Race-Blind Charging

Stanford CS224N Custom Project
Mentor: Megan Leszczynski

Madison Coots
Department of Computer Science
Stanford University
mcoots@stanford.edu

Abstract

In jurisdictions across the United States, prosecutors make highly consequential charging decisions using police incident narratives that contain information about the race of the suspect. Recent studies have shown that there is reason for concern that the judgments made by the prosecutor may suffer from explicit or implicit racial bias. Past work to develop an algorithm to redact explicit mentions of race and other race-related information has been limited to the use of regular expressions that mask a set of predetermined information classes (such as names and locations). While these methods have been successful in redacting the targeted information while also preserving the legibility of the narratives, they have been limited in their ability to obfuscate latent race-related information and their reliance on human input for near-perfect redaction of person names. In this paper, we apply several deep learning approaches to the problem of obfuscating a suspect's race through redaction. We make use of pretrained large language models to mitigate data availability issues and ultimately show that the use of unsupervised pretrained models fine-tuned on downstream tasks like named entity recognition are competitive with the performance of past algorithms designed for this redaction problem and notably do not require additional human inputs to the model.

1 Introduction

Under the Equal Protection Clause of the Fourteenth Amendment, defendants in the United States criminal justice system are constitutionally protected against having immutable features such as race and gender influence any of the decisions made at various stages in the criminal process. However, there is plentiful evidence that would suggest that these characteristics regularly factor into high stakes decisions [1], [2], [3]. One such decision point is that of deciding whether to charge or dismiss a criminal case and on what grounds. Due to the way in which prosecutors are presented information describing the alleged crime, there is reason for concern that the judgments made by the prosecutor may suffer from explicit or implicit racial bias.

There has been past work to develop an algorithm using regular expressions to redact explicit mentions of race and other race-related information from the incident narratives that are central to the prosecutorial decisions [4]. While this work has proven successful in redacting most race-related information and preserving the legibility of the narratives, machine classifiers trained on the redacted narratives have still been able to achieve high accuracy in correctly classifying the race of the suspect described in the narrative, indicating an imperfect redaction for the ultimate goal of obfuscating the race of the suspect. Moreover, these methods rely on additional input manually extracted and provided by a human on a per narrative basis in order to ensure that person names are consistently redacted from the narratives.

In this paper, we explore several approaches that leverage BERT, a pretrained language model, to make redactions without a specified rule set and without additional human input. Specifically, we investigate the usage of a BERT model fine tuned on the following tasks: token classification,

named entity recognition, and masked language modeling. Through these experiments, we show the benefits of using pretrained language models in data scarce settings such as this one, and ultimately demonstrate that the performance associated with the use of pretrained language models such as BERT for named entity recognition alone achieves results that are highly competitive with those of currently used methods. Moreover, our approach does not require human inputs, and thus guarantees the scalability of its performance.

2 Related Work

The concept of performing automated text redactions has been studied primarily in the context of redacting personally identifiable and otherwise sensitive information (i.e. names, birthdays, social security numbers), especially in the clinical setting. Past work has applied fine-tuned pretrained large language models, like BERT, on various natural language understanding tasks including named entity recognition, intent detection, and dialog act classification for the purpose of identifying information to redact [5]. Other work has presented alternative approaches to the task of text redaction that draw methods from differential privacy [6], active learning [7], and other forms of deep learning such as recursive neural networks [8].

Notably, however, "Blind Justice: Algorithmically Masking Race in Charging Decisions" [4] is the first paper to present a redaction method for the purpose of redacting explicit mentions of race and other race related information. This paper is what has primarily motivated the explorations presented in this paper and their application to real police incident narratives used in charging decisions. The algorithm presented in [4] distinguishes itself through its use of an unsupervised learning approach, namely the use of a predefined set of regular expressions to match the following types of information: 1) explicit mentions of race; 2) select physical descriptors, including hair and eye color; 3) individuals' names or nicknames; 4) location information, including neighborhood names and street addresses; and 5) officer names, given that prosecutors may remember where officers are stationed. In addition to using regular expressions, this algorithm uses a combination of a named entity recognition model and human input to ensure that person names are successfully redacted from the narrative. Specifically, as a separate input, the algorithm accepts a list of names of the people mentioned in the narrative that must be manually extracted and supplied by a human. The motivation for relying on an NER model in additon to human input is that the NER model used in [4] (`spaCy`) apparently struggled with consistently identifying non-European names. While the use of the human input enables the algorithm to achieve extremely high accuracy in redacting named entities from the incident narratives, the scalability of this algorithm and its performance are limited by its dependence on this human input.

Lastly, classification models trained on the redacted narratives produced in [4] reveal that the algorithm fails to fully redact all latent race information in the narrative. Specifically, the classifier trained to predict the presence of a black suspect in these redacted narratives still achieves an AUC of 0.75.

3 Approach

An effective model for redacting race information from free-text narratives needs to balance two objectives: 1) minimizing the number of redactions made so as to not impede legibility of the narrative, and 2) minimizing the amount of latent race information left in the narrative. There exists an inherent tension between these two goals. One could trivially minimize the number of redactions by making no redactions, but fail to remove any race information from the narrative. In contrast, one could trivially minimize the amount of latent race information by redacting everything, but this would render the narrative incomprehensible. Therefore the ideal deep learning model applied to this data would, in some sense, factor both of these goals into its loss function. However, this supervised approach would require a considerable amount of training data (likely orders of magnitude more than what was available at the time of experimentation) in order to reliably learn how to optimally perform redactions with the two aforementioned goals in mind.

In this section, we detail a supervised deep learning approach for learning how to make redactions using a pretrained BERT model for binary token classification. Through this experiment, we demonstrate the limitations of supervised deep learning approaches in the absence of larger amounts of training data. We then present two alternative unsupervised methods using pretrained BERT models

for named entity recognition (NER) and masked language modeling (MLM) to explore the viability of using pretrained language models without performing supervised learning for this redaction task.

3.1 Background on BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pretrained language representation model that was released in 2019 [9]. It was trained on unlabeled data extracted from the BooksCorpus (800M words) and English Wikipedia (2,500M words) using a "masked lan- guage model" (MLM) pre-training objective. When compared to other recent language representation models, BERT's key distinguishing characteristic is its use of deep bidirectional representations from unlabeled text. Specifically, during pretraining, BERT jointly conditions on both the left and right contexts across all layers of the network. This innovation results in two notable results: 1) a more nuanced understanding of language when compared to previous models, and 2) an ability to fine-tune BERT with just one additional output layer. This second feature allows BERT to easily be extended to a number of other natural language processing tasks including token classification, named entity recognition, and masked language modeling.

3.2 BERT Named Entity Recognition (NER)

In [9], the authors demonstrated that applying BERT to the CoNLL-2003 Named Entity Recognition (NER) task [10] achieves results competitive with the state of the art. Moreover, the results showed that the large BERT model trained on cased text data achieved the best performance. While not all race-related information may be classified as a named entity (e.g. hair color or eye color), our motivation for applying a fine-tuned BERT model for this task stems from issues described in [4] relating to issues with using off-the-shelf NER software for redaction of named entities in the narratives. Specifically, the authors in [4] employed `spaCy`'s NER model and encountered issues with its identification of many non-European names. Moreover, `spaCy`'s NER model was trained using deep convolutional neural network with residual connections [11], and we were interested in seeing how its performance compared to that of BERT, a transformer-based model.

To perform named entity recognition on our unredacted police incident narratives, we used a pretrained large BERT model fine-tuned on the NER task [12]. After feeding in the full narrative, the model returns a list of token labels for those classified as named entities. The named entities tagged by the BERT model included person names, organization names, locations, and other categories including nationalities, religions, and political groups. Using this list of tagged named entities, we then reconstructed the narrative, redacting those named entities identified by the model.

3.3 BERT Token Classification

Similar to the pretrained BERT model fine-tuned on the named entity recognition task, the pretrained BERT model for token classification outputs a label for each token in the narrative. However, because we desired a model that could correctly predict whether or not to redact a token, this model required labeled tokens, which stands in contrast to the other unsupervised approaches described in this paper. Therefore, in order to use this method, we used our gold standard redactions to produce token labels to provide as input to the pretrained BERT token classification model. Then, when applying the fine-tuned model to our test data, we used the outputted token labels to reconstruct the masked narrative.

3.4 BERT Masked Language Modeling (MLM)

Masked Language Modeling is a fill-in-the-blank task, in which a model uses the context (on both the left and right sides) surrounding a word that has been masked in order to try to predict that token [13]. For a word that has been masked, a masked language model produces a list of the top k most likely words.

In applying a pretrained BERT model fine-tuned on a MLM task, we reasoned that many race-related words that should be redacted from police-incident reports are likely hard to predict using MLM due to their specificity, i.e. person names, location names, and personal descriptors. Therefore, for words that did not appear in the top $k = 10$ predicted words generated by the BERT MLM model, we redacted that word from the narrative. In order to produce the entire redacted narrative with this

method, we first applied ia pretrained BERT tokenizer. We then masked each token in the narrative one at a time and stored the results. Lastly, we compared the top $k = 10$ most likely words predicted for a given masked token to the true token and redacted the word if and only if the true token was absent from the top k tokens. In Figure 1, we present a figure illustrating how different versions of the source text (where a different token is masked in each) are supplied as input to the BERT MLM model. We then process the outputs of the BERT MLM model on each version of the source text to produce a redaction.

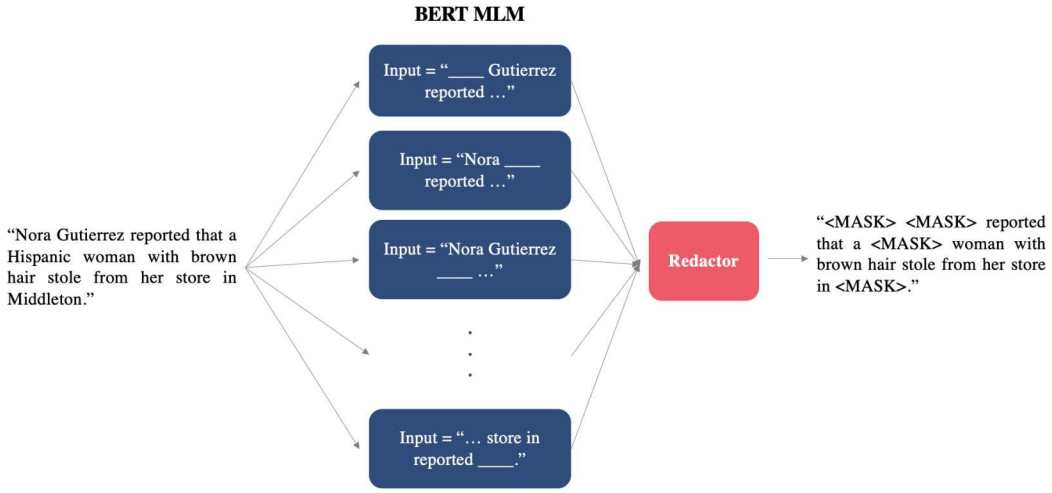

Figure 1: A figure illustrating how the outputs of the BERT MLM model are leveraged to produce a redaction. The source text on the left is fictional, but the output on the right is real output from our BERT MLM redactor model. We see that it successfully redacted person and locality names, as well as mentions of ethnicity.

3.5 Baselines

Regular Expression Algorithm

We used the regular expression algorithm that was described in [4] and used this as a primary baseline against which to measure the performance of our deep learning approaches. This algorithm uses regular expressions to match and redact the following types of information: explicit mentions of race; select physical descriptors, including hair and eye color; individuals' names or nicknames; location information, including neighborhood names and street addresses; and officer names, given that prosecutors may remember where officers are stationed. We note, however, that in order to consistently redact the names of individuals described in the narratives, this redaction algorithm accepts as input a list of the names of the people mentioned in the narrative to use for exact regular expression matching. In applying this algorithm as a baseline, however, we did not provide this list of person names to the algorithm in order to evaluate the algorithm's ability to perform redactions in the absence of human intervention. The authors of [4] kindly provided access to the repository containing the code for this algorithm for use as a baseline in this exploration.

spaCy Named Entity Recognition (NER)

We used the Python natural language processing package, `spaCy`, to perform named entity recognition on the raw narratives. Specifically, the algorithm redacted the named entities that fall under the following categories: person names, national or religious political groups, buildings, airports, highways, bridges, companies, agencies, institutions, countries, cities, states, mountain ranges and bodies of water. This baseline is useful for understanding the capabilities and limitations of off-the-shelf tools for named entity recognition, as well as understanding the amount of latent race information that remains after redacting only named entities.

Random Redaction

For each token in a narrative, we randomly redacted that token with some probability p. $p = 0$ corresponds to redacting no tokens and $p = 1$ corresponds to redacting every token. While unsophisticated, this method is capable of generating redactions useful for bounding the quality and performance of other redaction models. In our experiments, we set p equal to 0.15.

4 Experiments

4.1 Data

We use a set of real incident narratives provided by a jurisdiction in the United States. The original data collection provided by the jurisdiction is comprised of a total of 861 police incident reports originating from five different police departments. The narratives and race of the suspect involved in the incident are embedded in the incident reports, which were provided in a PDF format. Because of the format of the data, manual extraction of both the incident narratives and the race of the suspect was required. Efforts to automate extraction using optical character recognition (OCR) proved unfruitful due to inconsistencies in the formatting of the police reports and poor extraction quality.

Due to time limitations, we were only able to manually extract 100 narratives from the data. Each narrative contained approximately five to fifteen sentences. Once a subset of 100 (original narrative, suspect race) tuples were manually extracted from the data, we proceeded to manually redact the race-related information specified in [4] to create a set of narratives with gold-standard redactions. However, in contrast to the redactions described in [4] that still specified the type of information being redacted for increased legibility, we elected to instead perform "black bar" redactions to both speed up the manual redaction process and simplify the modeling task. In Figure 2, we show an example of how a fictional narrative would be redacted using this method. Once the manual redactions were completed, we had 100 (original narrative, redacted narrative, race)tuples to use for model training and evaluation.

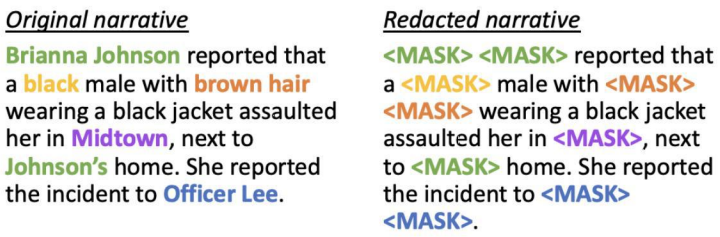

Figure 2: A fictional example showing the gold-standard "black bar" redaction and its unredacted counterpart. Mentions of race, physical descriptors, names, and locations are all identified and redacted with a mask token. Non-race-related descriptions (like "black jacket") are preserved.

4.2 Evaluation methods

When evaluating the quality of a model's redactions, there are two primary aspects of performance to consider: 1) the redacted narrative's quality (closeness to the gold-standard redactions), and 2) the model's ability to obfuscate the race of the suspect described in the narrative.

To evaluate the quality of the redacted narratives, we compare the model redactions to the gold-standard redactions, treating each token in the narrative as a binary decision of whether or not to redact. With this framework, we then compute the model's precision and recall as in [4], effectively evaluating the model's ability to emulate the gold-standard redactions.

In evaluating the model's ability to obfuscate the race of the suspect, we train a gradient boosted tree classifier (using `xgboost`) to determine the degree to which a machine classifier can correctly classify whether or not the narrative involved a Black suspect (as in [4]). To perform classification on the narratives, we first generate an embedding for each redacted narrative. This embedding was created by first removing stop words from the narratives and then mapping the remaining individual tokens to 300-dimension GloVe vectors [14]. When performing this mapping, the `<MASK>` token was skipped over. We then averaged the GloVe vectors for all tokens in the narrative to produce a 300-dimension

narrative embedding. This procedure for generating a narrative embedding is consistent with that described in [4]. However, due to the small number of labeled and redacted narratives available at the time of experimentation, it was necessary to perform feature reduction before training the `xgboost` classifier to mitigate the risk of overfitting. To accomplish this, we performed principal component analysis on the narrative embeddings to reduce their dimension from 300 to 30.

Effectively, this classification model serves as a quality assurance check, functioning as a proxy for the prosecutor to provide an estimate of the degree to which racial cues have been redacted from the narrative. To quantify the classifier's ability to correctly predict whether or not the suspect was Black, we use the AUC metric.

It is worth noting that there is an inherent tension between these two evaluation criteria. Because the gold-standard redactions were modeled off of those in [4], a model that perfectly reproduces these redactions would effectively mimic the behavior of the RegEx algorithm presented in [4], whose redactions fail to fully remove latent race information. On the other hand, a model could trivially minimize classification accuracy on whether a narrative involved a Black suspect by redacting every token in the narrative. Therefore, these two performance metrics together characterize the trade-off between redacting minimally and obfuscating latent race information to the greatest degree possible.

4.3 Experimental details

When fine-tuning the pretrained BERT model on the token classification task, we trained our model for 25 epochs using a learning rate scheduler.

As described in a previous section, when using the BERT Masked Language Model to inform redactions, we look at the top $k = 10$ words predicted by the model for a masked word. Additionally, when performing random redactions as a baseline, we redact a word with probability $p = 0.15$. Please refer to Section 4.2 for details on the implementation of the classification model used for redaction evaluation. In order to produce an unbiased estimate of the `xgboost` model's classification performance on unseen data, we train the model on approximately 60% of the labeled narratives, and test on the remaining 40%. The AUC scores reported in Table 2 are those achieved on the held-out test set.

4.4 Results

As presented in Table 1, the regular expression-based redaction algorithm from [4] achieves the best precision score. In this context, a high precision score means that the algorithm is not redacting many tokens that were not also redacted in the gold standard redactions. In other words, it is not over redacting. The BERT NER model achieves the best recall score. Here, a high recall score means that the model is correctly identifying and redacting most of the same tokens that were redacted in the gold standard redactions–essentially, the model is not under redacting.

In Table 2, we present the results of the classification model trained on the embeddings of the redacted narratives that serves as the audit on the efficacy of the redactions. Redaction models that successfully obfuscate latent race information should prevent the `xgboost` classifier from preventing high AUC scores. In other words, the closer the `xgboost` model's AUC score is to 0.5, the more effective a model's redactions are. From these results, we see that the random redaction model achieves the lowest AUC score. However, given that this was a fairly trivial model and baseline, we note also that the `spaCy` NER, BERT NER, and Blind Justice model redactions all achieve an AUC under 0.6, which indicate effective redactions. In Figure 3, we show the ROC curves for each of the models explored.

5 Analysis

Inspecting the redacted outputs of the models explored above provides key insights into what kinds of information the models redact easily, as well as those that they struggle with. For example, in the absence of a list of the names of the people involved in each narrative, the Blind Justice regular expression-based algorithm struggled to redact most person names. However, it nearly perfectly redacted other race-related pieces of information such as personal descriptors and explicit mentions

Model	Precision	Recall	F1
Blind Justice (RegEx)	**0.765**	0.279	0.409
`spaCy` NER	0.589	0.855	0.697
Random redaction ($p = 0.15$)	0.085	0.151	0.109
BERT NER	0.731	**0.917**	**0.814**
BERT token classification	0.082	0.348	0.133
BERT MLM redactor	0.263	0.632	0.371

Table 1: Evaluating redaction quality. Precision refers to the fraction of tokens that were redacted by the model and were also redacted in the gold standard. Recall refers to the fraction of tokens redacted in the gold standard that were also redacted by the model.

Model	`xgboost` AUC
No redactions	0.71
Gold standard	0.61
Blind Justice (RegEx)	0.58
`spaCy` NER	0.55
Random redaction ($p = 0.15$)	**0.54**
BERT NER	0.58
BERT token classification	0.55
BERT MLM redactor	0.64

Table 2: Auditing redaction efficacy. Here, we tabulate the AUC achieved by the `xgboost` classifier on the redacted narratives. The closer the AUC is to 0.5, the more effective the redactions are at obfuscating racial information in the narrative.

of race and nationality. It also consistently redacted location information such as street names and names of neighborhoods.

In slight contrast, both the `spaCy` and BERT NER models had very good performance on the consistent redaction of named entities, including people, places, and companies. However, these models struggled to consistently redact explicit mentions of race and ethnicity. In particular, the BERT NER model was able to redact ethnic descriptors tied explicitly to names entities such as "Hispanic" or "African American", but struggled to redact more ambiguous ethnic descriptors such as "white" and "black." The the `spaCy` NER model largely failed to redact any explicit mentions of race or ethnicity. Generally, it is surprising that the named entity recognition models achieved such considerably good performance across both evaluative metrics. The fact that the redactions produced by these two models were fairly close to the gold-standard redactions indicates that most of the information that should be redacted from these models are, in fact, just named entities. Additionally, the fact that redactions from both models achieved an AUC lower than the unredacted narratives confirms that the obfuscation of named entities from these narratives does tangibly reduce a classifier's ability to correctly predict race.

Examining the outputs of the BERT MLM redactor, we see that the model is fairly effective at redacting named entities, given that the likelihood of a specific name appearing in the top k most likely words for a masked token is exceedingly low. However, this model struggled significantly to redact most racial and ethnic descriptors, as well as typical physical descriptors such as hair and eye color.

Perhaps unsurprisingly, given the small amount of labeled training data available at the time of experimentation, the BERT token classification model (our only supervised approach) produced low quality redacted narratives when compared to the gold-standard. In fact, its precision and recall scores were better than only those of the random redaction baseline. This poor performance likely indicates that attempting to learn redaction rules from labeled data is simply unfeasible in the absence of large amounts of training data. Therefore, in the absence of additional data, this result underscores the need for an unsupervised approach for this redaction task.

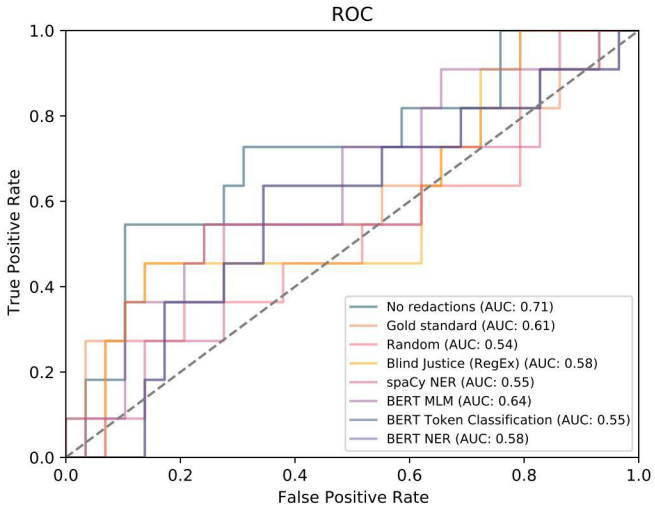

Figure 3: The ROC curves for the `xgboost` classifier trained on the redacted narratives from these redaction models.

6 Conclusion

In completing this project, we aimed to develop a redaction model that successfully obfuscated latent race information from police incident narratives while redacting as little information as possible. In contrast to pass work in this space, we experimented with several deep learning approaches that drew on the usage of large pretrained language models fine-tuned on various tasks to mitigate learning issues associated with having limited amounts of data. From these experiments, we discovered that supervised approaches to this redaction are largely unsuccessful due to limited training data. In order for supervised approaches to be successful, we would probably need a few orders of magnitude more data points. That is not to say, however, that obtaining large amounts of data is impossible. Recent work in the space of generative data synthesis has demonstrated that it is possible to artificially simulate infinite amounts of data using descriptive generative models, like probabilistic context-free grammars [15]. One can imagine writing such generative models to produce synthetic labeled data. With this data, it would be feasible to train a dual-objective deep learning model to simultaneously minimize redactions and minimize classification accuracy on the redacted narratives.

In the absence of such large amounts of labeled training data, however, there is still room for future work in improving the performance of unsupervised methods. Looking forward, we hope to further expand this work through the use of ensemble modeling. For example, for methods like the Blind Justice algorithm and BERT NER model, where one of which has high precision and the other high recall, it could prove fruitful to ensemble these methods to produce a redaction that obfuscates the union of the redacted token sets from each model. Moreover, we hope to improve our modeling and reliability of our results with the use of additional labeled training data. Ultimately, we hope that this work contributes to the conversation surrounding the promise for machine learning to play a prominent role in the mitigation of racial biases and other social injustices throughout the American legal system.

References

[1] L. Stolzenberg, S. J. D'Alessio, and D. Eitle. Race and cumulative discrimination in the prosecution of criminal defendants.

[2] B. Kutateladze, V. Lynn, , and E. Liang. Do race and ethnicity matter in prosecution? a review of empirical studies. technical report, vera institute of justice.

[3] T. W. Franklin. The state of race and punishment in america: Is justice really blind? journal of criminal justice.

[4] Alex Chohlas-Wood, Joe Nudell, Keniel Yao, Zhiyuan (Jerry) Lin, Julian Nyarko, and Sharad Goel. Blind justice: Algorithmically masking race in charging decisions. 2021.

[5] David Ifeoluwa Adelani, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. Privacy guarantees for de-identifying text transformations, 2020.

[6] Natasha Fernandes, Mark Dras, and Annabelle McIver. Generalised differential privacy for text document processing, 2019.

[7] Amir Feder, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, 107:103436, 2020.

[8] Jan Neerbek. Sensitive information detection: Recursive neural networks for encoding context. *ArXiv*, abs/2008.10863, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

[10] Erik F Tjong, Kim Sang, and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition.

[11] Matthew Honnibal. Spacy's named entity recognition model: Incremental parsing with bloom embeddings and residual cnns.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing.

[13] Wilson L. Taylor. Cloze procedure: A new tool for measuring readability.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation.

[15] Ali Malik, Mike Wu, Vrinda Vasavada, Jinpeng Song, Madison Coots, John Mitchell, Noah Goodman, and Chris Piech. Generative grading: Near human-level accuracy for automated feedback on richly structured problems, 2021.