STCS 6701: Foundations of Graphical Models: Final Project

Maxime TCHIBOZO (MT3390)

Monday December 14, 2020

Supervised LDA and Structural Topic Models to Measure Discriminatory Bias and Toxicity in Text

Abstract

By design, Machine Learning models are trained to maximize predictive performance on a given dataset. Models can achieve accurate predictions through incorrect associations if the data at hand is not adequately explored, interpreted and curated. Notably, machines trained on datasets containing racial, gender or disability bias may learn to internalize this bias, and perpetuate unwanted stereotypes when making predictions.

Our algorithms build on statistical methods such as Supervised LDA and Structural Topic Models trained on flagged social media comments to quantify the propensity for words or expressions to convey different types of discrimination. These interpretable methods achieve comparable results to Machine Learning alternatives on discriminatory classification tasks (>87% accuracy on gender, racial, religious and sexual orientation bias detection 7), are capable of identifying bias-charged vocabulary in unseen job descriptions (3), and could potentially automatically remove such bias (11).

Introduction

Deriving algorithms to perform an automatic flagging of toxic or discriminatory content has been an unfortunate but necessary focus of data scientists dealing with online content ((1),(2),(3),(4)). While deep learning classifiers or attention based language models are the favored solution to this challenge, the features they extract from text are rarely interpretable and may perpetuate bias from the datasets they were trained on ((5),(6),(7)). On the other hand, interpretable models offer exploratory capabilities to understand hidden and problematic stereotypes contained in data but are usually inaccurate, or require extensive labeling.

We develop a graphical model framework to create accurate and interpretable topic models capable of measuring discriminatory content in text. Using methods such as supervised LDA (sLDA (8)) and Structural Topic Models (STM (9)), we highlight the remarkable functionality offered by graphical models for subjective tasks such as online content policing. Particularly, the sLDA solves two traditionally mutually exclusive tasks: detecting which documents should be flagged with high accuracy, while identifying words susceptible of holding bias or derogatory connotations without any prior word labeling.

Background

The dataset used for training and fitting of models is the 'Jigsaw Unintended Bias in Toxicity Classification' dataset from Civil Comments containing 2M social media comments annotated with labels indicating their toxicity score (a real number between 0 and 1; Appendix.4) and the population targeted by the discriminatory language they may contain (male, female, transgender, and many more - 24 binary variables, Appendix.4). Appendix.5 illustrates examples of comments targeting specific populations. Appendix.6 details the metadata associated to one example comment which was flagged as containing female bias.

In this framework, we use only the comment text and its "type of discrimination" labels for training. Text data was minimally processed (lower casing, stopword and special character removal) and pronouns such as: 'he', 'him', 'his', 'she', "she's", 'her' or 'hers' were not removed as they could contain discriminatory bias. Model fitting and inference for sLDA models was performed in Python using the 'tomotopy' library. The 'stm' (10) package in R was used for fitting and inference of Structural Topic Models.

The presence of discriminatory text is particularly harmful in job advertisements (11). We tested our models on job postings from 22,000 US-based Job Listings from the Monster.com human resources website.

Methods

Supervised LDA topic models predict target variables using Linear Regression with latent regression parameters. Maximizing the variational objective function (ELBO) requires both learning latent topics which condense document information, and learning a latent regression parameter whose inner product with a document's topic proportions is equal to the target variable prediction (8).

STMs offers additional flexibility by incorporating metadata and exogenous covariates to evaluate their effect on documents' text (12). This allows us to answer questions such as: "How does the vocabulary for Female and Male discrimination differ from a racial discrimination vocabulary?" (Appendix.10),"Are words in a topic more associated to White or Black discrimination?" (Appendix. 11,12). sLDA and STM properties are detailed in Appendix.1,2.

Parameter	Description	Used for:
$\alpha = 0.1$	Dirichlet prior on the document-topic distributions.	STM, sLDA
$\eta = 0.01$	Dirichlet prior distribution on the topic-word distributions.	STM, sLDA
$K \in \{27, 67\}$	Number of Topics.	STM, sLDA
$\nu^2 = 2$	Variance of the regression parameters over topics.	sLDA only

Table 1: Hyperparameters/priors affect the distributions of latent parameters.

Model Fitting. We use several iterations of Box's loop to define priors, run inference, and criticize models. For the R 'stm' package (used for STM), optimization of the objective function (ELBO) done through Variational Inference, whereas Python's 'tomotopy' (used for sLDA) relies on Markov Chain Monte Carlo (MCMC). In both cases, we consider the objective has converged when its relative improvement on the previous iteration is less than 2.10^{-5} (Fig.4).

For a data-driven selection of K, we initially grid-search over values of K (Fig. 3) and evaluate the Held-Out Log-Likelihoods (13), Semantic Coherence (14), and ELBO. In a later iteration of Box's loop, we set the topic parameter to $K = 67$ using Lee and Mimno's spectral initialization method (15), which efficiently searches through hundreds of values of K in a single run.

Small Dirichlet parameter values for the topic-word distribution parameter η and the document-topic distribution parameter α generate sparse distributions (16). The large ν^2 prior on the regression parameter variance ensures both high classification accuracy and sparsity of the regression parameter, meaning a document is assigned to several classes based on only one or two topics (Fig. 8).

Model Checking. We combine quantitative and qualitative metrics for model checking of the STM on an unsupervised topic extraction task. The "meaningfulness" of inferred topics is visually (Fig. 5, Fig. 6, Fig. 7) and mathematically (Fig. 4) evaluated in the Appendix.

Results

Using model checking results from the STM to inform our model specification of sLDA, we now fully appreciate the power of a supervised topic model. For the following analysis, we restrict the Civil Comments dataset to the 94,326 comments having at least one non-zero value for the eight highest occurring binary variables indicating discrimination ('female', 'christian', 'male', 'white', 'muslim', 'black', 'homosexual_gay_or_lesbian' and 'jewish').

Accuracy and Dimensionality Reduction. We train an sLDA model ($\alpha = 0.1, \eta = 0.01, \nu^2 = 2, K = 67$) on 80% of the restricted dataset and test on the remaining 20%, randomly selecting comments and predicting the eight discrimination variables simultaneously. This sLDA model achieves scores of at least 87.3% accuracy on the test set for each of the eight variables (Table. 7). While it does note achieve state-of-the-art accuracy on discriminatory text classification, sLDA successfully condenses each document into a 67 dimension topic vector, reducing the dataset of over 40,000 words into 67 features, while methods used for comparison require at least 1,000 features.

Interpretable Predictions. sLDA gives us the ability to infer a generative story which explains the observations.

Document	Homosexuals are NOT pedophiles. The pedophiles in the catholic church are just that: pedophiles. Their sexual orientation doesn't indicate their likelihood to be a pedophile no more than your sexuality does.
Predictions	[0.000, 0.497, 0.012, 0.00, 0.00, 0.00, 1.00, 0.002]
Potential Types of Discrimination	Homosexual_gay_or_lesbian (class 7, 100%) Christian (class 2, 49.7%)
Most important topic	Topic 57
Regression Parameter with largest value for Most important topic	Homosexual_gay_or_lesbian
Highest likelihood words with their most likely topic assignment	pedophiles: 57, homosexuals: 57 orientation: 57, pedophiles: 8 pedophile: 8

Table 2: Illustration of how our model gives interpretable explanations to justify discrimination label assignments.

Note that the comment itself is not discriminatory. Since it contains topics usually associated to Homosexual_gay_or_lesbian or Christian discrimination, the sLDA assigns a high probability of potential discrimination for both classes.

We interpret the above example as follows: the algorithm assigns class 7 (Homosexual_gay_or_lesbian) to the comment with probability 100%, and almost assigns class 2 (Christian - probability 49.7%). For conciseness, we have not displayed the topic proportions (67-dimensional vector, K=67), but through figures similar to Fig.8 we find that the regression parameter for class 7 (homosexual_gay_or_lesbian discrimination) is maximal for the 57th topic. Topic 57 is therefore determinant in the assignment of this document to the Homosexual_gay_or_lesbian class. We look at the words most representative of Topic 57 and find that three words in the document are among the highest probability words in this topic. We can confidently deduce that the document was flagged as potentially homophobic because of these three words. We display additional examples of such assignment interpretations in Appendix.8.

Model Performance on the job postings dataset. We use the sLDA model trained on 'Jigsaw' social media comments to identify potentially biased words in 22,000 job postings from Monster.com

Job Posting	Predicted	Bias Words
"Are you a proven Recruiting Manager who has managed staffing teams and started up a recruiting department from scratch? If this sounds like you, please..."	Male	teams (topic 31) managing (41) recruiting (31)
"Progressive Design, Inc. is a Heavy Industrial Engineer and Design Consulting Firm. We are currently looking for Mechanical Designer/Machine Designer for a..."	Female	engineers (41) design (17) drawings (41)
"Are you looking to advance your career with one of the top Finance Companies in the United States? Please apply today for the Administrative Assistant position..."	Christian	prepares (36) administrative (36) experience (36)
"HIRING - PART TIME ALTERATIONS SPECIALISTA, we empowerour customers and our employees to stay true to their dreams and find the one, whether that means the wedding..."	Female	her (topic 10) garment (topic 61)
"Experis is working with a Pharmaceutical start up on a Direct Hire Sr. Process Engineer role opportunity. My client is investing heavily in the business, has a strong leadership team, a culture that fosters new ideas,..."	Male	equipment (63) strong (topic 41)

Table 3: This framework can also find potentially biased words in job postings.

Regression Parameter Analysis. sLDA Regression parameters capture information about the types of discrimination contained in the 'Jigsaw Dataset'. Each document prediction is the inner product of the document's topics (global) with its regression parameter (discrimination-specific). Fig.8 shows that regression parameters are sparse, but that different types of discrimination give importance to the same topics. This indicates overlap (or similarity) between the vocabularies of certain types of discrimination. STM is the ideal model to compare the effects of different treatment assignments on the topic vocabularies, but

STM works best when two types of discrimination are compared at a time. We use sLDA regression parameters to inform which pairs of variables are most similar and focus our analysis on similar pairs of variables identified by sLDA.

Measuring Overlap Between Discrimination Types. In Fig. 9, we use the cosine similarity of two regression parameters over topics as a measure of similarity between the discrimination categories. This analysis shows that the following pairs of variables give importance to similar topics: (white,black), (muslim,jewish), (christian,female). This indicates that similar words are used to discriminate pairs of subjects (e.g. White and Black). On the other hand, Male and Female discrimination regression parameters appear uncorrelated, and these classes are not discriminated against with similar words.

Exploratory Data Analysis of the similar biases with STM. We further analyze the (white,black), (muslim,jewish), (christian,male) pairs of variables which had high topic similarity, and also analyze the (male,female) pair which had a cosine similarity of 0 using STM. For each pair of bias classes, we answer the following questions: Which are the overlapping words between the two classes? How close are they to either class? Which are the words specific to each class? Figures 11 & 12 illustrate how we can use a treatment effect framework to answer the first two questions using the STM. Figure 10 illustrates how we can use a treatment effect framework to answer the third question using an STM.

Debiasing Strategies. Regression variable parameters can have negative values for specific topics, suggesting these topics "reduce" bias for a discrimination type. While this phenomenon is primarily due to correlation between topics (17), it opens possibilities for suggesting alternatives to biased vocabulary. Biased words could be replaced by words from "debiasing" topics, but this can introduce new types of discrimination (Fig.9). We counter this effect by constraining replacement candidates to words which do not create bias elsewhere (Fig.10). Unfortunately, these candidates do not hold much meaning in the initial context. We suggest combining the "debiasing" with a synonym-replacement system to ensure replacement words fit in the initial context 11.

Conclusion

The benefits of topic models are best highlighted on the job postings dataset: they perform a data-driven identification of discriminatory words, quantify the amount of bias each word adds to a document, and craft a story explaining which words might have caused content to be censored. A natural extension of this work is the identification of strategies which remove bias from documents without creating unwanted discrimination elsewhere. Causal Inference emerges as a worthy candidate to address this challenge.

Appendix: Jigsaw ’Unintended Bias in Toxicity Classification’ dataset

Variable Type	Variable Name	data type
Toxicity	’toxicity’, ’severe_toxicity’, ’obscene’, ’sexual_explicit’, ’identity_attack’, ’insult’, ’threat’	float ∈ [0,1]
Discrimination	’male’, ’female’, ’transgender’, ’other_gender’, ’heterosexual’, ’homosexual_gay_or_lesbian’, ’bisexual’,’other_sexual_orientation’, ’christian’, ’jewish’, ’muslim’, ’hindu’, ’buddhist’, ’atheist’, ’other_religion’, ’black’, ’white’, ’asian’, ’latino’, ’other_race_or_ethnicity’,’physical_disability’, ’intellectual_or_learning_disability’, ’psychiatric_or_mental_illness’,’other_disability’	Binary ∈ {0,1}
Text content	’comment_text’	String
Metadata	’id’, ’split’, ’created_date’, ’publication_id’, ’parent_id’, ’article_id’, ’rating’, ’funny’, ’wow’,’sad’ ’likes’,’disagree’, ’identity_annotator_count’, ’toxicity_annotator_count’	String, Date, int

Table 4: Description of the Jigsaw dataset variables

Discrimination Type	Biased Comment Example
Intellectual or learning disability	"she’s a retard"
White	"Just admit that the real reason that you and your ilk are upset is that you hate non-white immigration. You’d be flying high as a fairy if it were white europeans instead."
Jewish	"Trump is an honourary Jewish President 3 kids 2 married Jewish 1 married half Jewish Goodbye Obama’s anti-Israel - not very subtle"
Black	"Another rabidly racist post from an angry black male with a victim mentality looking for reparations."
Female	"Thunder thighs??? Dude, you just insulted half the women in Alaska! LOL At LEAST half of them. ROFLMAO"
Homosexual, gay or lesbian	"What I have noticed is that people who make comments like this are usually closet case HOMOSEXUALS."

Table 5: Examples of Jigsaw dataset comments with their associated discrimination types.

Id	5137126
comment_text	Blame men. There’s always an excuse to blame men for the failures of women.
split	train
created_date	2017-04-15 19:00:45.032674+00
publication_id	54
parent_id	5.13691e+06
article_id	327125
rating	rejected
funny	0
wow	0
sad	0
likes	0
disagree	0
toxicity	0.545455
severe_toxicity	0
obscene	0
sexual_explicit	0
identity_attack	0.636364
insult	0.272727
threat	0
male	1
female	1
transgender	0
other_gender	0
heterosexual	0
homosexual_gay_or_lesbian	0
bisexual	0
other_sexual_orientation	0
christian	0
jewish	0
muslim	0
hindu	0
buddhist	0
atheist	0
other_religion	0
black	0
white	0
asian	0
latino	0
other_race_or_ethnicity	0
physical_disability	0
intellectual_or_learning_disability	0
psychiatric_or_mental_illness	0
other_disability	0
identity_annotator_count	4
toxicity_annotator_count	11

Table 6: All the metadata associated to a single comment/document.

Appendix: Methods, sLDA, STM

sLDA Generative Story

1. Draw topic proportions $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   (a) Draw topic assignment $z_n \mid \theta \sim \text{Mult}(\theta)$
   (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \text{N}\left(\eta^T \bar{z}, \sigma^2\right)$, where $\bar{z} := (1/N) \sum_{n=1}^{N} z_n$

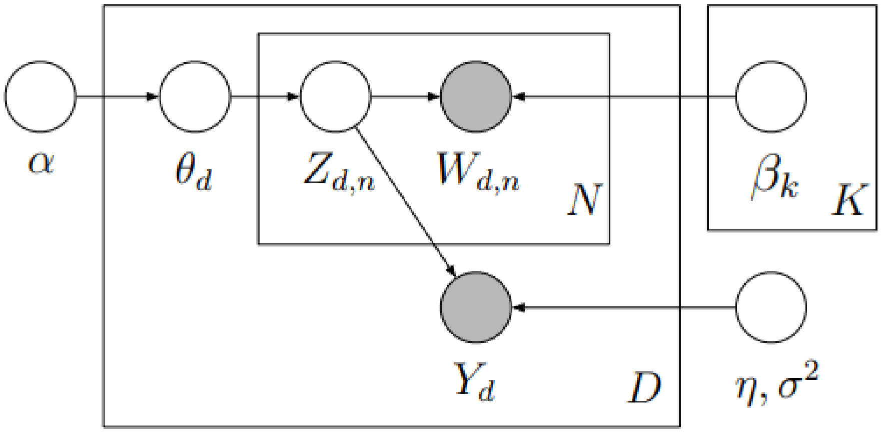


Figure 1: sLDA Plate Notation (8)

STM design

Adapted from the original work (12): "The model (Fig. 2) combines and extends three existing models: the correlated topic model (CTM), the Dirichlet-Multinomial Regression (DMR) topic model and the Sparse Additive Generative (SAGE) topic model [...]. The design matrix for the covariates X allows for arbitrarily flexible functional forms of the original covariates using radial basis functions. The distribution over words is replaced with a multinomial logit such that a token's distribution is the combination of three effects (topic, covariates, topic-covariate interaction) operationalized as sparse deviations from a baseline word frequency (m)."

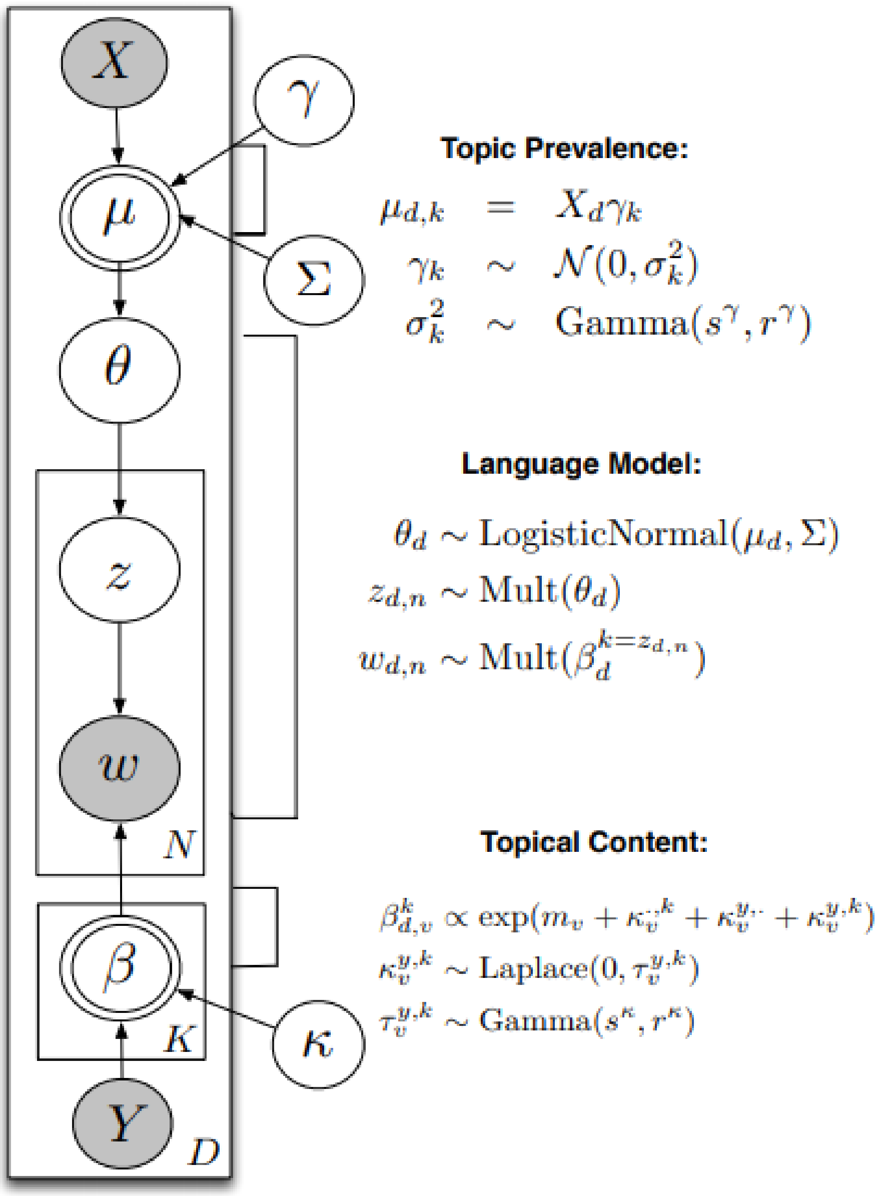

X
γ
μ
Σ
θ
z
w
N
β
K
κ
Y
D
Topic Prevalence:
$\mu_{d,k} = X_d \gamma_k$
$\gamma_k \sim \mathcal{N}(0, \sigma^2_k)$
$\sigma^2_k \sim \text{Gamma}(s^\gamma, r^\gamma)$
Language Model:
$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$
$z_{d,n} \sim \text{Mult}(\theta_d)$
$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$
Topical Content:
$\beta^k_{d,v} \propto \exp(m_v + \kappa^{\cdot,k}_v + \kappa^{y,\cdot}_v + \kappa^{y,k}_v)$
$\kappa^{y,k}_v \sim \text{Laplace}(0, \tau^{y,k}_v)$
$\tau^{y,k}_v \sim \text{Gamma}(s^\kappa, r^\kappa)$

Figure 2: STM Plate Notation (8). This model combines three separate modules: a Correlated Topic Model, a Dirichlet-Multinomial Topic Model, and a Sparse Additive Generative topic model. The STM assumes different latent parameters for each value of the metadata "treatment assignment" covariate X. The covariate-specific parameters are then compared to quantify the difference in outcome between the different groups assigned by X.

Appendix: Results Tables and Figures

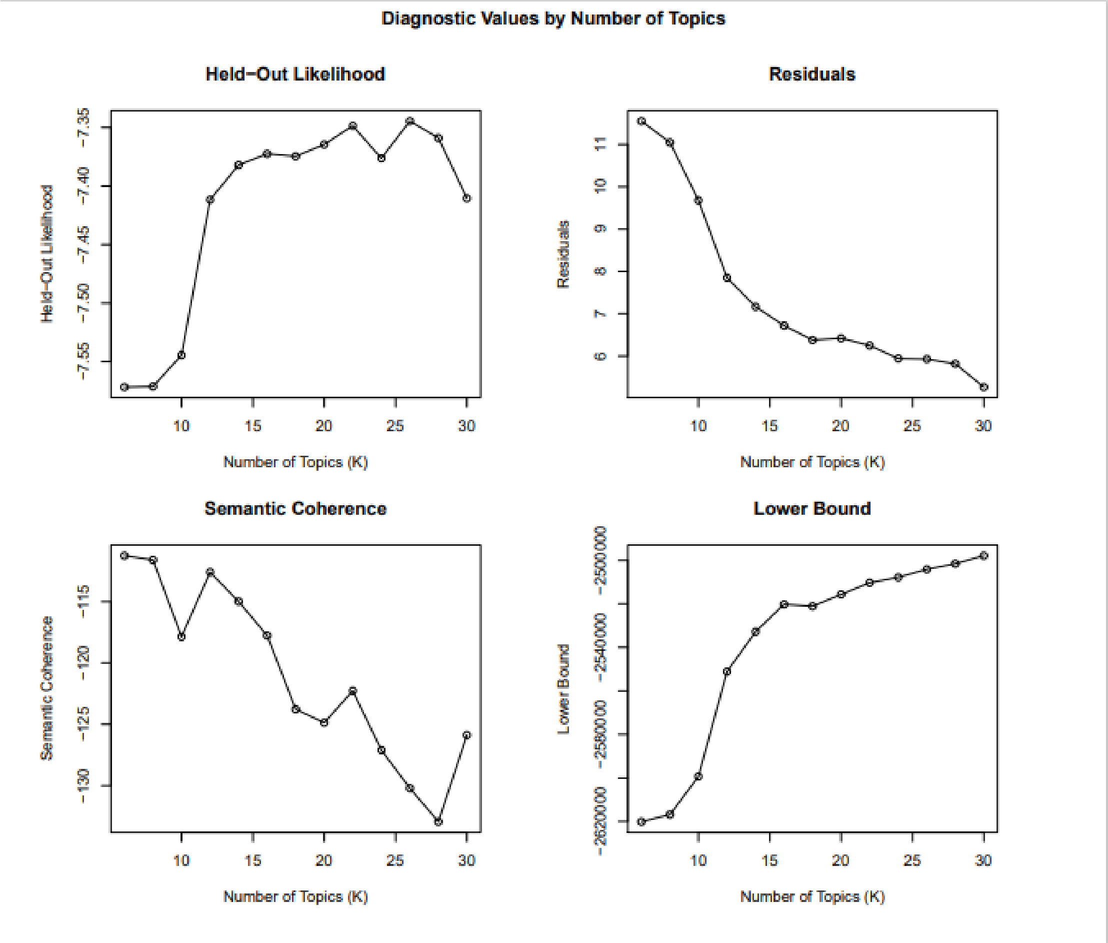
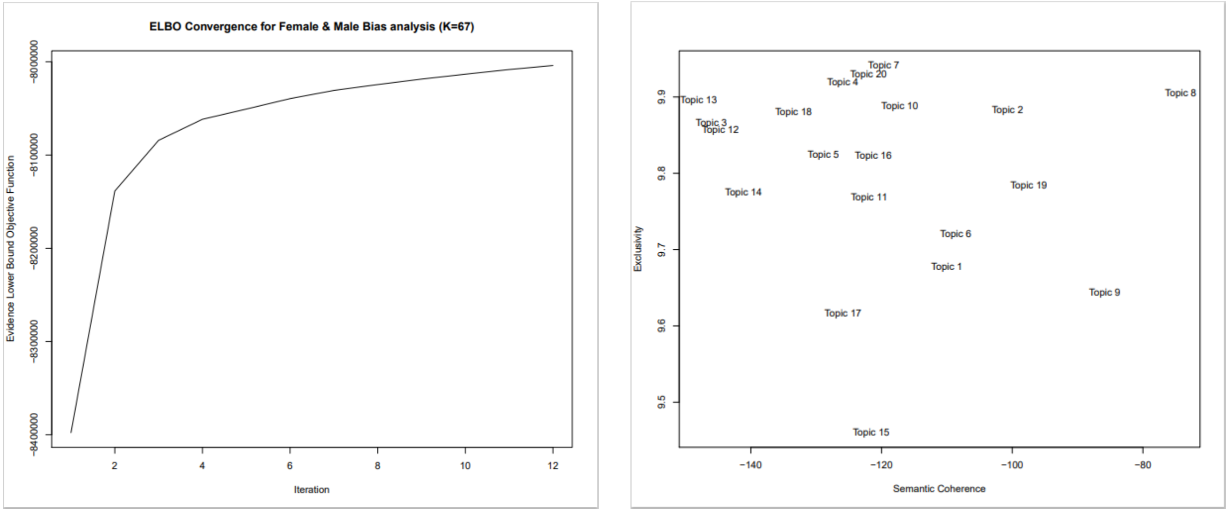


Figure 3: **Optimizing K**: We assess four quality criteria (Held-Out Likelihood, Residual distribution, Semantic Coherence and Evidence Lower Bound).



(a) Convergence of the STM ELBO on the Jigsaw dataset. First 12 iterations (K=67).

(b) Data-driven evaluation of quality for each topic

Figure 4: **Model Checking.** ELBO plots, semantic coherence and topic exclusivity are quantitative diagnostics for topic models such as STM and sLDA.

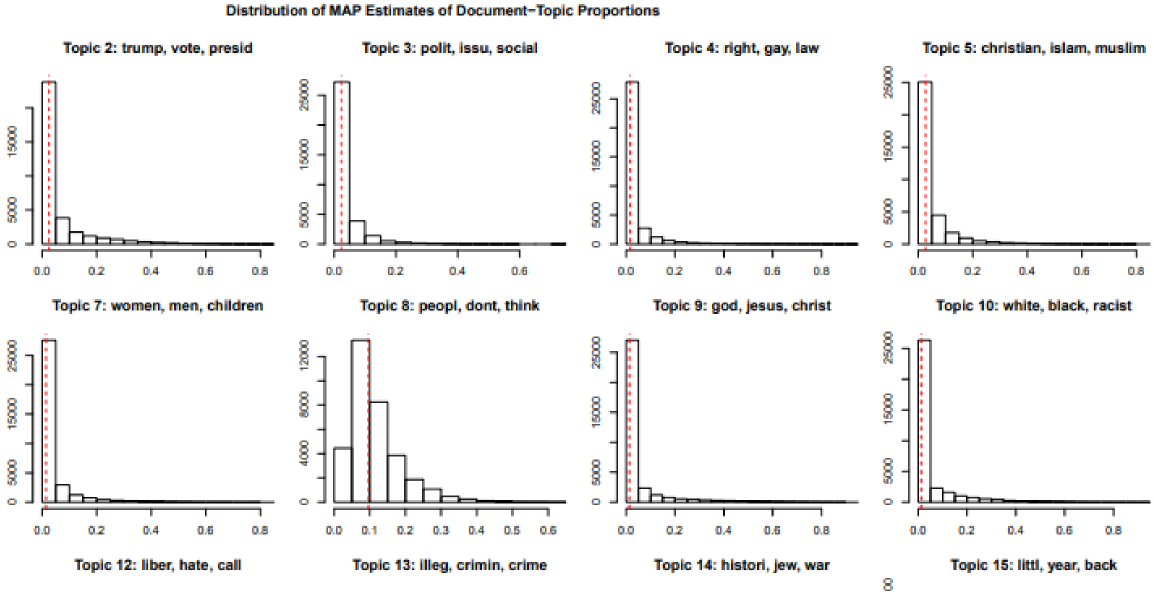
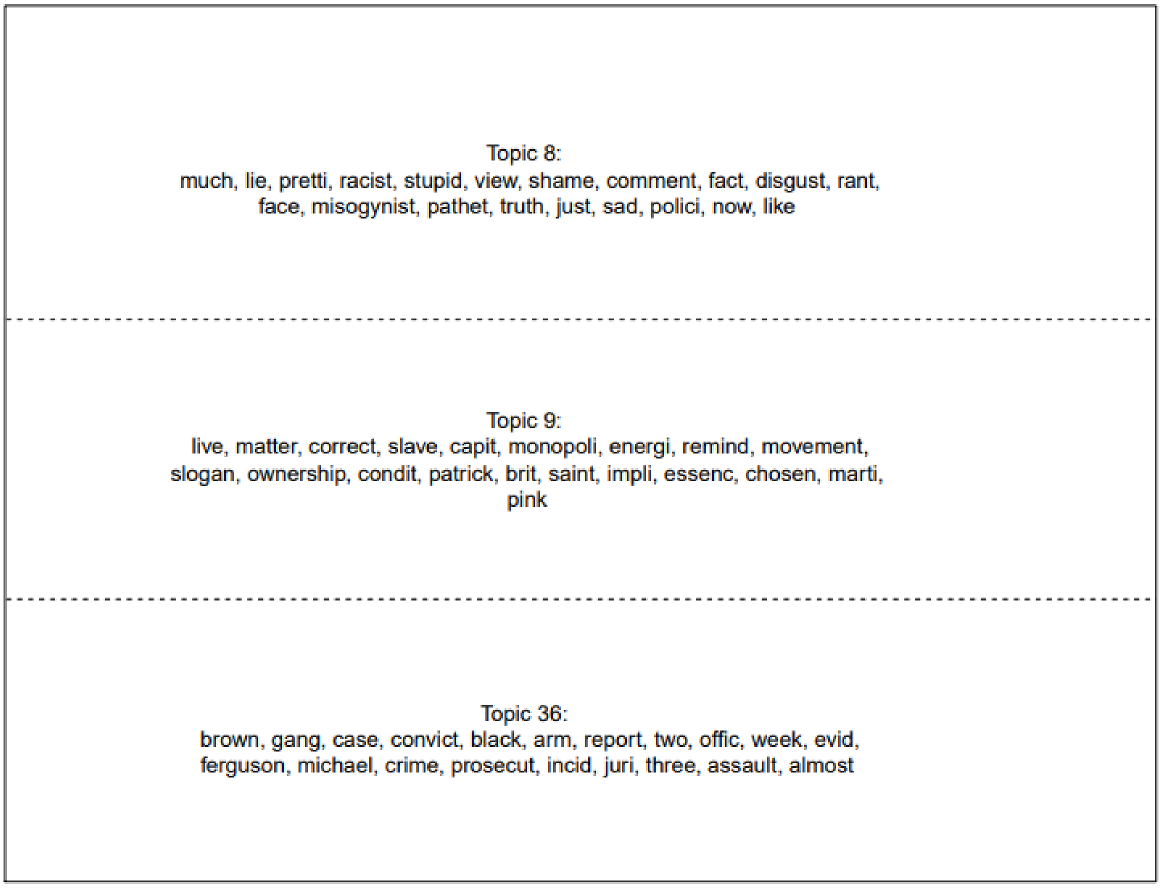


Figure 5: **Model Checking.** MAP assignment of documents to topics. Each document is represented by a mixture of topics. Some topics appear much more common than others. This plot was drawn for an STM with K=67 topics, but we only display 8 topics for visualization purposes.



Topic 8:
much, lie, pretti, racist, stupid, view, shame, comment, fact, disgust, rant, face, misogynist, pathet, truth, just, sad, polici, now, like

Topic 9:
live, matter, correct, slave, capit, monopoli, energi, remind, movement, slogan, ownership, condit, patrick, brit, saint, impli, essenc, chosen, marti, pink

Topic 36:
brown, gang, case, convict, black, arm, report, two, offic, week, evid, ferguson, michael, crime, prosecut, incid, juri, three, assault, almost

Figure 6: **STM Model Checking.** Words with highest probability within 3 specific topics. This method is the most common way of checking the quality of a topic model such as LDA, sLDA or STM. Topics 8, 9 and 36 are displayed here as they all seemed to relate to racism, police brutality and violence.

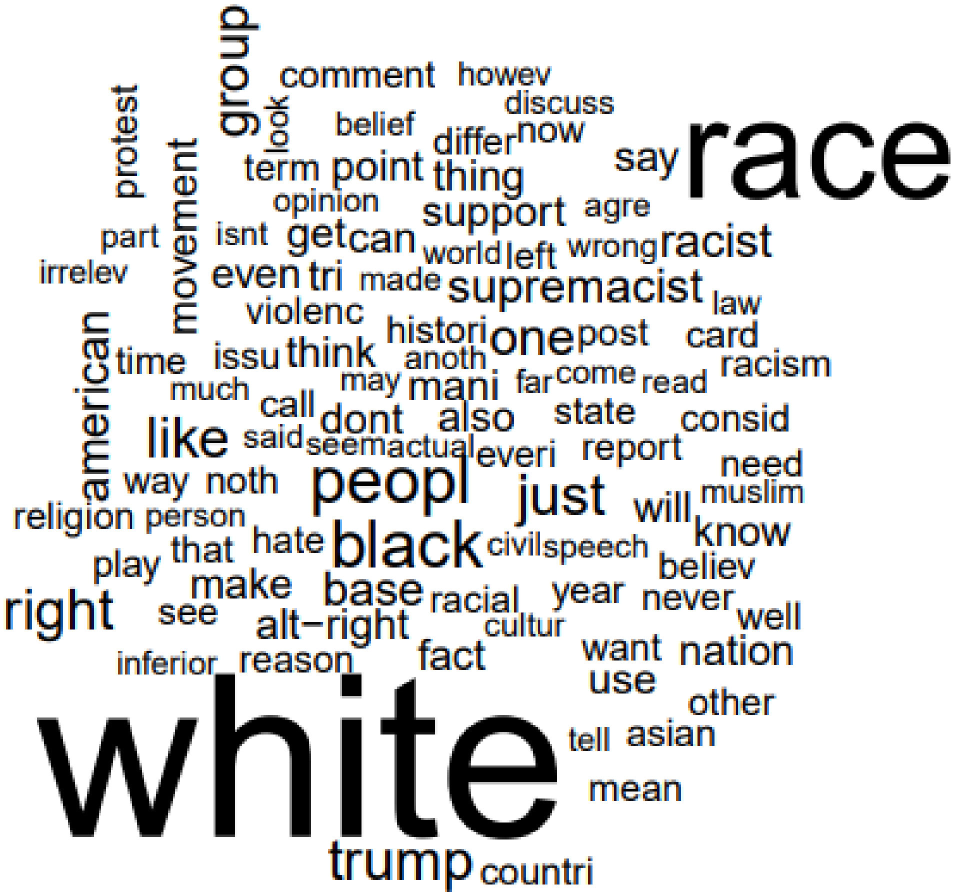


Figure 7: **STM Model Checking.** Wordcloud representation of the words most representative of topic #52 (K=67). This topic clearly addresses black and white racism. For this figure, we use the entire 'Jigsaw' dataset restricted to documents for which at least one the eight most prevalent discrimination classes was present (female, christian, male, black, jewish, white, muslim). STM automatically grouped these words together by considering only the binary bias type covariates and the text for each document. Although STM is an unsupervised method, the covariates inform the topics for a more meaningful clustering.

	sLDA	Log. Reg.	Rand. Forest	XGBoost	SVM
Female	87.3%	97.6%	95.2%	98.2%	95.7%
Christian	90.2%	96.6%	63.3%	97.3%	94.6%
Male	86.3%	95.5%	77.8%	96.3%	93.3%
White	93.5%	98.3%	94.5%	98.9%	96.4%
Muslim	95.3 %	98.9%	91.4%	99.1%	97%
Black	95.4%	98.7%	91.2%	99.1%	96.7%
Homosexual, gay or lesbian	96.4%	98.9%	92.8%	99.4%	97.1%
Jewish	96.7 %	98.8%	95.1%	99%	97.1%

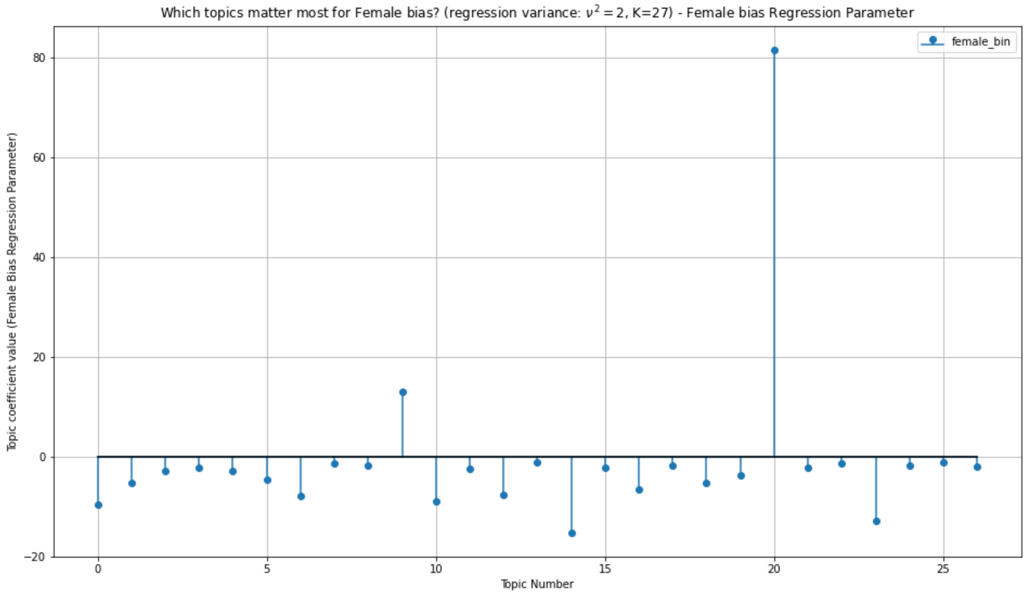
Table 7: **Results.** The sLDA Classification accuracy on the test set of the 'Jigsaw' dataset is comparable to the accuracy of other models when predicting the 8 most prevalent categories of discrimination. The sLDA uses K=67 features to represent each document, whereas for the other methods we use a CountVectorizer with the 1000 most common features to represent each document. Since 8 classes are predicted simultaneously, we use One-vs-Rest classification for Logistic Regression, Random Forest, XGBoost and SVM. sLDA is naturally capable of predicting all 8 classes simultaneously.

Document	Fogel did not mention the two-state solution because it is not a solution. All it will do is compromise Israel’s security and give the Palestinians military sovereignty to load up and attack Israel in the Arabs’ ongoing objective to eliminate Israel and eradicate Jews from the region.
Predictions	[0.00, 0.00, 0.00, 0.00, 0.11, 0.00, 0.00, 1.00]
Potential Discrimination	Jewish (100%)
Most important topic	Topic 52
Regression Parameter with largest value for Most important topic	Jewish
Highest likelihood words with their most likely topic assignment	solution (52) israel (52) israel (32)

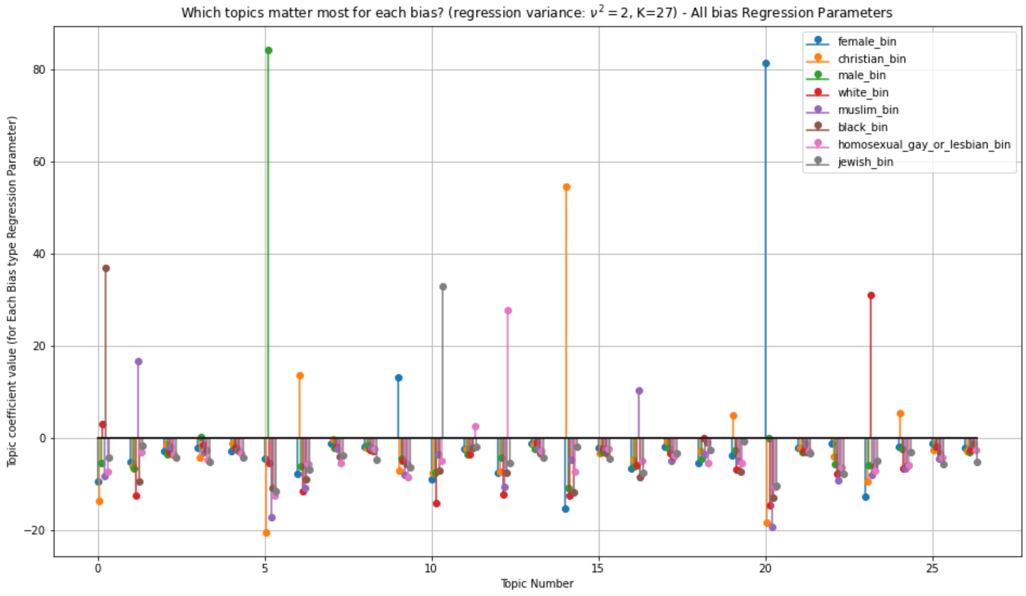
Document	Not only a Hawaii thing, majority of black kids have a fatherless home.
Predictions	[0.28, 0.04, 0.30, 0.10, 0.00, 0.68, 0.01, 0.02]
Potential Discrimination	Black (68%)
Most important topic	Topic 66
Regression Parameter with largest value for Most important topic	Black
Highest likelihood words with their most likely topic assignment	kids (33) home (33) black (66)

Document	"white trash"! Wow! If black was substitute for white the term would be seen as racist. But then again only whites can be racist, right?
Predictions	[0.00, 0.00, 0.02, 1.00, 0.00, 0.99, 0.00, 0.00]
Potential Types of Discrimination	White (100%) Black (99%)
Most important topic	Topic 7, Topic 0
Regression Parameter with largest value for Most important topic	White (topic 7) Black (topic 0)
Highest likelihood words with their most likely topic assignment	racist (7,0) white (7,0) trash (7)

Table 8: **Results.** Examples drawn from the ’Jigsaw’ dataset containing potentially discriminatory vocabulary (K=67). Prediction corresponds to the probability of containing discrimination related to the following populations: [Female, Christian, Male, White, Muslim, Black, Homosexual_gay_or_lesbian, Jewish]. Classes are not mutually exclusive. Each probability is the product between the same topic proportion vector and a class-specific regression parameter.



(a)



(b)

Figure 8: **Results.** Coefficients of the regression parameters for each topic. For visualization purposes, we use K=27 topics. As expected, regression parameters are sparse meaning only a few topics matter in assigning a document to discrimination variables.

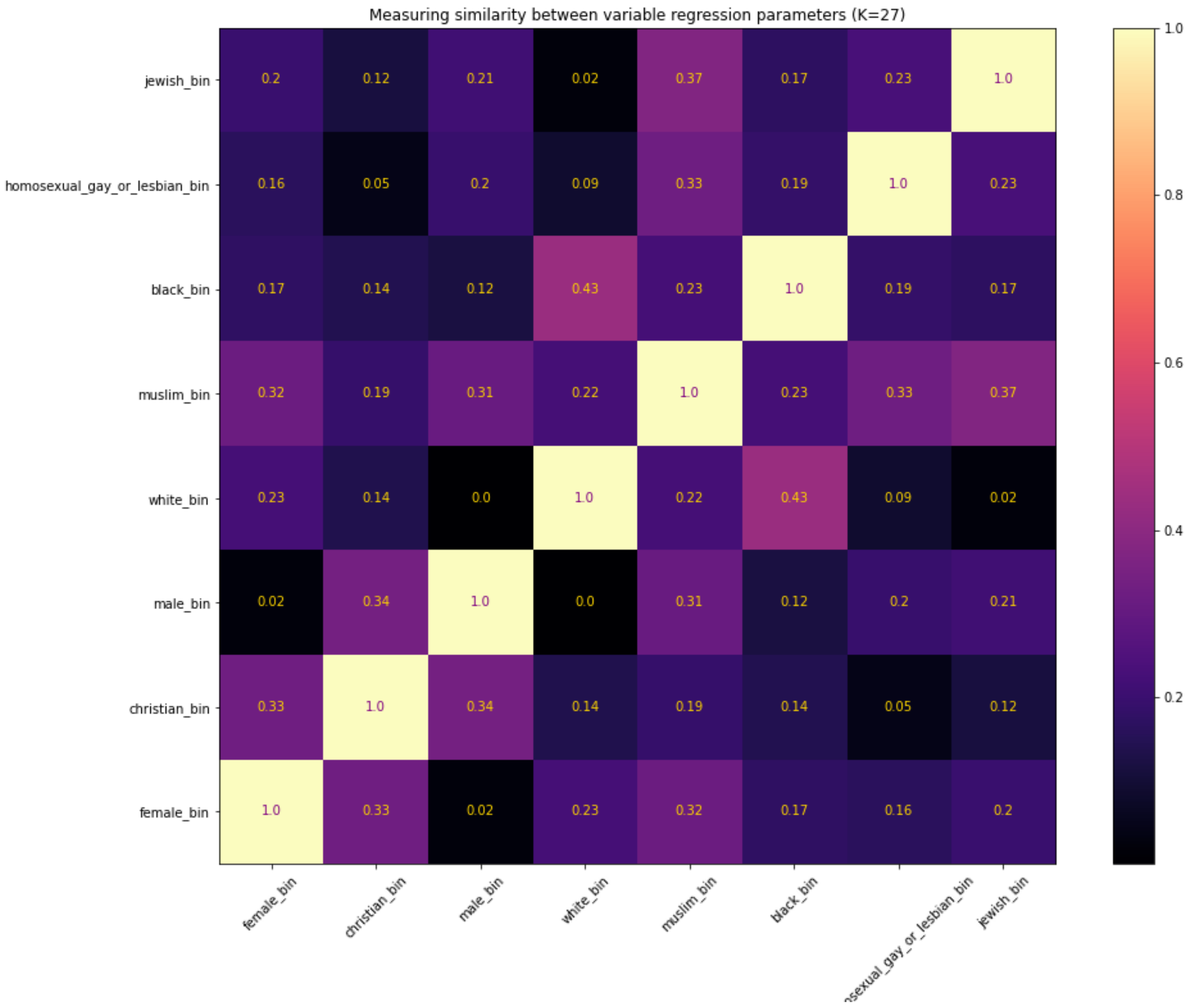
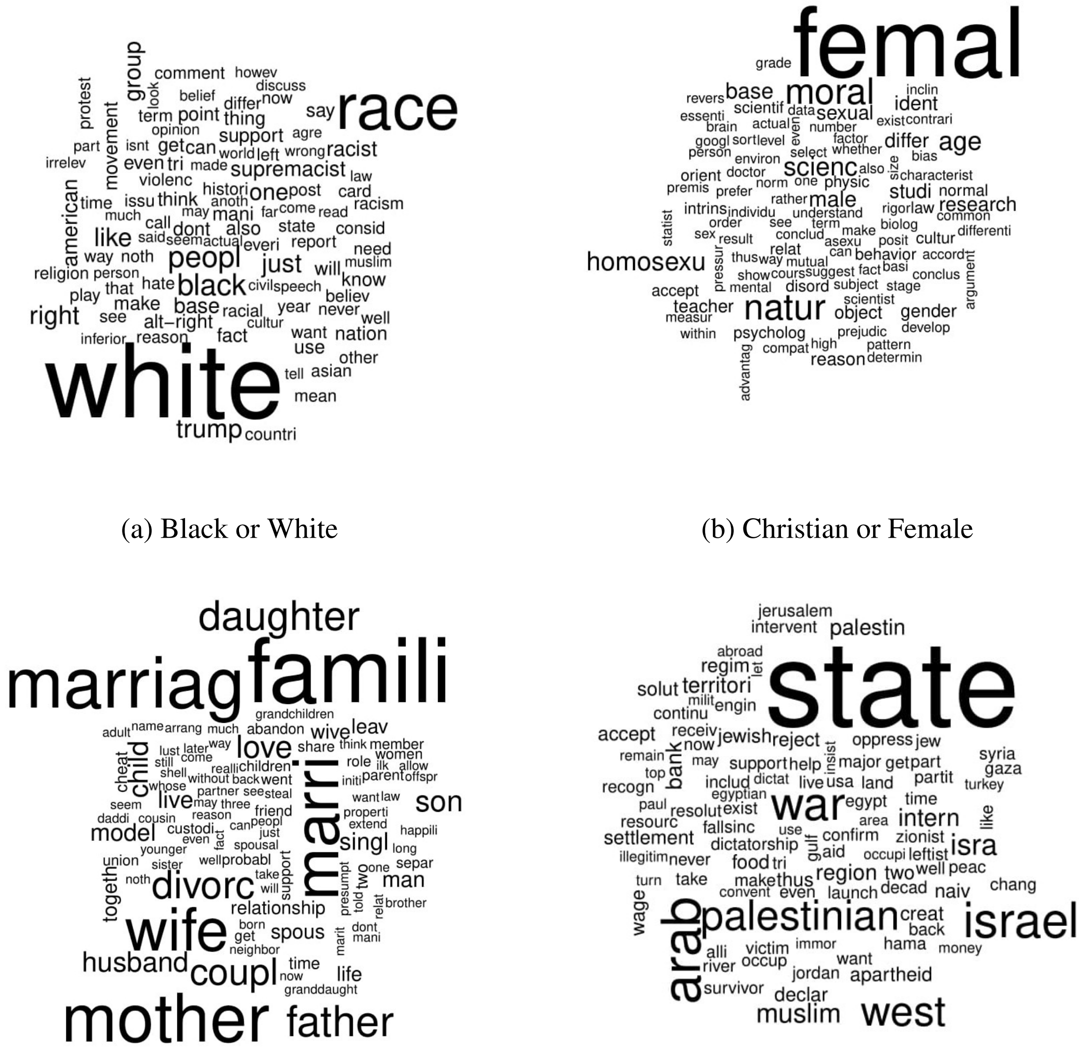


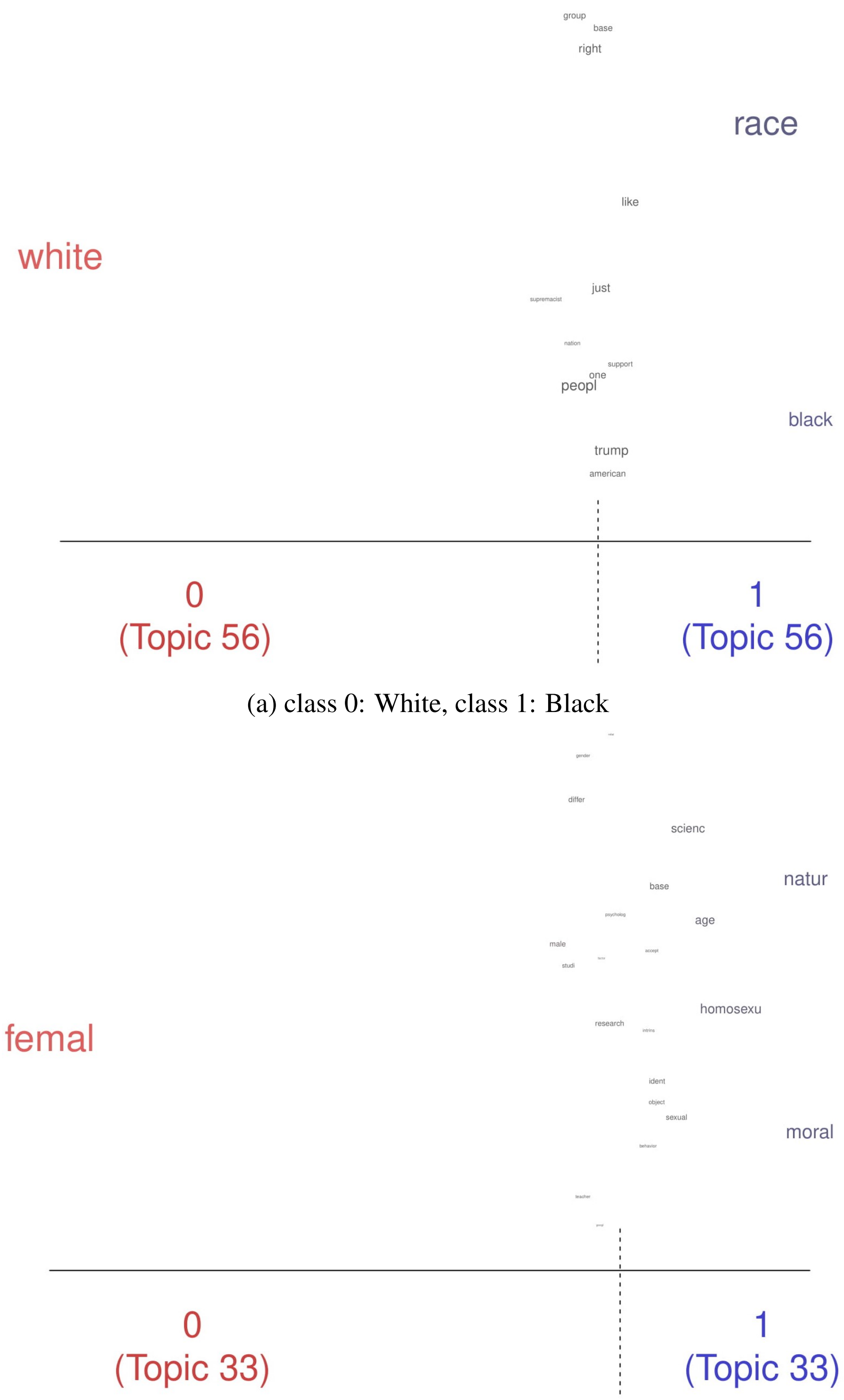
Figure 9: **Results.** Cosine similarity matrix between the sLDA regression parameters for each discrimination variable (k=27). The (white,black), (muslim,jewish) and (christian,male) pairs have high cosine similarity, whereas the (female,male) pair has a low cosine similarity. Similarity between regression parameters indicates that two types of discrimination give importance to similar topics.



(c) Female or Male

(d) Muslim or Jewish

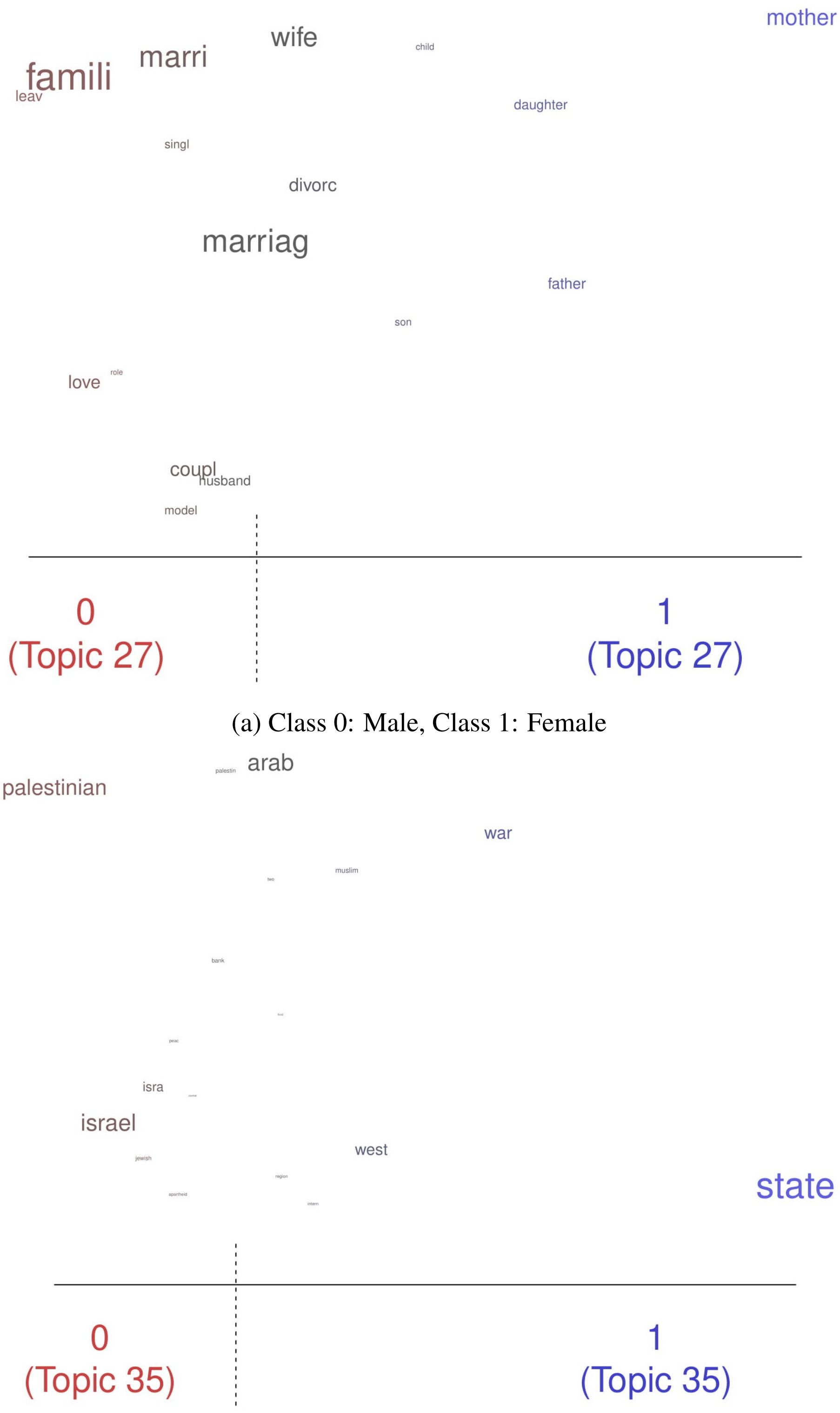
Figure 10: Different types of discrimination use different vocabularies, and mention different themes. STM Wordclouds resulting from different "treatment assignments" to one of two possible biases. For example, "Black or White" indicates the dataset was restricted to documents containing exclusively Black or White discrimination. This graph illustrates the effect of a bias variable "treatment assignment" on the vocabulary contained in the documents. The choice of these pairs of biases was determined by the cosine similarity between the sLDA regression parameters of the associated variables.



(a) class 0: White, class 1: Black

(b) class 0: Female, class 1: Christian

Figure 11: An STM for White or Black discrimination "treatment assignments" generated topic 56. This topic contains vocabulary leaning more towards Black discrimination (class 1). An STM with Female or Christian discrimination "treatment assignments" generated topic 33. This topic contains vocabulary leaning more towards Christian discrimination (class 1).



(a) Class 0: Male, Class 1: Female

(b) Class 0: Jewish, Class 1: Muslim

Figure 12: An STM with Male or Female discrimination "treatment assignments" generated topic 27. This topic contains vocabulary leaning more towards Male discrimination (class 0). An STM with Jewish or Muslim discrimination "treatment assignments" generated topic 35. This topic contains vocabulary leaning more towards Jewish discrimination (class 0).

Discrimination	Most biased coef.	Top bias words	Most debiasing coef.	Top debias words
Female	69.3 (topic 41)	women men woman	-17.4 (topic 7)	white people racist
Christian	53.5 (topic 25)	christian christians christianity	-23.2 (topic 60)	man men male
Male	82.0 (topic 60)	man men male	-10.7 (topic 60)	muslims muslim islam
White	60.7 (topic 7)	white people racist	-12.0 (topic 41)	women men woman
Muslim	54.1 (topic 12)	muslims muslim islam	-15.7 (topic 60)	man men male
Black	58.0 (topic 0)	black blacks people	-12.2 (topic 41)	women men woman
Homosexual_ gay_or_ lesbian	55.7 (topic 57)	gay marriage sex	-12.7 (topic 60)	man men male
Jewish	40.3 (topic 52)	jewish jews israel	-11.9 (topic 60)	man men male

Table 9: **1st debiasing strategy**. For each discrimination type, we observe the largest (most biased) and smallest (most debiasing) coefficients of the regression parameter, and view the 3 words which most belong to the associated bias and debiasing topics. Directly removing bias for one type of discrimination increases bias for another type of discrimination. For example, replacing "women" by "white" may remove Female bias, but will add White bias.

2nd debiasing strategy. For each discrimination type $disc.$ (e.g. $disc.$="Female", or $disc.$="Jewish", etc.), we multiply each topic's distribution over words by the $disc.$ regression parameter value for the same topic. This gives a vector of length $\mid V \mid = 88022$ (length of vocabulary) containing the weighted average bias score for each word with respect to $disc.$ denoted by $\Phi^{(disc)} = \{\phi_w^{(disc.)}\}_{w=1}^{88022}$.

$$\Phi^{(disc.)} = \begin{pmatrix} \phi_1^{(disc.)} \\ \\ \ldots \\ \\ \phi_{88022}^{(disc.)} \end{pmatrix} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,67} \\ \cdot & \vdots & \cdot \\ p_{88022,1} & \cdots & p_{88022,67} \end{pmatrix} \circ \begin{bmatrix} w_1^{(v)} & \ldots & w_{67}^{(v)} \end{bmatrix} = \begin{pmatrix} \sum_{k=1}^{67} p_{1,k} \cdot w_k^{(v)} \\ \ldots \\ \sum_{k=1}^{67} p_{88022,k} \cdot w_k^{(v)} \end{pmatrix}$$

Using this measure for word bias over a specific type of discrimination $disc.$, we define the bias score for words by: $\Delta_w^{(disc.)} = \frac{\phi_w^{(disc)}}{\sum\limits_{t \neq disc.} \phi_w^{(t)}}$.

This measure will give a high score to words which have bias for the $disc.$ category but not for other types of discrimination. If this measure is negative, this indicates a word debiases the $disc.$ category without creating bias elsewhere.

Discrimination	Most biased coef.	Max bias words	Min. Δ-debias val.	debias word
Female	69.3 (topic 41)	women men woman	-67103.5 -10218.8 -7760.2	wearer though misoginy
Christian	53.5 (topic 25)	christian christians christianity	-5582.6 -5537.1 -1469.2	pressed flout stinks
Male	82.0 (topic 60)	man men male	-2731.5 -1701.5 -1324.8	stragglers flout winnable
White	60.7 (topic 7)	white people racist	-6248.5 -3270.1 -3192.9	pressed flout stragglers
Muslim	54.1 (topic 12)	muslims muslim islam	-7193.5 -5603.5 -2571.2	flout pressed stragglers
Black	58.0 (topic 0)	black blacks people	-4567.5 -4053.4 -3381.7	pressed stressed pooh
Homosexual_ gay_or_ lesbian	55.7 (topic 57)	gay marriage sex	-6221.6 -2851.0 -1913.0	flout stragglers stinks
Jewish	40.3 (topic 52)	jewish jews israel	-6779.7 -5614.7 -2167.6	flout pressed stragglers

Table 10: **2nd debiasing strategy**. For each discrimination type, we observe the 3 most biased words and the three most debiased words. Debiased words are the most negative according to the Δ measure defined above. The debiasing words identified by this strategy do not create new bias, but are not very similar to the original word.

Discrimination	Most biased coef.	Max bias words	Min. Δ-debias val.	debias word
Female	69.3 (topic 41)	women men woman	-17.5 -17.5 -17.5	women women women
Christian	53.5 (topic 25)	christian christians christianity	-23.2 -23.2 -13.3	pakistanis pakistanis satan
Male	82.0 (topic 60)	man men male	-12.0 -56.0 -1324.8	someone males winnable
White	60.7 (topic 7)	white people racist	-16.8 -417.5 -43.2	blue someone fascism
Muslim	54.1 (topic 12)	muslims muslim islam	-21.7 -21.7 -21.7	kurds kurds kurds
Black	58.0 (topic 0)	black blacks people	-46.0 -23.7 -16.6	black blacks someone
Homosexual_ gay_or_ lesbian	55.7 (topic 57)	gay marriage sex	-61.6 -52.3.0 -21.8	queer polygamy bisexuality
Jewish	40.3 (topic 52)	jewish jews israel	-75.7 -19.1 -19.1	aryan serbs serbs

Table 11: **3rd debiasing strategy**. For each discrimination type, we observe the 3 most biased words and the three most debiased words under the constraint that the debiased words hold similar meaning to the biased words. We select the 2.5% most debiased words according to the Δ-measure defined above, compute their Word2Vec embedding, and find the word most similar to the initial word among the subset of debiased words. The replacement words suggested in this approach are closer in meaning to the original biased words than those from the 2nd debiasing strategy. The replacement of "man" by "someone" in the Male discrimination section is particularly encouraging. However, Word2Vec embeddings are not an appropriate measure for similarity between words in this context, and the replacement words still contain potentially discriminatory undertones. Future work could explore other methods of identifying synonyms. The definition for "debiased" using the Δ-measure is also highly arbitrary, and must be refined.

References

[1] E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams, "Multi-dimensional gender bias classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 314–331, Association for Computational Linguistics, Nov. 2020.

[2] Y. C. Tan and L. E. Celis, "Assessing social and intersectional biases in contextualized word representations," 2019.

[3] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov, "Rt-Gender: A corpus for studying differential responses to gender," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[4] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano, "Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application," 2020.

[5] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," 2016.

[6] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proceedings of the Third Workshop on Abusive Language Online*, (Florence, Italy), pp. 25–35, Association for Computational Linguistics, Aug. 2019.

[7] S. Kiritchenko and S. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, (New Orleans, Louisiana), pp. 43–53, Association for Computational Linguistics, June 2018.

[8] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, (Red Hook, NY, USA), p. 121–128, Curran Associates Inc., 2007.

[9] M. Roberts, B. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. Gadarian, and B. Albertson, "Structural topic models for open-ended survey responses," *American Journal of Political Science*, vol. 58, 03 2014.

[10] M. E. Roberts, B. M. Stewart, and D. Tingley, "stm: An r package for structural topic models," *Journal of Statistical Software, Articles*, vol. 91, no. 2, pp. 1–40, 2019.

[11] D. Gaucher, J. Friesen, and A. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality," *Journal of personality and social psychology*, vol. 101, pp. 109–28, 03 2011.

[12] M. E. Roberts, B. M. Stewart, D. Tingley, and E. Airoldi, "The structural topic model and applied social science," in *ICONIP 2013*, 2013.

[13] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, *Evaluation Methods for Topic Models*, p. 1105–1112. New York, NY, USA: Association for Computing Machinery, 2009.

[14] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, (USA), p. 262–272, Association for Computational Linguistics, 2011.

[15] D. Mimno and M. Lee, "Low-dimensional embeddings for interpretable anchor-based topic inference," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1319–1328, Association for Computational Linguistics, Oct. 2014.

[16] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1150–1159, 2003.

[17] J. Pearl, "Causal inference in statistics: An overview," *Statist. Surv.*, vol. 3, pp. 96–146, 2009.