EQUITBL: Usage Guide & Worksheet

H Devinney

Abstract

EQUITBL is a mixed-methods approach to analyzing bias in corpora using semi-supervised topic modeling. By combining a quantitative method for filtering relevant information out of a corpus with a qualitative analysis of the results, we can better understand *how* biases manifest in text corpora. This enables better risk analysis about how these biases may propagate or be amplified in NLP tools that use these corpora as training data.

Document Structure. This document contains three core elements: an introduction to the problem of implicit textual bias in language data; instructions for usage (of the support code *and* of the qualitative analysis method); and some information about how we have developed the method.

1 Introduction: Bias in Language Corpora

Bias in text data is a problem for Natural Language Processing (NLP) applications, because it leads directly or indirectly to harm against people. However, because most modern corpora are very large, it is difficult to detect the overall patterns of associations between concepts and groups or identities within these datasets.

The EQUITBL (Explore, Query, & Understand Implicit Textual Bias in Language data) package[1] provides a mixed-methods approach to analyzing large text corpora. By combining quantitative methods to discover themes and guided methods for qualitative analysis of these themes, we allow users to make a serious study of the contents of datasets which would otherwise be too large to critically assess. The results can inform stakeholders of the potential risks (e.g. lack of representation or stereotypical, offensive, or otherwise undesirable associations) within the data, allowing for curation of datasets to be more inclusive and fair.

1.1 Background

"Bias" as it relates to NLP has variable definitions (Blodgett et al., 2020), which may be categorized in different ways, for example by the types of harm (allocational vs representational) they cause or by how the performance of the system differs by group (e.g. counterfactual fairness (Kusner et al., 2017) or performance parity). Different definitions may also have varying formalizations and metrics, owing to the fact that "social bias" is a fundamentally complicated problem which resists any single, and therefore inherently oversimplifying, "tidy" calculation.

Bias in NLP models stems from multiple sources: imbalances or implicit biases in the source data; artefacts of the training method; and the combination thereof may all contribute to the end result of a model exhibiting biased behavior. We hope to contribute to lessening bias by providing methods to analyze source data for implicit biases before training models.

1.2 Bias Statement

We consider the problem of bias as one of systemic (and often systematic) injustice, and its presence in language data and technologies as a manifestation of the underlying power structures that construct society(ies). "Bias" therefore will appear differently across different contexts. Understood via Collins' matrix of domination, power structures operate on various levels of society and are fundamentally "characterized by intersecting oppressions" (Collins, 2000, p. 26), meaning that different 'types' of bias (e.g. gender and

[1] `https://github.com/hdevinney/EQUITBL`

racial bias) have a more than additive effect and cannot be understood as fully independent of each other or outside of the context of a particular place and time.

Following the first principle of Data Feminism ("examine power" (D'Ignazio and Klein, 2020)), we translate the issue of "identifying" bias in language data to one of "understanding" how power operates within the data. Because there is no one specific system to trained on the data to exhibit any particular behaviors, we instead search for hegemonic and or stereotypical themes within the underlying relationships of the words within the data, encouraging users to ask questions specifically tuned to different kinds of harm being done to different groups – both representationally within the data itself, and as various downstream harms from potential applications of the data.

For example, in the case of looking at gender biases we may ask questions about demographic parity within the data (*Are similar numbers of men, women, and nonbinary*[2] *people included? If not, does the representation correspond to general population statistics?*) as well as about the treatment of specific groups (*Are women more frequently associated with the concepts of home and family? Are men more frequently associated with politics and the economy? Are nonbinary people visible at all in the results?*). These questions should be informed by both the context of the dataset and the particular type of bias in question. For these reasons, we intentionally do not include any statistical measure or 'threshold' to define bias and instead provide a framework for qualitative analysis.

1.3 Related Work

The problem of identifying biases in training data for NLP is less well-studied than identifying biases in the resulting models. One common strategy is to try and ensure representational parity within datasets, i.e. make sure groups are represented with the same frequency in a corpus. In a gender-bias context this might entail counting the occurrences of different pronouns and then using a data augmentation strategy to add new material for underrepresented groups, see e.g. (Cao and Daumé III, 2020; Maudslay et al., 2019; Zmigrod et al., 2019).

Another related area dealing with the problem of "biased", stereotypical, and/or offensive content in datasets in NLP is labeling for the presence of hate speech or toxicity. Sap et al. (2020) provide a formalism for modeling pragmatic frames projecting or implying social biases or stereotypes, and establish a baseline for inferring these frames based on crowd-sourced annotations. These "Social Bias Frames" account for factors such as who is targeted (group vs individual); who the author is (in- or out-group?) and their perceived motivation; as well as the offensiveness of the statement or its implication(s). This requires close reading and annotation of texts within a corpus.

2 User Guide

In a nutshell, the workflow for EQUITBL is like this:

1. Convert your corpus into a nice, friendly format so that the system can read your texts – *section 2.2*
2. Preprocess your corpus – *section 2.3*
3. Train your topic model(s) – *section 2.4*
4. Get visualizations of your model(s) – *section 2.5*
5. Perform qualitative analysis – *section 2.6*

The process intentionally does not involve any "bias score" or similar metrics. This is done to encourage users to critically engage with the topics and patterns they find rather than merely being incentivized to chase "better" scores.

[2]We use 'nonbinary' as an umbrella term referring to all gender identities between or outside the 'binary' categories of men and women, following Devinney et al. (2020).

2.1 Before You Begin

For convenience in running scripts, EQUITBL uses config files (an example is provided). This is where you will set your preferred parameters, what you would prefer the output to be called, etc. You will also need to specify a few absolute paths (to the root of your project, and to a logfile).

Because training time and memory scale linearly with the amount of data in a corpus (see *section 3.2*), your hardware may limit how much data you can include in a topic model. Some 'tricks' to reduce this include pruning your vocabulary more strictly (reduce number of terms) and not using sliding context windows when chunking documents (reduce number of documents).

2.2 Corpus Formatting

If your corpus is not already preprocessed and saved in bag of words format (EQUITBL expects a `gensim` corpus and dictionary), you will have to take some steps to convert it into the schema EQUITBL expects when performing preprocessing steps.

EQUITBL expects corpora to be stored in `.json` files, with each document stored as a single string and given an associated ID. Formatting your corpus is therefore also an opportunity to define the size of your documents.

Example code can be found in `example_to_schema.py`

2.3 Preprocessing

At this step, you'll need to make some choices about how you would like your corpus to be preprocessed, and then save it to a `gensim` corpus and dictionary. We provide some suggested values in the example config file.

Document Size. In topic modeling, corpora are divided into documents which may in principle be any size. These documents essentially provide the context for which words are counted as co-ocurrences. We find that for gender bias detection in English, narrower contexts (in the form of *chunks*, a sliding window of a fixed number of tokens) are more effective. EQUITBL provides automated options for splitting texts into sentence-level or fixed-sized chunk documents.
default: 24-term chunks

Pruning (minimum frequency). To help keep the vocabulary size manageable and prevent very rare words from potentially skewing results, EQUITBL prunes infrequent terms. You can decide for yourself what "infrequent" means, although we set the default value to 3. Note that seed words will automatically be exempted from this pruning.
default value: 3

Stopwords. To help keep vocabulary size manageable and prevent very common words from cluttering up the results, you can define a list of stopwords, which will be ignored. For English, we suggest a modified version of the stopword list included in `nltk` (keeping third person pronouns, which are useful for investigating gender bias). Note that seed words may *NOT* be automatically exempted from this pruning.

Stop POS. To help keep vocabulary size manageable and only look at particularly "content-ful" terms when visualizing topics, EQUITBL takes a list of parts of speech which should be ignored. For English, we suggest keeping nouns (including proper nouns), verbs (including modals), adjectives, adverbs, pronouns, foreign words, and symbols. Thus, the default list contains all the other POS tags. (Note that POS tagging for English provided in EQUITBL uses the Penn Treebank tagset[3].)

[3] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

2.4 Training

To train the topic models, several hyperparameters must be provided (we suggest some defaults based on our experiments in section 3.2). The number of topics, t, must be equal to or greater than the number of seed word lists. Seed words should be chosen based on the aspects of bias/identity you wish to investigate in the corpus. α and β control the distribution of topics over documents and the distribution of topics over words, respectively (Blei et al., 2003). The z-weight parameter controls how much influence seed terms are given[4]

The models produced by pSSLDA are saved and a summary of the highest-ranking words in every topic is provided for analysis. We also provide information about the number of occurrences of the seed words.

Seed Words. Terms which define the start-state of a topic. Seed words are stored as text files (one term per line), and it is important that the seed words have undergone the *same preprocessing steps* as your corpus so that they will match tokens. Choosing seed words can be tricky, and we recommend checking the frequency of your terms before training (if a term appears not at all or very infrequently, you might want to replace it).
default value: We provide several pre-created lists in EQUITBL for investigating gender biases.

Number of topics (t). How many topics should be trained. You must have at least as many topics as seed word lists, i.e. the number of seeded topics must be less than or equal to the total number of topics.
default value: 10

z-weights. Controls how much "influence" a seed term has. Currently, EQUITBL sets all seed terms to uniform z-weight values, although there is some work in progress to allow variable values (e.g. in cases where seed terms should "belong" to more than one topic).
default value: 5.0

α. Controls the distribution of topics over *documents*. A value of $\alpha = 1$ creates a uniform distribution (any topic or combination of topics is equally likely). Values of $\alpha > 1$ encourage documents to be close to a (single) topic; values of $\alpha < 1$ encourage a mixture of topics (so a document is likely to be 'between' topics).
default value: 0.33

β. Controls the distribution of topics over *words*. Higher values of β will encourage each topic to be more strongly associated with fewer terms; lower values will "spread out" topics over large portions of the vocabulary.
default value: 0.2

Number of samplers. How many parallel samplers should be used during training.
default value: 10

Random Seed. For the purposes of repeatable experimentation, we manually set a random seed in the config file. If you want to train more than one model to get a better overview of the corpus (encouraged!) while keeping other parameters the same, you will need to vary the random seed.
default value: 237

Number of Samples. How many times Gibbs sampling should be repeated during training.
default value: 1000

[4]Note that where we refer to 'seed terms' the pSSLDA documentation uses 'z-labels.'

2.5 Visualization

To make interpretation of the results simpler, we provide visual representations of $p(w|t)$ (the likelihood of a word appearing in a given topic) and $p(t|w)$ (the "exclusivity" of a word to a topic) in the form of colored bar charts. We suggest producing two versions: one without seed terms visible (to avoid the 'intended' subject of the topic biasing your interpretation) and one with (in the final analysis, it is good to factor in how seed terms are weighted compared to other terms in the topic). Seed terms are identified with a *.

Example code can be found in `visualize_topics.py`

Exclusivity. Exclusivity represents the distribution of a word across various topics, $p(t|w)$. Higher values of $p(t_i|w)$ indicate that a word is particularly *unlikely* to appear in topics other than t_i, meaning that word may be a stronger indicator of the theme of topic t_i. By default, visualization of topics includes exclusivity of terms for that topic.

2.6 Qualitative Analysis

A series of questions (appendix A) is provided to facilitate qualitative analysis of topics given their visualizations. Take one topic at a time, and follow the questions in order.

At this point it is also possible to pull some documents from the corpus and do close readings of them in order to find out how the discovered themes actually appear in the texts.

3 Notes on Design and Default Parameters

Through a combination of experimental investigations (*section 3.2*) and informal workshops (*section 3.1*), we refined our default hyperparameters and the way we present the results of the topic models, respectively. Throughout these investigations, we used the English mainstream news corpus described in Devinney et al. (2020), consisting of 200000 news articles across a wide variety of subjects. We preprocessed the texts, lemmatizing and part of speech tagging tokens. We created separate topic models to investigate gendered and religious biases.

We set the number of topics (t) to be 15, as this seemed to work well for the context of our test corpus However, appropriate t values are extremely corpus-dependent, and will vary both with the size of the corpus and the range of material covered. As such, we recommend using unsupervised TM to see what captures the corpus well overall.

Not finding that z, the "seed weight" parameter, had much effect on the strength of the seed terms within our seeded topics, we kept the default setting from pSSLDA ($z = 5.0$).

Finally, we tested α and β values in combination with each other. We found that in general, α makes less of a difference in seeded topics than β. Larger β values gave us better focus, with topics having more exclusive terms; and smaller α gave us stronger themes. The combination produces topics that contain fewer but "more important" terms, which is desirable because the chunked documents are very short.

3.1 Workshops

Over the course of two workshops (one hybrid, one completely virtual) we gathered colleagues in both gender studies and computer science[5] and presented them with our baseline experiment ($t = 15, z = 5.0, \alpha = 0.33, \beta = 0.2$, and a consistent random seed). Based on feedback from the first workshop, we refined some of the visualization strategies (e.g. adding exact numerical data to the charts) and the way we explained what the different elements represent. Regardless of whether their expertise was in computer science or gender studies, all our participants found the visualizations easy to interpret and draw conclusions about when working in small discussion groups. Bringing the large group back together revealed that, while each group may have have noticed different nuances or reasoned slightly differently, the overall patterns discovered were consistent. This suggests that our method can be used to reliably (re)produce consistent analyses.

[5]Most participants had little to no professional overlap in these fields, but were interested in learning more about the "other" field.

3.2 Experimental Investigation

First, we explore how much data is required (and how much time and memory it takes) to train a useful topic model, by randomly selecting subsets of our full 200000 article (800MB) corpus. We find that for lists with sufficiently frequent seed words (such as gendered pronouns), even a small corpus of 25000 articles results in coherent topics, measured both by human evaluation and a C_{UCI} (see, e.g., Röder et al. (2015)) coherence score (Table 1). However, there is some drop off in the coherence and it is possible that issues with infrequent seed words may be exacerbated in smaller data sets. Figure 1 shows that both training time and the maximum amount of memory taken up during training scale linearly with the amount of data provided.

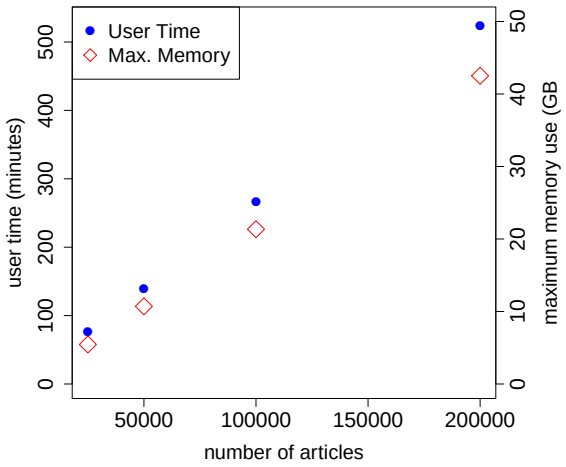

Figure 1: User time (actual computation time + operating system overhead) and maximum memory required for training a topic model by data size.

Model	Coherence
baseline (200000 articles)	0.435
baseline (unsupervised)	0.414
100000 articles	0.434
50000 articles	0.442
25000 articles	0.317

Table 1: Coherence scores by data size. The C_{UCI} coherence metric is calculated individually for each topic, and the arithmetic mean is reported.

Across the data size experiments, we find some variation in the actual subjects of the themes. However, further investigation with the full corpus and varying the random seed suggests that this is more due to the initial state of the model. Thus, training several topic models might produce a more thorough understanding of the patterns in the data.

We attempted to test the method on religious seed terms, but due to their infrequence in our corpus the resulting topics were largely incoherent or, when coherent, unrelated to religion. This relates to issues of identifying nonbinary themes rather than merely "ungendered" themes, due to underrepresentation.

We also test the effects of more directly manipulating the corpus based on our gendered categories, to see if it is possible to "curate" a corpus where particular gendered associations are minimized or the effects of data imbalance are eased. To do so, we tried both adding documents for the most underrepresented group (nonbinary people) and removing "strongly gendered" documents (i.e. those with the highest $p(d|t)$ for the gendered topics in the baseline model), both one topic at a time and all three gender categories at once.

The latter brought to light a particular methodological challenge: because the topics can vary by model a different random seed may have flagged other documents for removal.

In general, we find that the EQUITBL methodology is good at finding the hegemonic themes embedded in language corpora and presenting information in ways that are easy to discuss and digest. This allows us to examine problems of "data bias" from multiple angles. However, it has notable shortcomings. As observed in attempting to analyze our dataset for religious biases, when the seed terms used to try and find patterns are very rare or the group in question is not often discussed explicitly using those terms, TM fails to "pick up" the desired topics.

4 Conclusion

We suggest a mixed-methods approach for analyzing bias in text corpora, and provide tools for both the quantitative calculations and qualitative analysis aspects of this approach. We demonstrate some of the potential uses of these tools through exploring gendered patterns in a corpus of English-language news articles.

The 'downstream' effects of curating a dataset based on topic modeling results remains an open question, which we plan to investigate in future work. Although we have done some limited investigations on EQUITBL's efficacy in Swedish, it remains to be seen how it performs with other non-English languages (especially in terms of identifying gendered patterns in languages with grammatical gender or without gendered pronouns). We welcome any feedback or suggestions for improving the tool.

In addition to finding how biases and stereotypes manifest in a corpus, the codebase may also be of interest for researchers in other fields such as the Digital Humanities, as different seed words may be applied to explore a wide variety of potential themes within a text dataset.

4.1 Limitations

Although this method works for any language from which a topic model can be trained, we have only demonstrated it using an English corpus. For other languages, preprocessing, settings and strategies for seed terms will need to be adjusted. For example, in languages where third-person pronouns do not reliably correlate to with social gender (and not grammatical gender), it would not be appropriate to use them as seed terms for detecting gender bias. Additionally, seed terms need to be sufficiently frequent within the corpus for a coherent topic to form around them. When this does not occur, it may be a sign of *erasure*, a particular form of bias.

We chose topic modeling as it is computationally quite lightweight. However, our strategy of using sliding windows of relatively few tokens means there are more "documents" in our dataset, requiring larger matrices and therefore more memory during sampling. The random seed element of pSSLDA means that more than one model should be trained and analyzed, to ensure that results are robust and not merely due to chance initialization. We chose pSSLDA for our experiments because of its convenient seeding mechanism and efficient parallel implementation. We have so far not tried out other modeling options.

Finally, although we provide material to guide researchers through the qualitative aspects of the analysis, analyzing bias in this way will always require both time and expertise. Familiarity with the contexts, social structures, groups, and power dynamics at play are necessary. This being said, we believe that the methodology we have outlined, or variations thereof, will be useful for bias analysis in various settings.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blodgett, S. L., Barocas, S., III, H. D., and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Cao, Y. T. and Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Collins, P. H. (2000). *Black Feminist Thought*. Routledge, New York, NY, USA, 2nd edition.

Devinney, H., Björklund, J., and Björklund, H. (2020). Semi-supervised topic modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online). Association for Computational Linguistics.

D'Ignazio, C. and Klein, L. F. (2020). *Data Feminism*. The MIT Press, Cambridge, Mass., USA.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Maudslay, R. H., Gonen, H., Cotterell, R., and Teufel, S. (2019). It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5267–5275.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of The Eighth ACM International WSDM Conference*.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1651–1661.

A Worksheet for Interpreting Topic Models

Note: we recommend producing two versions of each topic visualization (one with seed words hidden, and one where they are visible) and looking first at the version with the hidden seed words. Extra space is provided to record your observations at each step.

For each topic visualization, look first at the lengths of the bars ($p(w|t)$) and consider the:

- **Nouns (NN)**: How are they related to each other? Do one or more specific themes emerge?

- **Verbs (VB)**: Are they, like the nouns, related to specific themes? Do they relate to activity or passivity (like running vs. sitting)?

- **Adjectives (JJ) and Adverbs (RB)**: How are they related to each other? Do one or more specific themes emerge?

Then, look at:

- **Exclusivity** ($p(t|w)$, represented by bar colors). Consider the words that are strongly exclusive to each topic – do they suggest particular themes?

And finally, consider the:

- **Themes** based on all of the above, is the topic associated with particular phenomena or characteristics (coherent themes)? Or is it represented as more "neutral" (no or less coherent themes)? *Try to summarize the theme(s) present in each topic in a few words or a short sentence.*

Based on these themes, which categories do you think belong to each topic?

Now, if you have been looking at visualizations without seed words, it is time to look at the versions where the seed words are not hidden. Look at the seed words for each topic.

How strongly do the seeded words appear in each topic?

Are pronouns in the topic more commonly nominative (subject form) or accusative (object form)?

Does this information change the evaluations of the summarized themes from the previous set of visualizations?