same structure (where closeness is defined by binary cross entropy/KL divergence). With sufficiently complex neural networks and enough data, we can approximate almost any model to an arbitrary degree. [8]

Furthermore, the outputs of models trained in this way can be interpreted as Bayesian posterior probabilities (Richard & Lippmann, 1991) [12], with some simplifying assumptions. Depending on the accuracy of the model and the validity of the assumption that the observations in the training set are a random sample of the population that the tweets in this project are drawn from, it is reasonable to treat the model's output as probabilities that given tweets contain sexist content.

## Appendix 2: Sexism scores

| state | n | score | standard error |
|---|---|---|---|
| Alabama | 43420 | 0.0659 | 0.00073 |
| Alaska | 28136 | 0.0603 | 0.00087 |
| Arizona | 85335 | 0.0631 | 0.00051 |
| Arkansas | 24614 | 0.0652 | 0.00097 |
| California | 579692 | 0.0604 | 0.00019 |
| Colorado | 57261 | 0.06 | 0.0006 |
| Connecticut | 31094 | 0.0576 | 0.0008 |
| Delaware | 33827 | 0.0586 | 0.00077 |
| District of Columbia | 38821 | 0.0553 | 0.00069 |
| Florida | 231794 | 0.064 | 0.00031 |
| Georgia | 125307 | 0.0726 | 0.00046 |
| Hawaii | 27665 | 0.0607 | 0.00086 |
| Idaho | 41701 | 0.0618 | 0.00071 |
| Illinois | 130595 | 0.0611 | 0.0004 |
| Indiana | 74153 | 0.0593 | 0.00053 |
| Iowa | 31997 | 0.0594 | 0.0008 |
| Kansas | 35802 | 0.0628 | 0.00078 |
| Kentucky | 41469 | 0.0611 | 0.00072 |
| Louisiana | 54745 | 0.079 | 0.00074 |
| Maine | 50787 | 0.0508 | 0.00056 |
| Maryland | 65894 | 0.0639 | 0.00059 |
| Massachusetts | 77733 | 0.0526 | 0.00047 |
| Michigan | 86920 | 0.061 | 0.0005 |
| Minnesota | 48284 | 0.0542 | 0.00061 |
| Mississippi | 92283 | 0.0707 | 0.00053 |
| Missouri | 61803 | 0.0581 | 0.00057 |
| Montana | 32694 | 0.0603 | 0.00079 |
| Nebraska | 23395 | 0.0547 | 0.00087 |
| Nevada | 56202 | 0.061 | 0.0006 |
| New Hampshire | 65957 | 0.0511 | 0.0005 |
| New Jersey | 79200 | 0.0596 | 0.00051 |

| state | n | score | standard error |
|---|---|---|---|
| New Mexico | 59388 | 0.061 | 0.00059 |
| New York | 257985 | 0.0579 | 0.00027 |
| North Carolina | 105045 | 0.0632 | 0.00046 |
| North Dakota | 43647 | 0.0615 | 0.00069 |
| Ohio | 139082 | 0.0636 | 0.0004 |
| Oklahoma | 42193 | 0.0661 | 0.00075 |
| Oregon | 45546 | 0.0558 | 0.00064 |
| Pennsylvania | 130313 | 0.061 | 0.00041 |
| Rhode Island | 49953 | 0.0528 | 0.00059 |
| South Carolina | 45962 | 0.0634 | 0.00069 |
| South Dakota | 53050 | 0.0611 | 0.00062 |
| Tennessee | 84447 | 0.0625 | 0.00051 |
| Texas | 393190 | 0.068 | 0.00025 |
| Utah | 28018 | 0.0623 | 0.00086 |
| Vermont | 32202 | 0.053 | 0.00073 |
| Virginia | 83206 | 0.062 | 0.00051 |
| Washington | 77704 | 0.0575 | 0.0005 |
| West Virginia | 65358 | 0.0613 | 0.00057 |
| Wisconsin | 37335 | 0.061 | 0.00077 |
| Wyoming | 49899 | 0.0483 | 0.00055 |

## References


[1] Badjatiya, P., Gupta, S., & Varma, V. (2017). "Deep Learning for Hate Speech Detection in Tweets."

[2] Bierly, M. (1985). "Prejudice toward contemporary outgroups as a generalised attitude." *Journal of Applied Social Psychology*, vol. 15, pp. 189-199.

[3] Bollen, J., Mao, H., & Pepe, A. (2011). "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena." AAAI Publications, *Fifth International AAAI Conference on Weblogs and Social Media.*

[4] Braithwaite, S., Giraud-Carrier, C., West, J., Barnes, M., Hanson, C. (2016). "Validating machine learning algorithms for Twitter data against established measures of suicidality." *JMIR Mental Health.*

[5] Gentzkow M., Kelly, B., & Taddy, M. (2017). "Text as Data," NBER Working Paper No. 23276.

[6] Gentzkow, M., Shapiro, J., & Taddy, M. (2016). "Measuring polarization in high-dimensional data: Method and application to congressional speech." National Bureau of Economic Research.

[7] Kouloumpis, E., Wilson, T., & Moore, J. (2011). "Twitter Sentiment Analysis: The Good the Bad and the OMG!" *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*

[8] Liang, S. & Srikant, R. (2017). "Why Deep Neural Networks for Function Approximation?" Presented at 5th International Conference on Learning Representations (ICLR).

[9] Michelman, V., Price, P., & Zimmerman, S. (working paper). "The Distribution of and Returns to Social Success at Elite Universities." Retrieved from https://harris.uchicago.edu/files/inline-files/MPZ_Main.pdf.

[10] Mao, H., Counts, S., & Bollen, J. (2011). "Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data." Retrieved from https://arxiv.org/pdf/1112.1051.pdf.

[11] Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). "Detecting Offensive Language in Tweets Using Deep Learning"

[12] Richard, M., & Lippmann, R. (1991). "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461-463.

[13] Lansdall-Welfare, T., Lampos, V., & Christiani, N. (2012). "Effects of the recession on public mood in the UK." *Proceedings of the 21st International Conference on World Wide Web*, 1221-1226.

[14] Twenge, J. (1997). "Attitudes toward women, 1970–1995." *Psychology of Women Quarterly*, vol. 21, pp. 35–51.

[15] Waseem, Z. & Hovy, D. (2016). "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics.

[16] Wu, A. (2017). "Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum." Working paper, Princeton University - Center for Health and Wellbeing.