A Bayesian Quantification of Aporophobia and the Aggravating Effect of Low–Wealth Contexts on Stigmatization

Ryan Brate,[†] **Marieke van Erp,**[†] **Antal van den Bosch**[⊕]
[†]DHLab, [⊕]Utrecht University
KNAW Humanities Cluster, DHLab, Amsterdam, the Netherlands
Utrecht University, Institute for Language Sciences, Utrecht, the Netherlands
{ryan.brate, marieke.van.erp}@dh.huc.knaw.nl
a.p.j.vandenbosch@uu.nl

Abstract

Aporophobia, a negative social bias against poverty and the poor, has been highlighted as an overlooked phenomenon in toxicity detection in texts. Aporophobia is potentially important both as a standalone form of toxicity, but also given its potential as an aggravating factor in the wider stigmatization of groups. As yet, there has been limited quantification of this phenomenon. In this paper, we first quantify the extent of aporophobia, as observable in Reddit data: contrasting estimates of stigmatising topic propensity between low–wealth contexts and high–wealth contexts via Bayesian estimation. Next, we consider aporophobia as a causal factor in the prejudicial association of groups with stigmatising topics, by introducing people group as a variable, specifically *Black people*. This group is selected given its history of being the subject of toxicity. We evaluate the aggravating effect on the observed n–grams indicative of stigmatised topics observed in comments which refer to Black people, due to the presence of low–wealth contexts. We perform this evaluation via a Structural Causal Modelling approach, performing interventions on simulations via Bayesian models, for three hypothesised causal mechanisms.

Disclaimer: This paper contains derogatory words and phrases. They are provided solely as illustrations of the research results and do not reflect the opinions of the authors or their organisations.

1 Introduction

Aporophobia, from the Greek *áporos* meaning *without resources* and *phobia* meaning *fear*, describes a negative social bias against poor people. In communicative contexts, one could imagine this taking the form of direct statements which express negative sentiment, such as, "I dislike beggars"; or take the form of negative bias elicited through an implied or asserted propensity to some negatively–perceived attribute, situation or behaviour: such as, "you can't be poor and be intelligent" or "poor people are more likely to be criminals"; or simply the act of associating poor people with some negative stereotyping in the same context.

The recent position paper, *Aporophobia: An Overlooked Type of Toxic Language Targeting the Poor* (Kiritchenko et al., 2023), makes the argument for the need for greater attention to aporophobic attitudes in discourse in the NLP sub–field of toxic speech analysis. The arguments put forward are three-fold: 1) aporophobia is an observable social phenomenon; 2) aporophobia may be an aggravating factor in the stigmatization of people groups; and 3) existing toxicity datasets offer too few aporophobic instances and/or targeted human annotations for adequate modelling. In the study, aporophobia was demonstrated according to associations with negatively biased topics: identifying such topics, via a BERTopic analysis on a subset of tweets containing n-grams proposed as highly indicative of *poor* or *low–wealth* instances.

There remains, however, open questions as to how disproportionate the associations between poverty contexts and negative topical associations are; and how strong an effect aporophobia is as an aggravating factor in the context of other forms of toxicity. Our contribution to this research area is twofold: firstly, we quantify the relative propensity of stigmatising topics with low–wealth contexts as opposed to high–wealth contexts. Secondly, we quantify the aggravating low–wealth status referenced in comments, on the observed rate of topical n–grams indicative of stigmatising topics associated with Black people. This group has been selected for their history of being subject to negative bias.The analysis is performed in the context of a corpus of publicly available Reddit content. We ask the following research questions: 1) *How statistically distinct is the co-occurrence of identified*

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 234–243

negatively biased topics in low–wealth contexts versus high–wealth contexts?; and, 2) *Can we estimate quantitatively, a non–negligible aggravating causal effect of low–wealth references on negatively biased topic rates, in respect of comments also referencing Black people?*

The first research question is one of statistical associations, e.g., the probability of occurrence of some negatively social biased, or stigmatising topics given some wealth context, and requires a subjective classification of associated topics as negatively socially biased or not: we ground this subjectivity in literature related to notions of stigmatising associations, detailed in Section 2.

The second research question is concerned with aggravation, which implies causation: i.e., some event increasing the incidence of some result. To answer this question we adopt the methodology of Structural Causal Modelling (SCM). This methodology allows us to evaluate the strength of causal interactions according to a presumed causal model. Thus, to answer the research question we must introduce a further subjectivity, *the causal mechanism under consideration*: how we represent this mechanism of aggravation of stigmatising topic association against some people group due to low–wealth status. The introduction of further subjectivity may give the reader pause; however, we argue that notions of prejudicial associations, and aporophobia are relatively straightforward concepts in regards their causal implications, thereby representing a clear starting point for causal analysis and a spring–board for further analysis and discussion.

2 Related Literature

In this research, we quantify prejudice against a group via stigmatising contextual associations. The suggestion of behaviours, attributes or situations as having implicit sentiment attachment is not controversial, nor is the idea of a behaviour, attribute or situation which is viewed negatively, being prejudicial when applied to a group as a stereotype. (Katz and Braly, 1933)

Various definitions are proffered in literature and in law to define stigmatising and stigma, however, most appear to conform in broad terms to the frequently cited Goffman, who defines stigmatization simply as, "as an attribute that is deeply discrediting" (Goffman, 1963). Albrecht et al. measured this discredited position on the notion of perceived social distance. Analysis of survey responses identified *social deviants*; i.e., ex-convicts, the mentally ill, and alcoholics as the both most social distanced and as physically threatening and offensive. The study highlighted a link between perceived disruption to social interaction and perceived social distance. Weiner et al. investigated sentiments towards stigmas perceived as onset-controllable (behavioural) or onset-uncontrollable (physical disability), where perceived onset-controllable stigmas are relatively strongly linked to anger, judgement and lack of pity. There are clear parallels between the outcomes of these aforementioned studies. Similar themes are revealed in Taylor and Dear, who based on analysis of surveys, linked mental health problems with perceptions of dangerousness, social isolation and lack of trustworthiness.

The second research question is concerned with measuring a causal effect, where we must address the need, limitations and successful use cases of Structural Causal Modelling. The gold standard for causal inference is the randomised controlled trial (RCT) (Eldridge et al., 2016). Observational data, however, precludes real–world intervention. Toxic speech analysis is one such field where practical and ethical considerations limit the scope for RCT studies. Such observational data is adequate for modelling statistical associations as the basis of predictive models, but falls short of being able to explore the interaction between explanatory features in a causal manner. However, the field of Structural Causal Modelling (SCM) (Pearl, 2009) offers a solution: a statistical framework for *simulating* the causal influence of interrelated features, given some assumed causal model. SCM has its roots in fields such as genetics (Wright) and econometrics (Haavelmo, 1943). Since, the explanatory value of its outcomes are predicated on the validity of the presumed causal model, the method is best suited to instances where the causal models have a high degree of apriori confidence. We argue for its applicability in quantifying aporophobia as an aggravating factor of prejudicial association , owing to the near self–evident causal nature of both aporophobia and prejudicial association of people group, in relation to stigmatising topics.

3 Data

In the absence of the Twitter data from (Kiritchenko et al., 2023), we use the subset of 266,268,920 separate public comments, from January 2015 to May 2015, from the Reddit social news ag-

gregation, content rating, and forum social network (Stuck_In_the_Matrix, 2015).[1]

We identify a likely low–wealth subset of Reddit comments via the presence of one of more of the n-grams: *poor people*, *poor folks*, *poor families*, *homeless*, *on welfare*, *welfare recipients*, *low-income*, *underprivileged*, *disadvantaged*, *lower class*. We identify a high–wealth subset of Reddit comments via the presence of one or more of the n-grams: *the rich*, *rich people*, *rich ppl*, *rich men*, *rich folks*, *rich guys*, *rich elites*, *rich families*, *wealth*, *well-off*, *upper-class*, *millionaires*, *billionaires*, *elite class*, *privileged*, *executives*. We differ from Kiritchenko et al. in respect of the low and high wealth n–grams only in the omission of the bigram, *the poor*, which a cursory examination hinted at a high frequency of associated non–wealth contexts in which it is used as an adjective, e.g., *the poor kittens*. There are 215,405 comments matching the low–wealth context seed n-grams and 258,124 comments matching the high–wealth seed n-grams. A sample of comments not flagged as low–wealth or high–wealth contexts were sampled with a Bernoulli probability of 0.4%, yielding a control sample of unspecified wealth contexts of 1,063,729 comments.

Additionally, we identify comments referencing Black people according to the presence of one or more of the seed n-grams: *blacks*, *Black people*, *black ppl*, *black kids*, *black guys*, *black men*, *black women*, *black families*; and separately, comments directly referencing Black people via the derogatory n-grams: *negro*, *negros*, *nigger*, *niggers*. There are a total of 248,108 comments the non–derogatory, Black people n–grams, and 73,586 comments referencing the derogatory Black people n–grams. The total size of this comment set is approximately 1.8M comments.

4 Methodology

Firstly, we perform topic analysis on the assembled sub-corpus. We then identify those low-ambiguity n–grams corresponding to topics, presumed indicative of suggested stigmatising topics with negative social biases. We make an estimate of the rate at which comments containing these n–grams demonstrate the stigmatising topic in question. In respect of the first research questions, we estimate the propensity of each of identified negative social bias, with respect to each of low–wealth and high-wealth comment subsets, and estimate their relative propensities. In respect of research question 2, we analyse the aggravating effect of references to a low-wealth context on co-occurrence frequencies of observed negatively biased topics with Black people: we analyse the aggravating effect according to three distinct possible causal models. All code use to generate the data and perform the analysis can be found on the GitHub repository accompanying this paper. [2]

4.1 Topic Analysis

Topic analysis is performed separately on: i) the low-wealth comments subset only; and ii) the whole set of approximately 1.8M comments, via BERTopic (Grootendorst, 2022) to identify emergent topics resulting from analysis on a small and large data set. As per the original study, we use the *all-MiniLM-L6-v2* embedding model; a vectorizer model , removing english stop–words and terms that appeared in less than 5% of sentences; and a minimum topic cluster size of 170 is specified (i.e., scaled down to approximately 1/3 of the original study's 500 owing to the available low–wealth comment set size being approximately 1/3 of Kiritchenko et al.).

4.2 Topical n–grams corresponding to presumed stigmatising topics

We rank the top-50 topics identified by BERTopic, ranked descending according to their frequency in the low–wealth subset. Within this ranked list, we select topics which we hypothesise as being strongly indicative of some underlying stigma. For each of these topics, and their corresponding BERTopic–provided most strongly predicting n–grams, we identify the least semantically ambiguous. For each n–gram set, we then estimate the rate at which the stigmatising topic is observed, with respect to 50 randomly sampled comments. The n–grams are listed in Table 1, where bold face denotes the low ambiguity n–grams sampled against, together with the count of observed stigmatising topics (as indicated in the table), from inspection of the random samples. We generally observe the bold face n–grams to result in high estimates of likely observance of the stigmatising topic. In the case of *addition, addict, addicts*, the *unspecified* meaning instances were overwhelmingly indicative of some addition, possibly substance abuse, but not

[1] https://en.wikipedia.org/wiki/Reddit

[2] https://github.com/ryanbrate/WOAH_2024_aporophobia

clearly specified. Thus, when considered in terms of the general topic of *some addiction*, the observed rate is 46/50.

Top 10 n–grams By BERTopic Topic	**Presumed Stigmatising Topic**	**Rate Observed**
police, **cops**, officer, **cop**, officers, gun, **police officers**, homeless man, force, shooting	interaction with law enforcement	39/50
prison, **jail**, court, lawyer, justice, lawyers, trial, guilty, **prisons**, legal	as related to incarceration	49/50
food, healthy, **fast food**, eat foods, cook meal, **mcdonalds** (**mcdonald's**, **McDonalds**, **McDonald's**), fast, healthy food	ultra–processed food consumption	22/50
drug, **drug testing**, testing, recipients, welfare recipients, welfare, drugs, **drug test**, test, tested	testing for drug use	50/50
fat people, weight, **obese obesity**, **overweight**, skinny, people fat, **fatties**, healthy	obesity	50/50
relationship, **attractive**, sex, women, **dating**, date, girl, girls, married, divorce	perceived eligibility	41/50
marijuana, drugs, drug, prohibition, **cannabis**, legalization, weed, illegal, alcohol, pot	association with marijuana	50/50
mental, mentally, **mentally ill**, ill, **mental illness**, illness, mental health, health, homeless, homeless people	mental illness	50/50
heroin	association with heroin	50/50
addiction, drugs, drug, sober, **addict**, life, drinking, **addicts**	substance add. unspecified gambling	27/50 15/50 4/50

Table 1: Presumed stigmatising topics, and the counts they are observed in a random sample of the assembled corpus, corresponding to the bold face n–grams of the most relevant n–grams to each identified topic.

4.3 Estimation of the relative propensity of stigmatising topics with wealth context

For each comment, the presence of topical n–grams which are interpretable in context as a stigmatising topic, is a binary event. Accordingly this can be represented as the outcome of a Bernoulli trial, according to some latent propensity, or probability of occurrence. Using the data of Tables 1 and 2, we can estimate this propensity, $P(\text{stig., } n\text{–grams} \mid \text{wealth cont.})$.

Table 1 lists counts of presumed stigmatising topics, and the rate they are observed in random samples which contain the bold–face, low–ambiguity, topical n–grams listed. We denote this count $C_{\text{stig.} \mid \text{sample}}$ with respect to a total count, C_{sample}, for each sample set. Using these counts, we compute a posterior estimation of the probability of observing the stigmatising topic given the presence of the n–grams, $P(\text{stig.} \mid n\text{–grams})$. We do this via via Bayesian Estimation (Kruschke, 2012) using PyMC (Oriol et al., 2023), assuming an effectively uniform prior probability, according to equation set 1.

Topical n–grams	**Count in Low Wealth Context**	**Count in High Wealth Context**
police, cops, cop, police officers	7737	5228
prison, jail, prisons	5082	4100
fast food, mcdonalds, mcdonald's, McDonald's McDonalds	2067	1036
drug testing, drug test	694	78
fat people, obese obesity, overweight, fatties	1450	750
relationship, attractive, dating	3685	6022
marijuana, cannabis	536	573
mentally ill, mental illness	3331	359
heroin	979	316
addiction, addict, addicts	3708	625

Table 2: Co-occurrence counts of the selected n–grams, presumed indicative of the stigmatising topics in Table 1, with low and high–wealth contexts.

$$\begin{aligned} P(\text{stig.} \mid n\text{–grams}) &\sim Logistic(Normal(0, 1.5)) \\ C_{\text{stig.} \mid \text{sample}} &\sim Binomial(P(\text{stig.} \mid n\text{–grams}), C_{\text{sample}}) \end{aligned} \tag{1}$$

Table 2 lists the frequencies of the these same topical n–grams with both the low–wealth contexts, $C_{n\text{–gram} \mid \text{low–wealth}}$ and high–wealth contexts, $C_{n\text{–grams} \mid \text{high–wealth}}$. We use these counts, with respect the the total available comments for each wealth context, to estimate the probability of an n–gram set given each wealth context, $P(n\text{–grams} \mid \text{wealth cont.})$. We do this via Bayesian Estimation according to Equation set 2.

$$\begin{aligned} P(n\text{–grams} \mid \text{wealth cont.}) &\sim Logistic(Normal(0, 1.5)) \\ C_{n\text{–grams} \mid \text{wealth cont.}} &\sim Binomial(P(n\text{–grams} \mid \text{wealth cont.}), C_{\text{wealth cont.}}) \end{aligned} \tag{2}$$

The Bayesian posterior estimate of $P(\text{stig., } n\text{–grams} \mid \text{wealth cont.})$ is then estimated via the chain rule of Equation 3. This is predicated on the simplifying assumption that $P(\text{stig.} \mid n\text{–grams, wealth cont.})$ is approximately equal to $P(\text{stig.} \mid n\text{–grams})$.

$$\begin{aligned} p(\text{stig., } n\text{–grams} \mid \text{wealth cont.}) = \\ P(\text{stig.} \mid n\text{–grams, wealth cont.}) \times \\ P(n\text{–grams} \mid \text{wealth cont.}) \end{aligned} \tag{3}$$

We compare these estimates of stigmatising topic propensity, for each of the low–wealth and high–wealth contexts according to the Relative Risk ratio, given by Equation 4. We apply the Risk Ratio to paired samples of the posterior estimates of $P(\text{stig., } n\text{–grams} \mid \text{wealth cont.})$, for the low–wealth and high–wealth contexts, yielding a Bayesian posterior estimate of the Risk Ratio. The

outcomes of the analysis are given in Section 5.1, Table 3.

$$\text{Risk Ratio} = \frac{\text{P(stig., } n\text{–grams | low–wealth context)}}{\text{P(stig., } n\text{–grams| high–wealth context)}} \quad (4)$$

4.4 Poverty as an aggravating factor of people group stigmatisation

The presence of a reference to a *low–wealth* context, some *people group* and some *stigmatising topic* are binary events. However, in regards to the notion of aggravation of stigmatising topic association, the causal process by which one binary event influences another is not found in the data: it must be proposed. With this in mind, we note the following foundational assumptions which follow naturally from the concepts of prejudice and aporophobia: individuals or groups may be stigmatized via low–wealth associations: individuals or groups may be stigmatized outside of low–wealth associations; and, association with certain topics may act as proxies for stigmatization.

Supplementary to this, we propose three separate suppositions regarding how *people group* and *low–wealth* context occurrences are causally related with each other. Figures 1, 2 and 3 are plate models of the generative regression models representing these suppositions. Equation sets 5,6, and 7 are the corresponding equations defining each regression model. In each, the observable binary variables as to the occurrence of people group (G_i), low–wealth context reference (W_i) and stigmatising topic (T_i), corresponding to each separate comment (of index i) are shaded grey: considered on their own, the observable variables and the edges between them can be considered as Directed Acyclic Graphs (DAGs), indicating the direction of influence between them.

Supposition 1: any joint references to Black people and references to low–wealth status are incidental, however, both influence the chance of observing a stigmatising topic. Supposition 2: the chance of observing low–wealth status references is influenced by the presence of Black people references. Both influence the probability of observing a stigmatising topic. Supposition 3: the chance of observing references to Black people is influenced by the presence of low–wealth status references. Both influence the probability of observing a stigmatising topic.

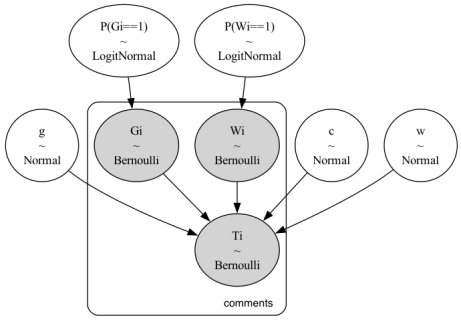


Figure 1: Bayesian regression model for causal supposition 1: that for each comment, i, the probability of occurrence of either a reference to the people group of interest, G_i or low–wealth context, W_i, are not directly influenced by one another. However, both people group and low–wealth references influence the probability of occurrence of a stigmatising topic.

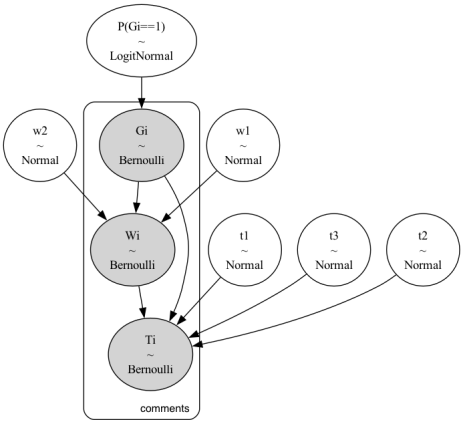


Figure 2: Bayesian regression model for causal supposition 2: that for each comment, i, the probability of occurrence of reference to a low–wealth context, W_i, is influenced by the presence of the people group in question, G_i. Both in–turn influence the probability of occurrence of a stigmatising topic, T_i.

$$\begin{aligned} T_i &\sim Bernoulli(P(T_i = 1)) \\ P(T_i = 1) &= Logistic(t1 + G_i.t2 + W_i.t3) \\ t1, t2, t3 &\sim Normal(0, 5) \\ G_i &\sim Bernoulli(P(G_i = 1)) \\ W_i &\sim Bernoulli(P(W_i = 1)) \\ P(G_i = 1) &\sim Logistic(Normal(0, 1.5)) \\ P(W_i = 1) &\sim Logistic(Normal(0, 1.5)) \end{aligned} \quad (5)$$

$$\begin{aligned} T_i &\sim Bernoulli(P(T_i = 1)) \\ P(T_i = 1) &= Logistic(t1 + G_i.t2 + W_i.t3) \\ W_i &\sim Bernoulli(P(W_i = 1)) \\ P(W_i = 1) &= Logistic(w1 + G_i.w2) \\ G_i &\sim Bernoulli(P(G = 1)) \\ P(G = 1) &\sim Logistic(Normal(0, 1.5)) \\ w1, w2, t1, t2, t3 &\sim Normal(0, 5) \end{aligned} \quad (6)$$

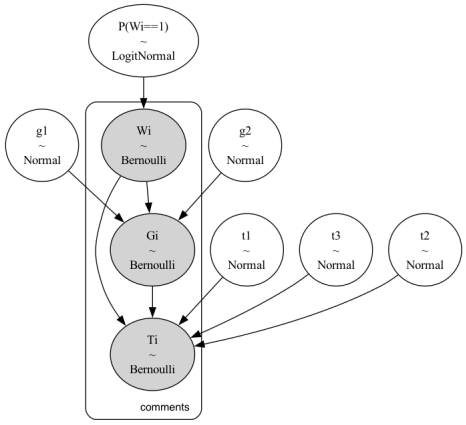


Figure 3: Bayesian regression model for causal supposition 3: that for each comment, i, the probability of occurrence the people group in question, G_i, is influenced by the presence of low–wealth context references, W_i. Both in–turn influence the probability of occurrence of a stigmatising topic, T_i.

$$\begin{aligned} T_i &\sim Bernoulli(P(T_i = 1)) \\ P(T_i = 1) &= Logistic(t1 + G_i.t2 + W_i.t3) \\ G_i &\sim Bernoulli(P(G_i = 1)) \\ P(G_i = 1) &= Logistic(g1 + W_i.g2) \\ W_i &\sim Bernoulli(P(W_i = 1)) \\ P(W_i = 1) &\sim Logistic(Normal(0, 1.5)) \\ g1, g2, t1, t2, t3 &\sim Normal(0, 5) \end{aligned} \tag{7}$$

Each generative (regression) model corresponding to a supposition, is fitted to the data via PyMC (Oriol et al., 2023). We measure low–wealth context and stigmatising topic presence via the indicative n–grams previously outlined. Black people are considered as the people group, whose occurrence is measures via the indicative n–grams previously outlined. The result of the model fitting are posterior estimates of the probability distributions of each latent model parameter. We use these parameters as the basis for simulating the causal effect of changes to the observed rates of low–wealth context instances, on stigmatising topic co–occurrence. The implementation of the generative (regression) models, has been checked against simulated data for each of the causal models

In evaluating, *can we estimate quantitatively, a non–negligible aggravating causal effect of low–wealth references on negatively biased topic rates, in respect of comments also referencing Black people?*, we consider the the outcomes of the Bayesian simulations for each of the causal models in terms of the statistics given by Equation 8 and Equation 9. Both of these statistics measure the effect of simulated interventions, on the observed rate of stigmatising topic co–occurrence. The intervention in question, being a factoring of the expectation of low–wealth context occurrence, $P(W_i = 1)$. Equation 8 contrasts the effect of intervening vs not intervening, in the presence of people group of interest references. Equation 9, contrasts the effect of an intervention of the same magnitude, in the presence of people group of interest references versus in their absence. The combination of both statistics enables us to measure how *disproportionate* the aggravating effect of low–wealth status is on the vilification of some people, according to topical associations. For each causal model and topic separately we simulate both, the intervention cases and the non-intervention case over 4000 times, for a comment set size of 1000, as per the PyMC defaults. We record the maximum likelihood point estimates of $P(T_i = 1|G_i = 1, \text{intervention})$ and $P(T_i = 1|G_i = 0, \text{intervention})$, for each simulation. Thus, giving us a posterior distribution of these statistics, from which to calculate credible intervals with respect to the statistics given by Equations 8 and 9. Several variations on the intervention, a factoring of the models' latent $P(W_i = 1)$, are considered, to help identify the general trend. The outcomes of the analysis can be found in Section 5.2.

$$\frac{P(T_i = 1|G_i = 1, \text{intervention})}{P(T_i = 1|G_i = 1, \text{no intervention})} \tag{8}$$

$$\frac{P(T_i = 1|G_i = 1, \text{intervention})}{P(T_i = 1|G_i = 0, \text{intervention})} \tag{9}$$

5 Results and Evaluation

Section 5.1 corresponds to the first research questions according to the methodology detailed in Section 4.3. Section 5.2 corresponds to the second research question according to the methodology detailed in Section 4.4

5.1 Estimation of the relative propensity of stigmatising topics with wealth context

Table 3 lists the posterior estimates of the Risk Ratios, according to Equation 4, a measure of the relative propensity of each stigmatising topic between low–wealth and high–wealth subsets. The Risk Ratio is reported according to the 99% most credible interval. It is evident that *mental illness*, *testing for drug use*, *addiction* and *association with heroin* demonstrate the most extreme estimated propensities for low–wealth contexts as opposed for high–wealth contexts, with respect to their lower–bound Risk Ratio estimates.

The outcomes of Table 3 estimate the skew by wealth context, in regards to the *contextual asso-*

Stigmatising Topics	Est. Risk Ratio
interaction with law enforcement	1.2 to 2.4
as related to incarceration	1.2 to 1.7
ultra–processed food consumption	1.1 to 4.3
testing for drug use	**6.8** to 14.3
obesity	1.9 to 2.7
perceived eligability	0.69 to 0.78
association with marijuana, cannabis	0.93 to 1.4
mental illness	**9.0** to 13.2
association with heroin	**3.0** to 4.6
addiction	**5.4** to 8.8

Table 3: 99% Credible Interval Risk Ratios comparing credible estimates of the relative propensity of each stigmatising topic for low–wealth (as opposed to high). Bold denotes the lower–bound estimates of the most severe skews in association.

ciation between the listed stigmatising concepts and the wealth contexts. We extend this by estimating the wealth context skew of stigmatisation, not just on contextual co–occurrence, but according number of instances that the stigmatising topic *directly marks* person or group representing the wealth context. I.e., a person of the corresponding wealth contest described as: being subject to a drug test; having a mental illness, using heroin use, or having an addiction. Thus, we estimate a Risk Ratio based not on an estimate of $P(\text{stig., } n\text{–grams} \mid \text{wealth cont})$, but on $P(\text{dir., stig., } n\text{–grams} \mid \text{wealth cont.})$. As per the chain rule expansion of Equation 10, we require an estimate of $P(\text{dir., } \mid \text{stig., } n\text{–grams, wealth cont.})$.

$$\begin{aligned} P(\text{dir., stig., } n\text{–grams} \mid \text{wealth cont.}) = \\ P(\text{dir., } \mid \text{stig., } n\text{–grams, wealth cont.}) \times \\ P(\text{stig, } \mid n\text{–grams, wealth cont.}) \times \\ P(n\text{–grams} \mid \text{wealth cont.}) \end{aligned} \tag{10}$$

For each of *mental illness*, *testing for drug use*, *addiction and association with heroin*, we further sample 50 comments containing the corresponding topical n–grams of Table 2 for each of low–wealth and high–wealth contexts. From these samples and for each wealth context, we obtain counts of: i) the number of sample comments for which the topical n–grams are demonstrative of the stigmatising topic in question, $C_{\text{stig., } \mid \text{sample}}$; and ii) of those comments for which the topical n–grams are demonstrative of the stigmatising topic, a count of the subset for which the stigmatising topic is *directed marks* people representative of the wealth context in question, $C_{\text{dir. } \mid \text{ stig., sample}}$. These counts are reported in Table 4. We then make a posterior Bayesian estimate, for each of low–wealth and high wealth contexts of, $P(\text{dir. } \mid \text{ stig., } n\text{–grams, wealth cont.})$, as per Equation set 11. We the subsequently obtain a posterior estimate of the propensity of directed stigmatisation, $P(\text{dir., stig., } n\text{–grams} \mid \text{wealth cont.})$ as per Equation 10.

directed stigmatisation	low–wealth context	high–wealth context
having mental illness	39/50	16/50
tested for drug use	48/50	16/50
having addiction	43/50	23/50
using heroin use	42/47	20/45

Table 4: $C_{\text{dir. } \mid \text{ stig. sample}}$ / $C_{\text{stig. } \mid \text{ sample}}$ counts. Where $C_{\text{stig. } \mid \text{ sample}}$ is a count of comments where the stigmatising topic is observed, and $C_{\text{dir. } \mid \text{ stig., sample}}$ is a count of where this observed stigmatising topic is directed at people representative of the wealth context.

$$\begin{aligned} P(\text{dir. } \mid \text{ stig. , } n\text{–grams, wealth cont.}) \sim Logistic(Normal(0, 1.5)) \\ C_{\text{dir., } \mid \text{ stig., sample}} = \\ Binom(P(\text{dir. } \mid \text{ stig. , } n\text{–grams, wealth cont.}), C_{\text{stig. } \mid \text{ sample}}) \end{aligned} \tag{11}$$

We present these updated Risk Ratios, reflecting the relative propensity of directed stigmatiation according to wealth context in Table 5. We observe an even greater skew towards low–wealth contexts of directed stigmatisation with respect to the analysed topics than of the contextual association with stigmatising topics of Table 3.

directed stigmatisation	Est. Risk Ratio
having mental illness	**14.3** to 51.0
tested for drug use	**14.9** to 58.0
having addiction	**7.9** to 22.5
heroin use	**4.5** to 12.5

Table 5: 99% Credible Interval Risk Ratios comparing estimate of the relative propensity of directed stigmatisation for low–wealth (as opposed to high). Bold denotes the lower–bound estimates of the Risk Ratios.

Closer inspection of the comment random samples, demonstrates the low–wealth contexts with respect to these high association topics, to be highly specific : homelessness is overwhelmingly the low–wealth n–gram related to *mental illness* and *addiction* and *association with heroin*; and welfare (as in receipt of government aid) in respect of *drug testing, drug test* topical associations.

5.2 Poverty as an aggravating factor of people group stigmatisation

As per the analysis of Section 4.4, for each proposed causal model, the propensity of low–wealth contexts was directly factored as an explanatory

intervention as to the effect of low–wealth context on stigmatising topic association with reference to Black people. The statistic given by Equation 8 estimates the relative increase in expected stigmatising topic occurrence, given the presence of the people group of interest, due to the intervention. The statistic is calculated as a 99% Credible Interval. Thus, where the lower–bound estimate of this statistic exceeds 1.0, for some level of intervention on the expected rate of low–wealth contexts, the implication is that there is a non-zero effect on observed stigmatising topic rates due to the intervention, with a 99% probability. Figures 4a and 4b give the lower bound estimate with respect to causal suppositions 1 and 2. Figure 4a gives estimates of this lower bound statistic for the topical n–grams *police, cops, cop, police officers*, given causal supposition 1. Figure 4b gives estimates of this lower bound statistic for the topical n–grams *police, cops, cop, police officers* and *prison, jail, prisons*, given causal supposition 2. In both cases, the causal link, between low–wealth references and observed frequency of those specific stigmatising topics is weak: a very large intervention is needed before the lower–bound estimated measure of the effect is non–negligible.

The statistic given by Equation 9, for some causal supposition, estimates the relative increase in expected stigmatising topic occurrence at some level of intervention: contrasting comments containing and omitting the people group. Where this statistic exceeds 1.0, for some level of intervention on the expected rate of low–wealth contexts, the implication is that there is a non-negligible relative increase. Figures 4c and 4d gives the lower bound estimates of the 99% Credible Interval estimates of this statistics. We see lower bound estimates of this statistic exceed 1.0 for both *prison, jail, prisons* and *police, cops, cop, police officers*, given either causal supposition 1 or 2.

To further contextualise the results, we again randomly sampled comments. We sample 50 samples from the pool of 629 comments where Black people, low–wealth references and *police,cop, cops, police officer* topical n–grams are present; and 50 samples from the pool of 348 comments where Black people, low–wealth reference and *prison, jail, prisons* topical n–grams are present.

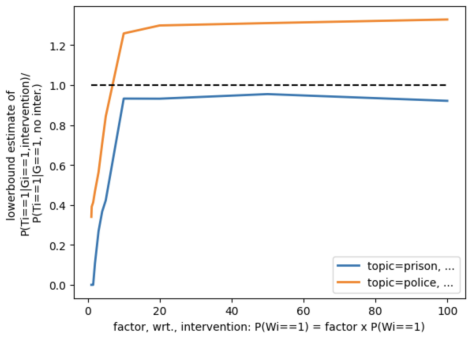
The Black people, low–wealth, *police, cop, cops, police officer* samples have the following observed implications: 38/50 discuss the targeting of Black people by the police, and a further 2/50 are related in that they imply a disproportionate response by the judicial system. The following quote typifies the common referencing of low–wealth and Black people in stigmatised contexts, "Do poor people commit more crimes? Yes. Are there more poor Black people? Also yes. Does that mean police target blacks more harshly? No."

In the Black people, low-wealth, *prison, ...* topical n–grams, 45/50 explicitly refer to the incarceration of Black people.

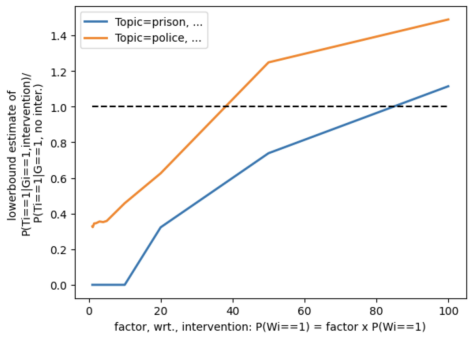
6 Limitations and Conclusion

With regards the research question, *how statistically distinct is the co-occurrence of identified negatively biased topics and low–wealth contexts versus high–wealth contexts?*, we see evidence of support of aporophobia for several proposed stigmatising topics: *mental illness*; *testing for drug use*; *addiction*; and *association with heroin*. Based on the incorporation of estimates, of the probability of topical n–grams indicative of a stigmatising topic actually being that topic, each was estimated as highly disproportionately associated with low–wealth contexts. Additionally, in further incorporating estimates of the probability of a stigmatising topic being directed at people of groups representative of the wealth context, an even greater skew towards low–wealth contexts was shown. E.g., heroin is more likely to contextually occur with low–wealth contexts than high–wealth; but low–wealth people or groups are even more likely to be characterised as *using heroin* than high–wealth users. These results are predicated on the wealth context n–grams being suitable proxies for the respective wealth contexts. However, it should be noted that what was observed in these strongest of outcomes, corresponded to highly specialised manifestations of aporophobia, in respect of highly specific social discussions: E.g., drug testing in the context of welfare receipt. This is somewhat expected: the selected n–grams were chosen for high precision in respect of the context they predict: to promote strong signals to facilitate detection. There remains an open question as to how to address aporophobia as a phenomenon related to less polarising depictions of low–wealth status, in terms of relatively more ambiguous language.

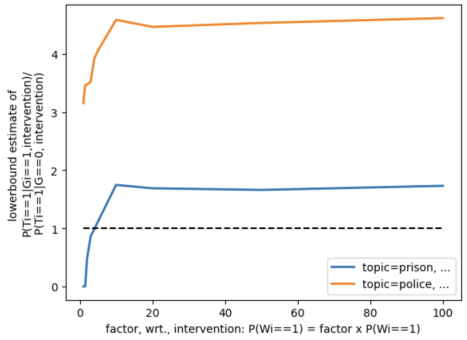
With respect to the second research question, *can we estimate quantitatively, a non–negligible aggravating causal effect of low–wealth references on negatively biased topic rates, in respect of com-*



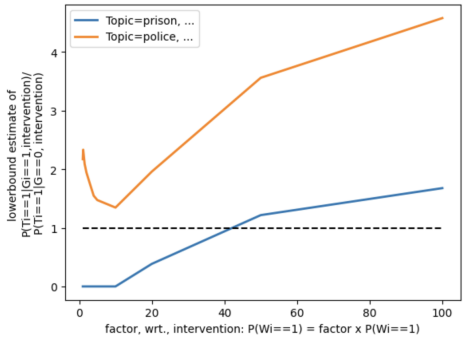
(a) lower bound estimate of the statistic given by Equation 8, with respect to causal supposition 1, according to a 99% Credible Interval.



(b) lower bound estimate of the statistic given by Equation 8, with respect to causal supposition 2, according to a 99% Credible Interval.



(c) lower bound estimate of the statistic given by Equation 9, with respect to causal supposition 1, according to a 99% Credible Interval.



(d) lower bound estimate of the statistic given by Equation 9, with respect to causal supposition 2, according to a 99% Credible Interval.

ments also referencing Black people?: we detected an aggravating causal relationship between low–wealth status and i) *police, cops, cop, police officer* assuming causal supposition 1; and ii) both *police, cops, cop, police officer* and *prison, jail, prisons* according to supposition 2. Inspection of random samples of these coincident contexts demonstrated a high estimate of clearly directed negative implications. However, the analysis suggested a *weak causal relationship*. No causal relationship was found between low–wealth status and any of the analysed stigmatising topics, for the model related to causal supposition 3.

The positive results from the Bayesian models corresponding to supposition 1 and 2, imply the detection of aporophobia, albeit weakly, in regards to the assumed causal models and predicated on the analysis assumptions. In contrast, as was the case for the causal model corresponding to supposition 3 and the other stigmatising topics; a failure to detect aporophobia via SCM, implies the proposed causal model and the data are incompatible: i.e., an incorrectly framed causal model; or a dataset or data features not reflective of the phenomena.

The proposed causal models, were proposed based on the almost self-evident expressions of prejudice against a group and aporophobia. In contrast, the analysis highlights a problem of data sparsity, in balancing feature precision and recall: i.e., from the Reddit subset of approximately 266M comments, only 0.1% referenced the selected low–wealth n–grams; of which only 2%reference the Black people n–grams. The pool is further shrunk according to the considered topics, which could explain the relatively few stigmatising topics for which aggravation was detected: *prison, jail, prisons* and *police, cops, cop, police officers*. These topics have the highest representation in the low–wealth and Black people common context wealth pools. We interpret these results as further support for the need for annotation schemes and corresponding datasets specifically tailored towards aporophobia for sensitive detection of the phenomena in regards toxic speech.

7 Future Work

It would be interesting to extend this study to other dataset domains, but moreover, to incorporate a modified feature set benefiting from any future human–annotated datasets dedicated to aporophobia. This would facilitate both a wider and more sensitive analysis of the topic.

References

Gary L. Albrecht, Vivian G. Walker, and Judith A. Levy. 1982. Social distance from the stigmatized : A test of two theories. Social Science & Medicine, 16(14):1319–1327.

Sandra M Eldridge, Claire L Chan, Michael J Campbell, Christine M Bond, Sally Hopewell, Lehana Thabane, and Gillian A Lancaster. 2016. Consort 2010 statement: extension to randomised pilot and feasibility trials. BMJ, 355.

Erving Goffman. 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice-Hall.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

T. Haavelmo. 1943. The statistical implications of a system of simultaneous equations. Econometrica, 11:1.

Daniel Katz and K. W. Braly. 1933. Racial stereotypes of one hundred college students. The Journal of Abnormal and Social Psychology, 28:280–290.

Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C. Fraser. 2023. Aporophobia: An overlooked type of toxic language targeting the poor. In The 7th Workshop on Online Abuse and Harms (WOAH), pages 113–125, Toronto, Canada. Association for Computational Linguistics.

John Kruschke. 2012. Bayesian estimation supersedes the t test. Journal of experimental psychology. General, 142.

Abril-Pla Oriol, Andreani Virgile, Carroll Colin, Dong Larry, Fonnesbeck Christopher J., Kochurov Maxim, Kumar Ravin, Lao Jupeng, Luhmann Christian C., Martin Osvaldo A., Osthege Michael, Vieira Ricardo, Wiecki Thomas, and Zinkov Robert. 2023. Pymc: A modern and comprehensive probabilistic programming framework in python. PeerJ Computer Science, 9:e1516.

Judea Pearl. 2009. Causality: Models, Reasoning and Inference, 2nd edition. Cambridge University Press, USA.

Stuck_In_the_Matrix. 2015. Reddit public comments (2007-10 through 2015-05).

S. Martin Taylor and Michael J. Dear. 1981. Scaling Community Attitudes Toward the Mentally Ill. Schizophrenia Bulletin, 7(2):225–240.

Bernard Weiner, Raymond P Perry, and Jamie Magnusson. 1988. An attributional analysis of reactions to stigmas. Journal of personality and social psychology, 55 5:738–48.

sewall Wright. Correlation and causation. Journal of agricultural research, 20(3).