Using machine learning to create economically useful measures of sexist sentiment from Twitter data

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.

2 Machine learning

To build a machine learning classifier capable of recognizing sexism in tweets, I used a data set from Waseem and Hovy (2016) containing 12215 tweets, each labeled as sexist or not sexist (2949 were labeled as sexist). The data set also included labels for racism, although I chose to focus only on sexism. According to Waseem and Hovy, their criteria for labeling offensive tweets was the following:

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

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. (2017) [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.

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 (2018) [1] [2], as shown in the Table 1.

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

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 these scores are not directly comparable since the other scores were computed using classifiers that had been trained using both sexist and racist tweets, whereas my classifier was trained only for sexist tweets. The scores also change depending on the cutoff value used. 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.

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

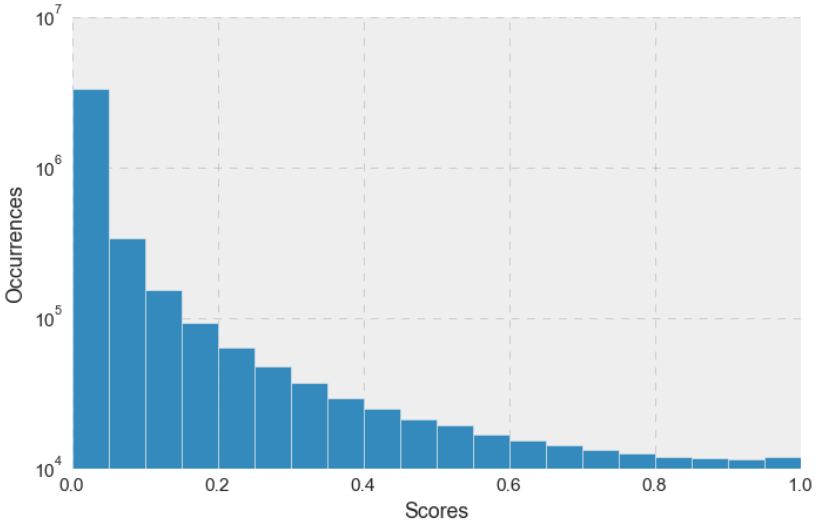
4 Data analysis

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 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).

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

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

Figure 1: Distribution of sexism scores, all 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 a clearly some states with higher rates of sexist language: Louisiana in particular 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: Sexism scores*.

When the relative scores are displayed on 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.

5 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.

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

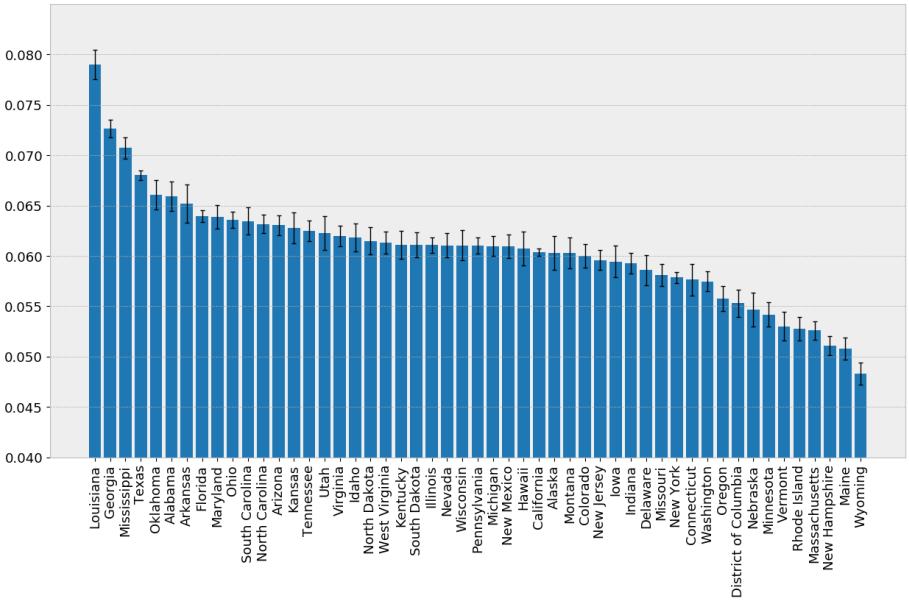
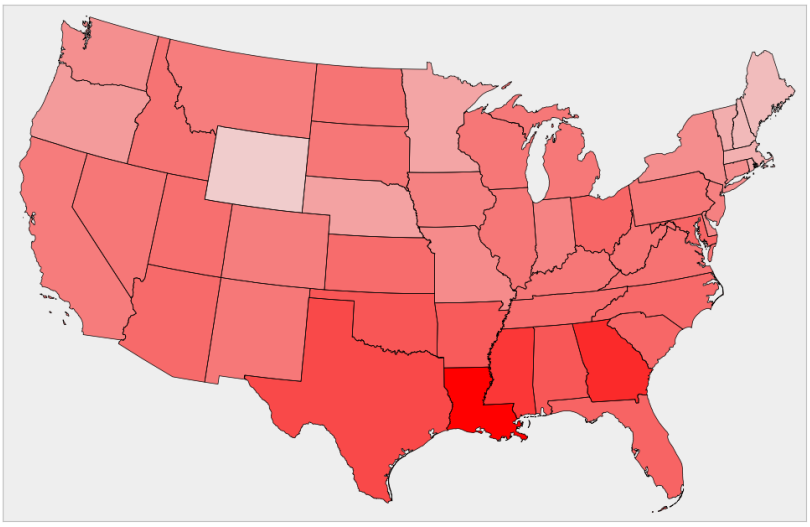


Figure 3: Heatmap of sexism scores. Darker shades of red correspond to higher Twitter sexism scores.



Using data from the Institute for Women's Policy Research (IWPR),[3] I deter-

[3]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.

mined the correlations between the Twitter sexism scores and various measures of women's economic status. The results of this correlative analysis are listed in Table 2.

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. 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.

6 Conclusion

Machine learning is a useful tool for identifying trends in sentiment in online data. The analysis presented here suggests that such sentiment data is useful not just on its own terms but as a potentially useful economic indicator. 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.

Appendix: 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
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

state	n	score	standard error
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., and Varma, V. *Deep Learning for Hate Speech Detection in Tweets*. 2017.

[2] Pitsilis, G. K., Ramampiaro, H., and Langseth, H. *Detecting Offensive Language in Tweets Using Deep Learning*. 2018.

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