Examining risks of racial biases in NLP tools for child protective services

Anjalie Field*
afield6@jhu.edu
Johns Hopkins University
Baltimore, MA, USA

Amanda Coston
acoston@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Nupoor Gandhi
nmgandhi@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Alexandra Chouldechova†
achoulde@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Emily Putnam-Hornstein
eph@unc.edu
The University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

David Steier
steier@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Yulia Tsvetkov
yuliats@cs.washington.edu
University of Washington
Seattle, WA, USA

ABSTRACT

Although much literature has established the presence of demographic bias in natural language processing (NLP) models, most work relies on curated bias metrics that may not be reflective of real-world applications. At the same time, practitioners are increasingly using algorithmic tools in high-stakes settings, with particular recent interest in NLP. In this work, we focus on one such setting: child protective services (CPS). CPS workers often write copious free-form text notes about families they are working with, and CPS agencies are actively seeking to deploy NLP models to leverage these data. Given well-established racial bias in this setting, we investigate possible ways deployed NLP is liable to increase racial disparities. We specifically examine word statistics within notes and algorithmic fairness in risk prediction, coreference resolution, and named entity recognition (NER). We document consistent algorithmic unfairness in NER models, possible algorithmic unfairness in coreference resolution models, and little evidence of exacerbated racial bias in risk prediction. While there is existing pronounced criticism of risk prediction, our results expose previously undocumented risks of racial bias in realistic information extraction systems, highlighting potential concerns in deploying them, even though they may appear more benign. Our work serves as a rare realistic examination of NLP algorithmic fairness in a potential deployed setting and a timely investigation of a specific risk associated with deploying NLP in CPS settings.

*Also with Carnegie Mellon University.
†Also with Microsoft Research NYC.

This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '23, June 12–15, 2023, Chicago, IL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594094

KEYWORDS

NLP, bias, race, child protection system, CPS, text processing

ACM Reference Format:
Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in NLP tools for child protective services. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3593013.3594094

Natural Language Processing (NLP) models are well-known to absorb and amplify data biases. A plethora of research has shown that models exhibit gender bias [58], with more recent work also examining dimensions like race [19], disability [25], and mental health status [33]. Despite these findings, understanding of model biases in realistic settings remains limited. Most work focuses on developing benchmark tasks and draws data from laboratory psychology studies or large public records [1, 19]. Curated data sets facilitate reproducibility and controlled experiments, but they can fail to articulate what is being measured, contain ambiguities, and may not reflect real deployment settings [2].

At the same time, despite concerns about bias, practitioners are increasingly turning to machine learning in high-stakes settings such as public services, hiring, education, and criminal justice [7, 9, 46, 52, 60]. We focus on one such setting: the child protection system (CPS). CPS agencies have copious free-form text notes written by CPS workers, which contain extensive details, including professional assessments of family situations and needs [14, 52]. Undirected manual reviews of notes is challenging for caseworkers and supervisors when, for example, making time-sensitive decisions or examining a newly assigned case [42]. Further, needing to spend time on administrative tasks like reviewing notes rather than working directly with families is associated with higher caseworker turnover and worse experiences for affected children [56, 57]. As the potential for NLP to aid in processing expert-written notes has been demonstrated in domains like healthcare and law [26, 28, 59, 66],

CPS agencies are actively seeking NLP tools to extract and deliver information from unstructured data [24, 40, 42, 61], and some research has additionally discussed implications of leveraging these data in predictive risk models [52, 53]. Predictive risk models have already been implemented to inform various aspects of CPS practice, such as which allegations are screened-in for investigation [10, 38, 52] or which investigations are prioritized for supervisory review [44], but existing models primarily rely on tabular structured data [52].

In this work, we examine the algorithmic fairness of NLP technology in CPS settings. Research on NLP model performance over real high stakes data is extremely difficult, given challenges around data privacy and forming partnerships with practitioners. Unlike research that constructs data sets to probe NLP model bias, our work serves as a rare opportunity to benchmark biases in a realistic setting. Furthermore, families involved in CPS already express mistrust in "the system" [3] and there is pronounced criticism of using algorithmic tools in any capacity [18]. Thus, our work is also a timely exploration of one potential risk of deploying NLP in CPS settings: algorithmic unfairness.

We focus specifically on disparities in model performance over case notes about black and white families based on decades of research demonstrating racial disparities in the child protection system in the United States [15, 22, 39, 47, 48, 62]. In comparison to white children, black children are disproportionately involved in the system [12], they may be more likely to be reported for abuse by doctors even with similarly severe injuries [31], their referrals are investigated at higher rates [5], and in some states they are placed in foster care at higher rates [6]. Institutional racism in child protective services has also been attributed to the links between the child protection system and other systems like mental health services, criminal justice, and education [22]. We investigate how deployed NLP models may amplify these disparities.

In Section 1, we first describe our primary data set: more than 3 million contact notes written by caseworkers, supervisors, and service providers in the Department of Human Services (DHS) in one anonymous USA county.[1] Next, we present initial statistics surfacing possible racial disparities in the text data (Section 2). We then examine two types of NLP models that could be deployed in this setting: the incorporation of text data into an existing risk predictive tool (Section 3) and information extraction tools, specifically named entity recognition (NER) and coreference resolution (Section 4). Despite substantial research on gender bias in coreference systems [50, 65], to the best of our knowledge, our work is the first consider racial bias. Our results show consistent racial bias in NER, possible biases in coreference resolution, and no evidence of increased racial bias in risk assessment. Thus, while risk assessment systems are already heavily scrutinized [53], our work highlights one way deployment of information extraction systems could also result in direct harms, even though they are further removed from direct decision-making and may appear more benign than risk prediction.

Finally, we emphasize algorithmic fairness is only one risk associated with developing NLP technology in CPS settings. Our focus on this particular risk is motivated by extensive NLP research on synthetically probing model biases with few examinations of model performance in realistic settings, but broader sociotechincal forces must be considered [1]. Thus, we conclude by discussing some of the risks beyond algorithmic fairness identified in prior work and surfaced in our analysis [53]. Our work is a rare documentation of direct harms that can result from algorithmic bias in deployed NLP systems. To the best of our knowledge, it is the first work to consider algorithmic unfairness in NLP technology for CPS settings.

[1]All research was conducted with IRB approval and under a data sharing and protection agreement with the county.

	Average	Min.	Max.
# tokens per note	156.59	0	2,915
# notes per case	128.36	1	4,831
# notes per referral	8.55	1	672

Table 1: Overview statistics for 3.1M contact notes associated with cases and referrals.

1 DATA: CPS CONTACT NOTES

Child welfare cases typically begin with a referral, where someone (the "reporter") contacts social services with concerns about a child. A call-screening staff member then makes a *Call Screen Decision*: whether or not to investigate the allegations made in the referral. If the referral is *screened in*, a caseworker then conducts an investigation, which may involve interviewing relevant contacts and conducting assessments. In many states, caseworkers must complete the investigation in a fixed amount of time, such as 60 days [4]. Based on the investigation, the caseworker decides whether or not the family should be accepted for services by the child protection agency (*Service Decision*). If the family is accepted, a case is opened. Cases can stay open for varying lengths of time, and families may receive a range of services, such as housing support or addiction treatment often through external *service providers*. Caseworkers, supervisors, and service providers write copious notes throughout the duration of a case, including during the investigation phase before a case is actually opened.

In this work, we investigate a data set of 3,105,071 contact notes, which consists of all *contact notes* written by CPS workers in the Department of Human Services (DHS) in an anonymous county from approximately 2010 to November 23, 2020 (notes prior to 2010 were inconsistently digitized in the current system). Table 1 presents overview statistics of the data. Contact notes log communication between CPS workers and families. As shown in Figure 1, most notes record telephone or face-to-face contacts, such as visiting families at home or school, but notes can also be created for other forms of contact, such as emails. In addition to the primary data, we also reference associated meta-data, such as case open and close dates, lists of associated clients, and mappings between cases, referrals, and clients. Throughout this work, when we refer separately to *referral* notes and *case* notes we generally do not duplicate data, e.g. for a referral that turned into a case, we would not consider notes associated with the referral as also associated with the case, which is consistent with DHS data organization.

Research was conducted with full-board IRB approval and under a data sharing agreement with DHS. In accordance with these protocols, data was exclusively stored on a remote disk-encrypted

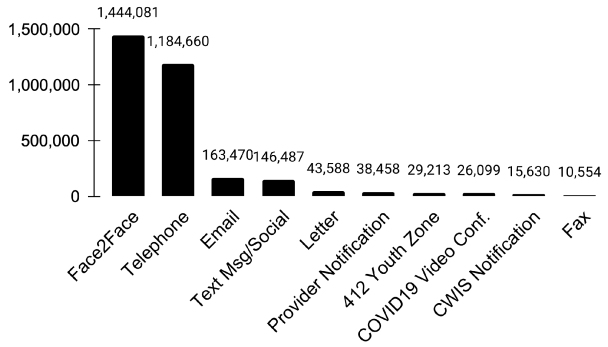


Figure 1: Histogram of contact types for 3.1M contact notes associated with cases and referrals. Contact types with <1000 notes are omitted for brevity.

	Num.	Avg. Word Count
Black-assoc. Referrals	50,490	1,485.13
White-assoc. Referrals	65,393	1,323.96
Black-assoc. Cases	7,525	50.81
White-assoc. Cases	6,637	69.86

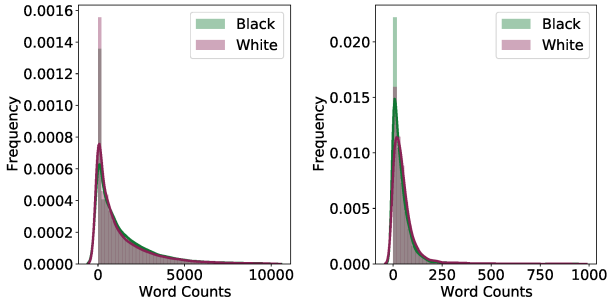


Figure 2: Length of notes associated with cases and referrals by race. For cases, word counts are normalized by the number of days the case has been open. Left histogram shows word counts associated with referrals and right shows for cases (truncated for readability). There are not consistent length differences by race.

server, with access restricted to approved members of the research team who underwent IRB-determined training, and researchers only accessed the server through secure encrypted connections. In general the standards of security and privacy upheld throughout this research were higher than those mandated for DHS contractors.

1.1 Conceptualization of Race

As discussed, we focus on investigating algorithmic racial bias motivated by limited research on algorithmic racial bias in realistic NLP systems and documented disparities in CPS settings. This investigation necessitates information about races of families involved in CPS. Our primary source of information about race is metadata entered by CPS workers, which means our conceptualiztion of race is race as perceived by CPS workers. This conceptualiztion is appropriate in some contexts, e.g. how implicit bias of CPS workers may manifest in text data, but as perceived race can differ from self-identified race [49], it suggest our results may not be fully reflective of experiences of families.

We focus specifically on white and black families and clients due to well-document disparities between white and black families in CPS settings and for clarity of analysis. Although a clear over-simplification, this focus does include much of the available data: out of all 234,818 clients listed on referrals, meta-data specifies 47.3% as "White", 33.7% as "Black or African American", and and 11.1% as "Unknown/Did Not Ask/Declined to Answer", with <8% of the data specifying client race as mixed or other than white or black. In general, rerunning statistics (e.g., Section 2) where we include other/mixed race clients as either black or white does not change results. Nevertheless, this simplification is a limitation in our work. We provide additional details on exact processing of race information for each task in task-specific sections (Section 2.1, 4.1).

2 RACIAL DISPARITIES IN CPS CASE NOTES

We first consider possible racial disparities in the raw text data, which could have implications for both training and deploying NLP models in this setting. For example, if there tend to be more and longer notes written about white families than black families, any models *trained* on this data may learn patterns disproportionally representative of white families. Similarly, any models *deployed* on this data could disproportionally benefit white families by retrieving or summarizing more information for white families from the underlying data. We investigate the research question: are there systemic differences in ways notes about black and white families are written? We focus on two types of differences: quantitative (volume of data) and qualitative (word choice).

2.1 Methodology

To examine differences in data quantity, we compute the average number of words in contact notes for referrals and cases for black and white families. We consider a case or referral to be black-associated if DHS metadata specifies >50% of associated clients as "Black or African American" and analogously define white-associated. We do not include cases and referrals where neither of these criteria are met (e.g. race of clients is unspecified, is other than black or white, or is mixed such that no one race is common to >50% of clients). We exclude 4,075/18,237cases and 27,015/142,898 referrals using this criteria, where almost all of them have mixed or unknown race information, i.e. only 91/4,075 excluded cases have a client with a race other than Black, White, or Unknown specified. Referrals have fixed time frames, but cases can span variable lengths of time. Thus, for cases, we normalize word counts by the number of days the case was marked as open in DHS systems. Normalizing by number of notes may seem more intuitive; however, we are interested in total volume of available data, not average note length, e.g., a case with many short notes may have lower average note length but more total data than a case with a few long notes.

To examine differences in data quality, we identify word-level language differences. We compute to what extent each word in the data is overrepresented in notes about black or white families as compared to all other notes using log-odds with a Dirichlet prior

Black-assoc.	Score	Nearest Match	White-assoc.	Score	Nearest match
Referrals					
she	52.19	he,that,stated	he	54.64	she,said
belt	47.37	spatula,paddle,spanked	heroin	41.87	Crack,ecstasy,crack-cocaine
her	45.39	his,him,and	PGF	36.08	PGGM,MGGM,GPGM
BM	37.90	BP,BPs,KCG	treatment	36.16	services,Outpatient,PND
bus	30.95	trolley,rides,stop	anxiety	34.25	unmedicated,schizophrenia,schizoaffective
shelter	25.11	motel,courthouse,AVAC	using	27.45	utilize,utilized,smoking
whooped	23.96	spanked,smacked,paddled	therapist	26.05	therapist-,school-based
Cases					
school	56.80	work,he,that	F	130.67	M,MGM,CW
housing	42.01	sub-iodized,HUD,housing/shelter	parents	59.26	mother,mom,children
informed	37.76	stated,reported,also	drug	37.65	D+A,DOA,Screens
pass	35.75	re-issued,passed,fail	methadone	36.55	crack-cocaine,THC,Crack

Table 2: Words overrepresented in notes about black and white families computed using log-odds with a Dirichlet prior [37]. Nearest-matching words that are not associated with the specified race are identified using cosine similarity of word embeddings. There are noticeable differences in topics and terminology in notes about children of different races. Commonly used DHS-specific acronyms are, BM: Birth Mother, BP(s): Birth Parent(s), KCG: Kinship Caregiver, PGF: Paternal Grandfather, PGGM: Paternal Great Grandmother, MGGM: Maternal Great Grandmother, D+A: Drugs and Alcohol. Others generally refer to service providers or common usage.

[37]. We then compare overrepresented words with common alternatives using 100-dimensional word embeddings trained over all 3.1M contact notes using skip-gram Word2Vec with a context window of 5. For each of the 100 most-overrepresented words, we identify the 3 words with the most similar (using cosine similarity) word embeddings that are not overrepresented (e.g. log-odds score < 0), discarding words that occur < 30 times in the data set. These common alternatives offer perspective on what terminology note writers could have used, which can yield broader insights. For example, if hypothetically we found that the term "deplorable" was more common in black-associated notes, this result would be difficult to interpret. By comparing with common alternatives, we might find that "disrepair" is not black-associated but is used in similar contexts, which would suggest that note writers use more negative language when describing housing needs of black families.

2.2 Results

Figure 2 reports the average number of words and word-count histograms in contact notes associated with clients of different races on referrals and cases. Generally, there are not consistent differences by race in data set sizes. Although there tends to be more text data for white-associated cases (69.86/day compared to 50.81/day), there tends to be more text data for black-associated referrals (1,485.13 compared to 1,323.96).

Table 2 presents a subset of the 100 terms most overrepresented in case and referral notes about black and white families and their 3 nearest neighbors. Numerous words are overrepresented in both case and referral notes (duplicates not shown for brevity). Associations do reveal possible content and style differences along racial lines. Words common in notes about black families focus on behavior, punishment, and basic needs, while words common in notes about white families focus on drug use. Words related to women and female caregivers are also more common in notes for black families whereas notes for white families contain more references to male caregivers. While these metrics likely reflect differences in events and reasons that families are referred to CPS, there are also racial associations in some near-synonyms, likely more reflective of style than content: notes about black children use "whooped" over "spanked", and "informed" over "stated". Manual examination of notes containing overrepresented terms shows that note writers often directly and indirectly quote clients and sources they interview, for instance notes about black families often contain African American English (AAE). Some notes quotation marks, e.g., *C reports that F tells him "sit down and listen"*, but language likely still reflects terms used by clients and sources even when not marked with quotation marks, e.g., *he said he was trying to get out of the room so he can go outside.*[2] Thus, terminology differences result both from terms used by note writers themselves and by people they quote. It is well documented that such subtle differences in language can lead to harmful biases in downstream tasks, e.g., off-the-shelf NLP models for toxic language classification are more likely to falsely classify AAE as offensive [13, 19, 51]. While much scope remains for further investigation into the origins and effects of these differences as well as closer examination of language variance beyond word-level metrics, these word statistics do suggest that there are systemic differences in notes about children of different races which could be absorbed and amplified by NLP models. We explore some such models in the following sections.

3 RACIAL DISPARITIES IN RISK ASSESSMENT

Public agencies are increasingly turning to algorithmic models with the goal of improving the consistency and accuracy of time-sensitive high-stakes decisions [7, 9, 16, 46, 52, 60]. These models reflect evolution from operator-driven checklists derived from regressions to machine-learning methods that draw on hundreds of

[2]Examples were modified to preserve privacy, not directly drawn from the data.

pieces of information [35]. Risk assessments tools in general have been criticized for failing to account for relevant individual context and for automating biases in the data [18, 48, 63]. In a survey of CPS algorithms, Saxena et al. [52] find that current tools rely on structured tabular data and suggest that augmenting the data features with natural language could be one way to incorporate context, improve model performance, and reduce bias. Though they dispute this suggestion in follow up work [53], the initial suggestion evidences how contact notes may appear useful data for risk prediction: they generally contain numerous details not captured in structured data. However, their incorporation could exacerbate many of the concerns around risk assessment: text is written by people and reflects their perceptions of events, which may or may not accurately reflect reality [17, 53]. There are numerous risks associated with incorporating text notes into risk predictive models, including reducing transparency, overfitting, increasing surveillance and privacy violations. In this section, we focus on algorithmic bias as one specific risk, and we provide additional discussion in Section 5.

In the anonymous county, DHS uses a predictive risk assessment tool to aid in call-screening. For an incoming referral, the tool presents call-screening staff with a score from 1 to 20 that aims to reflect the likelihood that the child will be placed (removed from home) within 2 years conditional on the referral being screened in. While the model has undergone changes, the original version was a logistic LASSO that selected 71 features from > 800 variables providing demographics, past welfare interaction, public welfare, county prison, juvenile probation, and behavioral health information on all persons associated with each referral. Some features are derived from previous interactions with the child welfare system, (e.g. the number of previous referrals associated with people on the new referral).

Chouldechova et al. [10] show that call-screening tools can exhibit miscalibration by race, and Cheng et al. [8] show that without caseworker oversight, they can result in a much higher screen-in rate for black children than white children. In this section, we integrate text features into the county's existing tool and analyze the impact on model performance, focusing on racial disparities. We examine several related research questions: Does integrating text features:

(1) increase model performance disparities for black and white children?
(2) increase model miscalibration with respect to race?
(3) increase the proportion of black children flagged as high risk by the model?

3.1 Methodology

Data and Features. We base our models on the same data and features used in the original version of the model, which encompasses referrals screened in for investigation from April 2010 to July 2014. The basic data unit is a referral–child pair: if a maltreatment referral had multiple children or if the same child was included in multiple referrals, we treat each unit as a separate data point. We use pre-constructed test and training splits, which were constructed to ensure no overlap in children or referrals between the training and test set, and we reserve 10% of the training set as a validation set. Although all child–referral observations contain structured features based on the current referral, not all families have had prior interactions with DHS and prior interactions could have been expunged. Thus, not all observations contain associated text notes. Our final data set consists of 28,769 training instances, 7,893 of which contain text data, and 14,417 test data points, 4,133 of which contain text data.

Models are trained to predict out-of-home placement within 2 years using structured and text features. As structured features, we use the same 818 features as early versions of the model. For text features, for each child in each referral, we pull contact notes for prior cases and referrals where the child was listed as a client, restricted to notes from the previous 365 days that contain the child's first name. We restrict data to the preceding year based on suggestions from DHS employees and early experiments with the validation set, which suggested that data from this time frame most improved model performance when compared to longer or shorter time frames. We preprocess the text by expanding common acronyms (e.g., F = father) using a list manually curated in consultation with DHS. We further remove first and last names of clients as listed in associated metadata and mask any additional people and location named entities identified using SpaCy.

Models. We examined four standard classification models: logistic regression, random forest, GatedCNN (a convoluted neural network (CNN)-based neural classifier shown to perform well on a similar task over medical text) [26], and a RoBERTa-based neural classifier [34]. As the random forest model achieved the best classification performance, likely due to data imbalance and limited high-dimensional training data, and it is most similar to previously investigated call-screening tools [10], we describe implementation and results from this model here and describe other models in Appendix A. We report results for the random forest model without text features (structured) and with text and structured features (hybrid). For the structured model, the feature-input is 818 dimensional structured feature vectors. To incorporate text features, we first trained a text model using a logistic regression classifier with TF-IDF-weighted bag-of-words features and a 10,000 word vocabulary. We then took the 500 words with the highest learned coefficients and the 500 words with the lowest (most negative) coefficients and constructed TF-IDF features from this refined 1,000 word vocabulary. We then concatenated these features with the 818 structured features, constructing 1,818-dimensional feature vectors. We do this feature selection because we found in early experiments over the validation set that concatenating all text features with the 818 structured features caused the model to ignore the structured features. All random forest classifiers used 500 trees.

Fairness Metrics. In order to examine (1), if text features increase model performance disparities, we examine the difference in model predictions with and without text features for black and white children (e.g., accuracy equity and error rates) using several performance evaluation metrics: area under the ROC curve (AUC), false positive rate (FPR), false negative rate (FNR), and the raw model scores outputted for children who were placed out of home (Average Positive Score) and who were not (Average Negative Score). In order to both provide realistic estimates of how these systems may operate when deployed in practice and highlight differences in performance when incorporating text features, we report metrics

Test set	Model	AUC	Avg. Pos Score	Avg. Neg Score	FPR	FNR
Full	Structured	75.77 ± 0.02	13.85 ± 0.00	8.70 ± 0.00	19.59 ± 0.01	43.96 ± 0.05
	Hybrid	76.27 ± 0.02	13.94 ± 0.00	8.69 ± 0.00	19.53 ± 0.01	43.60 ± 0.07
Examples with notes	Structured	69.83 ± 0.03	14.54 ± 0.01	11.19 ± 0.01	31.78 ± 0.06	40.51 ± 0.09
	Hybrid	71.88 ± 0.04	15.84 ± 0.01	13.24 ± 0.01	40.34 ± 0.08	29.80 ± 0.13

Table 3: Performance of structured and hybrid models trained and evaluated on predicting out-of-home placement. The feature inputs to the structured model are the tabular structured data, and the feature inputs to the hybrid is both the structured data and contact notes. Avg. Pos/Neg Score report the average predicted risk scores for true positive (placement occurred) and true negative (no placement) test data, where risk scores are computed by bucketing test predictions into ventiles. *Top:* Differences in model performance across the full test set ($n = 14,417$) are small. *Bottom:* Differences across the test set that contains text data ($n = 4,133$) show reductions in false negatives, but not in false positives.

Test Set	Model	$AUC_{black-white}$	Avg. Pos $Score_{b-w}$	Avg. Neg $Score_{b-w}$	FPR_{b-w}	FNR_{b-w}
Full	Structured	−0.26 ± 0.04	0.35 ± 0.01	0.66 ± 0.01	1.09 ± 0.06	−2.87 ± 0.13
	Hybrid	0.72 ± 0.04	0.34 ± 0.01	0.31 ± 0.01	0.35 ± 0.06	−4.00 ± 0.16
w/ notes	Structured	−1.83 ± 0.06	0.14 ± 0.01	0.72 ± 0.01	2.72 ± 0.10	−0.92 ± 0.18
	Hybrid	−1.05 ± 0.08	0.35 ± 0.01	0.68 ± 0.01	6.17 ± 0.13	−4.44 ± 0.22

Table 4: Performance disparities by race for the structured and hybrid models. The predictive disparities are largely comparable for the structured and hybrid models. We do not therefore find evidence that incorporation of text features increases aggregate measures of algorithmic unfairness in this setting. Raw performance values are reported in Appendix B.

over the full test set and over only the subset of the test data that contains text features. When computing metrics requiring a classification decision (e.g., FPR, FNR), we consider the 25% of test data with the highest raw output scores as having positive predictions. This percentage corresponds to the mandatory screen-in threshold that has been used by DHS. Additionally, we conduct bootstrap sampling with the training data and report average metrics and standard error values computed from 100 samples.

To evaluate (2), if text features increase model miscalibration with respect to race, we plot true placement rate for each bracket of model-predicted risk score, following Chouldechova et al. [10]. Both (1) and (2) compute results based on out-of-home placement values, but this proxy outcome may itself reflect racial bias. Thus, we examine (3), if text features increase the proportion of black children flagged as high risk by the model, by comparing the averages and distributions of predicted risk scores for black and white children. Overall, these metrics aim to capture if incorporating text features into the tool is likely to disproportionally harm black families referred to CPS by increasing the chance their referral is investigated or by creating racial disparities in model errors.

3.2 Results

Table 3 reports overall performance of the structured and hybrid models. Changes in AUC and in predicted risks scores are small, though the hybrid model does consistently outperform the structured model. The improvements are largely driven by increases in the risk scores for data points where out-of-home placement occurred, which decrease the false negative rate. However, the incorporation of text features does not decrease the false positive rate for data points with text features, nor result in lower risk scores for data points where out-of-home placements did not occur. In general, the model interprets the presence of associated text as an indicator of risk, as test data points with text are assigned higher scores by the hybrid model than the structured model, regardless of whether or not out-of-home placement occurred.

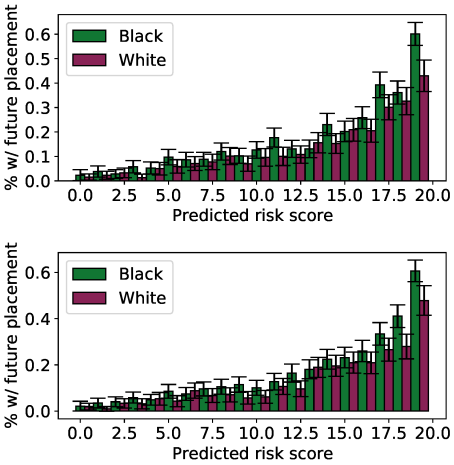


Figure 3: Calibration plots using structured (top) and hybrid (bottom) features (Chouldechova et al. [10] provides details on computing/interpreting calibration plots). We infer predicted risk scores by grouping the full data set into ventiles, but only display data points for black and white children in these figures. Both the structured and hybrid models display signs of miscalibration in the highest ventile.

Table 4 reports results for question (1), if text features increase performance disparities. Differences in model performance for black and white children are small overall, and the hybrid model does not

Model	$\%race_{b-w}$	$\%flagged_{b-w}$
Structured	3.18 ± 0.06	13.30 ± 0.11
Hybrid	2.73 ± 0.07	12.50 ± 0.12

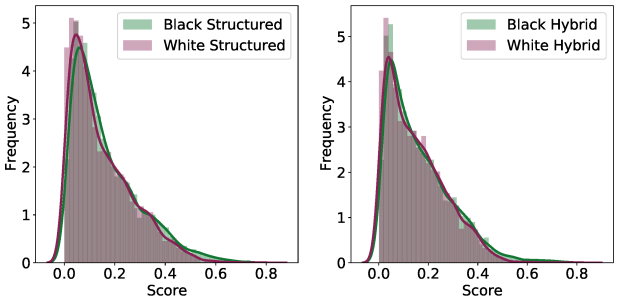


Figure 4: *Top:* difference in the percent of children of each race (black - white) that are flagged as high-risk under the specified model (%race), difference in the racial composition (black - white) of those who are flagged as high risk (%flagged). The hybrid model shows a reduced disparity in the percentage of black children who are flagged as high-risk. *Bottom:* histograms of the raw scores outputted by each model, divided by race. Under the structured model (left), score predictions for black children are right-shifted compared to white children. Under the hybrid model (right), the score distributions are nearly identical.

show greater performance disparities than the structured model. In identifying referrals without future out-of-home placement (e.g. FPR), the hybrid model slightly reduces disparities (1.09 to 0.35). In identifying referrals with out-of-home placements (e.g. FNR), the hybrid model slightly improves performance for black children more than white children.

Figure 3 addresses question (2) and compares risk scores using calibration plots. Both the structured and hybrid models display signs of miscalibration in the highest risk bracket. Specifically, a higher percentage of black children assigned the highest risk were placed out-of-home than white children, but the hybrid model does not appear any more miscalibrated than the structured model.

Finally, to address question (3), Figure 4 displays the percent of children of each race that are flagged as high-risk as well as the racial composition of those that are flagged for both models, allowing us to examine if the hybrid model is liable to increase the number of black families involved in CPS. Under both models, a higher percentage of black children are flagged as high-risk relative to white children, but the hybrid model reduces this disparity. This reduction is also visible in the histograms of raw model scores: under the structured model, scores for black children are right-shifted compared to white children, while under the hybrid model, the score distributions are nearly identical.

Overall, the changes in model performance when text features are incorporated suggest that text features do not increase aggregate measures of racial disparities in model predictions and may actually improve them. Prior work has shown racial bias is undoubtedly a concern in deploying algorithms in CPS settings [8, 10]. However, in the context of NLP research priorities, our results suggest that model debiasing may not be useful in this particular task, and focus on algorithmic bias may be over-estimating perceived risks and distracting from true risks, like reducing transparency and risking privacy violations. Our work offers corroborating empirical evidence to more theoretical discussions on how NLP research on debiasing can be misplaced [1]. Nevertheless, although we do not find evidence of increased bias in this specific experimental setup, we emphasize that this finding should not be interpreted as a generalizable lack of racial bias in the contact notes or in the predictive models used on those notes; results may differ under different models or subsets of the data.

4 RACIAL DISPARITIES IN INFORMATION EXTRACTION SYSTEMS

While risk assessment models reflect technology already in use in CPS settings, for text processing, CPS agencies are more actively interested in developing NLP systems that focus on aiding CPS workers in retrieving and organizing information rather than direct decision-making. For example, in 2018 one agency solicited proposals for tools that mined information from text, including identifying family support systems (e.g., names of extended family members who can provide care) and issues of concern (e.g., substance use, intimate partner violence) [40]. In the research community, Perron et al. [42] provide an example: using NLP to determine if a case note mentions a substance-related problem; in follow up work, Perron et al. [43] similarly explore using a rule-based entity recognition system for identifying opioid mentions.

We investigate racial disparities in named entity recognition (NER) and coreference resolution, which are necessary components of information extraction systems. While a domain-specific model might identify substance use, NER and coreference are necessary for resolving who in a case is involved: substance abuse for a parent is a different situation than for a child. There are also specific imminent use cases: although the county maintains structured records of people involved in cases, they are not always up-to-date. In the event of a crisis situation and a court orders a child to be removed from their home, NER models could aid a caseworker in finding immediate kinship placement options over a non-kinship or group home. For coreference, we focus on use cases previously explored for processing expert-written notes in the medical domain. This framework aims to directly identify mentions of *People*, *Problems*, *Treatments*, and *Tests* in addition to coreference links between them [59]. Our focus on these two tasks is additionally motivated by conversations with DHS employees about what NLP tools they would find useful and are actively considering deploying.

Furthermore, NER and coreference resolution are both well-established NLP tasks with existing off-the-shelf-deployable models, which makes their potential deployment a realistic scenario. While prior research has investigated gender or racial bias in these tasks over synthetic data [36, 50, 65], to the best of our knowledge, no work has explored more realistic settings. Here, we investigate: do existing NER and coreference resolution models exhibit performance disparities over notes about black and white families? This question aims to capture possible direct harms that could occur from model deployment. An NER system that has higher recall for names of white clients than black clients could result in black

		Full Names					First Names				
		# names-notes	SpaCy	NLTK	FlairNLP (ConLL)	FlairNLP (OntoNotes)	# names-notes	SpaCy	NLTK	FlairNLP (ConLL)	FlairNLP (OntoNotes)
Referrals	Black	95K	78.3%	83.5%	98.0%	95.6%	314K	68.0%	83.8%	97.2%	96.0%
	White	108K	83.4%	86.9%	99.1%	97.2%	368K	76.5%	88.4%	98.3%	97.3%
	B - W		-5.1%	-3.4%	-1.1%	-1.6%		-8.5%	-4.6%	-1.1%	-1.3%
Cases	Black	858K	72.85%	78.61%	97.18%	94.67%	6.7M	61.47%	81.67%	96.24%	95.21%
	White	538K	77.99%	83.16%	98.87%	96.76%	4.2M	72.79%	86.68%	97.99%	97.06%
	B - W		-5.14%	-4.55%	-1.69%	-2.09%		-11.32%	-5.01%	-1.75%	-1.85%

Table 5: NER model recall at recovering names in structured data from contact notes. Across all models, recall is higher for names of white clients than black clients. All performance differences between names of white and black clients are statistically significant.

children being placed initially in non-kinship homes more often, even when kinship placement options were available. In contrast, a coreference system that has higher recall at identifying *Problems* for black families and *Treatments* for white families could unfairly highlight risks for black families and supports for white families.

4.1 Methodology

NER. We evaluate four NER models: SpaCy "en_core_web_sm" default NER model (a transition-based parser trained on OntoNotes), NLTK's currently recommended NER module (a MaxEnt classifier trained on the ACE corpora) and two neural models based on document-level XLM-R embeddings, one trained on CoNLL-03 and one trained on OntoNotes implemented with the fairNLP packages [54]. We selected these models to cover a range of training data sets and architectures and also due to their ease of implementation, which makes them more likely for CPS agencies to adopt. For each model, we examine entities tagged with *PER* or *PERSON*.

For evaluation, we use existing county records of clients involved in cases and referrals as gold named entities, and we evaluate the ability of NER models to recover these names from contact notes. We conduct this analysis at a case/referral-client-note level, which best replicates a realistic scenario: e.g. a CPS worker may search through notes on a case to find all information about possible caregivers. We again focus on black and white clients and do not include clients who are of mixed race or other races. Our primary evaluation metric is recall, as output from an NER system would be read by a CPS worker who could easily disregard any incorrect names. CPS workers often use acronyms, so many clients listed on cases are not mentioned at all by name in contact notes. To account for this, we searched for the client's name in the relevant notes and only computed model recall over names that appeared in the text. In practice, NER models are more useful for identifying people not already listed in structured data; however, names in structured data offer us an existing evaluation set that is reflective of how well NER models can capture people described in notes. We focus on comparing model recall for names of black clients with names of white clients.

Coreference. We examine two neural coreference systems. The first is an extremely common coreference architecture where an encoder, mention detector, and antecedent linker are trained end-to-end [32]. The second is a state of the art variant of Lee et al. [32] that relies only on the start and endpoint of the span [30]. For both models we use a SpanBERT encoder [29] and train on OntoNotes [23]. As OntoNotes differs greatly from CPS contact notes, we employ continued training on more similar domains, which is a standard approach to overcoming established domain-transfer challenges for coreference [64]. We focus on two settings: (1) continued training on a limited subset of annotated contact notes (keeping a test set held-out) and (2) continued training on clinical notes from the i2b2/VA corpus [59], which also consists of expert-written notes and uses the same annotation scheme as the annotated contact notes.

As evaluation and continued training data, we use a data set of 200 contact notes annotated for coreference resolution in prior work [20], with train/dev/test sets of sizes 100/10/90. These notes were annotated using the same schema as prior work on medical notes [59] and examples include, *People*: names, pronouns, acronyms; *Problems*: housing insecurity, physical abuse, substance use; *Tests*: structured assessments carried out by CPS workers, records of school attendance; *Treatments*: housing services, sources of support, treatment plans. For all models, we report evaluation over the held-out test set over notes where the majority of clients are white (31 notes) or black (49 notes), using the same definitions as in Section 2.1. For models without continued training where we do not need to reserve data for training, we additionally report results over all black-majority (87) and white-majority (78) notes in the annotated data. We adopt the standard approach of averaging three coreference metrics: MUC, B^3, $CEAF_{\phi_4}$.

4.2 Results

Table 5 presents recall scores for all NER models. Recall scores for flairNLP models in particular are high. For all models, recall is significantly higher for names of white clients than for names of black clients. These findings are similar to the analysis of synthetic data conducted by Mishra et al. [36], who demonstrate that NER performance for several models is higher over white first names than black first names. Our work shows that this finding holds in a real and much a larger data set and over full names as well as first names. In examining names missed by the models, models typically fail to identify uncommon spellings, such as "Emilie" or names that

	Kirstain et al. [30]	Lee et al. [32] (i2b2/VA)	Lee et al. [32] (contact notes)
Black	58.82%	43.96%	-
White	57.11%	41.92%	-
B - W	1.71%	2.04%	-
Black	56.98%	43.58%	66.81%
White	57.12%	41.22%	68.24%
B - W	-0.14%	2.36%	-1.43%

Table 6: Average F1 for coreference models in annotated set of 165 contact notes (top) and in held-out test set of 80 notes (bottom). (i2b2/VA) and (contact notes) indicate the data set used for continued training. There is no statistically significant difference in performance over contact notes about majority-white and majority-black families for any models.

are also non-proper nouns, such as "Precious" or "Ruby".[3] Although performance differences are smaller for better-performing models, we examine a very large data set, so even a 1% score difference reflects 1000s of name-instances and is statistically significant.

In contrast, the test set of coreference annotations is quite small. Collecting more data is extremely difficult, as annotations are time-consuming and require domain expertise [20]. While we discuss performance in the absence of a larger data set, we caution that results are generally not statistically significant. From Table 6, for the two off-the-shelf-style models that are not trained on in-domain data, Kirstain et al. [30] and Lee et al. [32] (i2b2/VA), performance is poor, and manual examination of outputs suggests the models are not accurate enough to be usable. Training on in-domain data improves model performance (Lee et al. [32] (contact notes)), but reverses the direction of model bias: while Kirstain et al. [30] and Lee et al. [32] (i2b2/VA) perform slightly better or equal on notes about black-majority families, Lee et al. [32] (contact notes) performs better on notes about white-majority families.

Table 7 provides a finer-grained breakdown of model performances, presenting model recall of each entity type for the Lee et al. [32] (i2b2/VA) and Lee et al. [32] (contact notes) models, which are trained on data annotated with these entities. Both configurations exhibit stronger recall over *Person* and *Problem* entities for notes about black-majority families. Lee et al. [32] (contact notes) additionally achieves better recall of *Treatment* entities for notes about white-majority families. The origins of these disparities is difficult to untangle: they could reflect correlations between the types of *Problem*/*Treatment* entities mentioned in notes about families of different races and the types of entities that models recall better, e.g. from Section 2, notes about white families tend to discuss substances abuse more than basic needs. They also could reflect biases absorbed from external model training data or biases from ways notes are written. Nevertheless, regardless of origin, a coreference system that retrieves a smaller proportion of *Problem* entities in white-majority contact notes and a smaller proportion of *Treatment* entities for black-majority notes might lead to disproportionate focus on risks for black families and supports for white families.

[3]To preserve privacy, these examples are fabricated based on general patterns observed in the data, not real names of clients.

5 DISCUSSION

Despite numerous associated risks, in reality CPS agencies are actively seeking to deploy NLP tools. At the same time, NLP benchmark data sets are not sufficient for assessing model performance. While much literature has critiqued risk assessment systems [18, 53], little work has explored the impact of NLP in these tools, nor potential biases in assumedly more benign information extraction systems. We find significant racial bias in NER systems and possible biases in coreference systems (Section 4), but we do not identify racial bias as a core concern with integrating text features into an existing risk assessment system (Section 3). We additionally document evidence of different language use by race in contact notes (Section 2), which could have impacts on any models trained or deployed over this data.

Our results suggest that risks other than racial bias are more important considerations regarding incorporating text into risk assessment systems. Through an in-depth analysis of contact notes, Saxena et al. [53] discourage using this data in risk assessment models and emphasize ways CPS systems deprive families of agency, over-surveil parents, and problematize the data collection processes. Our finding that the models we train tend to assign higher risk scores to any referrals with pre-exisiting text data (Section 3) also suggest that incorporating text data could increase over-surveillance by encouraging repeated investigation of families that have already been involved in CPS. There are also risks that come from text as a data type. High-dimensional text data are liable to increase overfitting to "proxy" values used to train risk models. Also, text data are extremely difficult to anonymize. Because algorithmic systems are typically built by external researchers or contractors, text processing requires sharing it externally, which reduces the privacy of families and increases the potential harms associated with any data breaches. Finally, text data are liable to reduce the transparency and accountability of these systems. Community members have already objected to algorithmic risk assessment systems because they have no way to contest inputs or outputs and have little visibility into how tools are constructed or different features are weighted [3]. Text-processing systems are even more difficult to interpret than statistical classifiers: the volume of ongoing research on interpreting NLP models evidences that interpretability is an unsolved problem. Furthermore, although families likely have some knowledge of the values contained in structured data (e.g., if a caregiver has a criminal history), they typically have no knowledge of what contact notes contain nor the ability to contest them.

Information extraction systems focus on organizing information over direct decision-making, making them appear safer, and CPS agencies are more actively interested in deploying them; however, we do identify racial bias as a concern in this setting, and there are additional risks. Like with risk predictive models, deployment typically requires sharing data externally, and information extraction systems in particular tend to require much annotated data to achieve reasonable performance [64]. Further, even with a hypothetical perfect model, usefulness is entirely dependent on data quality. Saxena et al. [53] demonstrate that notes often contain descriptions of *perceived* risks that may not reflect reality, high turnover leads to inexperienced caseworkers, and caseworkers are incentivized to practice defensive decision-making over objective

	# of Entities				Lee et al. [32] (i2b2/VA)				Lee et al. [32] (contact notes)			
	Per.	Treat.	Test	Prob.	Per.	Treat.	Test	Prob.	Per.	Treat.	Test	Prob.
Black	424	269	195	290	50.81%	54.59%	59.39%	54.56%	-	-	-	-
White	366	238	163	248	48.04%	54.57%	59.43%	51.55%	-	-	-	-
B - W					2.77%	0.02%	-0.04%	3.01%	-	-	-	-
Black	59	42	30	40	51.14%	53.17%	55.61%	54.21%	82.21%	83.89%	90.22%	83.59%
White	39	29	24	29	46.11%	52.81%	54.75%	50.17%	81.32%	86.08%	89.19%	81.64%
B - W					5.03%	0.36%	0.86%	4.04%	0.89%	-2.19%	1.03%	1.95%

Table 7: Average recall for coreference models for entity types over 165-note annotated data (top) and held-out 80-note test set (bottom). For both models, recall of *Problems* is higher for notes about black-majority families than white-majority ones. For Lee et al. [32] (contact notes), recall of *Treatments* is higher for white majority families.

recording. Our data reveals similar evidence that caseworkers write notes explicitly to document and justify decisions (Appendix C). Word statistics (Section 2) also demonstrate that language can differ depending on the race of people involved. A CPS worker may be able to distinguish reliable from unreliable information more easily when manually reading notes than when viewing outputs of information extraction systems, which reduce relevant context. Finally, more investigation is needed to understand effect on clients. For example, DHS workers may find information extraction useful for preparing for court hearings, possibly even reducing administrative burden enough to decrease caseworker turnover. However, families and their advocates could be unfairly disadvantaged without access to similar tools and information. Even if algorithmic racial bias could be mitigated, models can still perpetuate harms by retrieving biased information from underlying data and increasing the power imbalance between CPS agencies and affected families.

We focus on risk assessment and information extraction models, because they have clear use cases and are of active interest to CPS agencies, but recent advances in interactive generative models, like the highly publicized ChatGPT have resulted in much interest around using these models as well. Our work does include investigations of pre-trained language models: we examine RoBERTa for risk assessment (Appendix A), the flairNLP NER models use XLM-R embeddings (Section 4.1, [11, 54]), and the neural coreference models use SpanBERT (Section 4.1, [29]). Generative models may have the potential to be useful for tasks like summarization, court preparation, or perhaps even information extraction. However, current models are prone to hallucinating fictional information [27], severely imitating their usability in high-stakes domains. Popular models are additionally closed-sourced and require interfacing through APIs rather than running models locally, which precludes evaluating them over protected data. Our results suggest that should these models be considered for deployment, evaluating them for racial bias is essential.

Conclusions. Studying the performance of NLP models in realistic high-stakes settings is extremely difficult due to concerns about data privacy and lack of transparency from many practitioners. Nevertheless, advances in NLP model performance over benchmark datasets and interactive demos like ChatGPT have generated intense interest in deploying these types of models, which mandates understanding how they actually perform. Our work aims to a provide a more realistic examination of algorithmic unfairness in NLP models than current research focused on synthetic benchmark data. We do uncover some evidence of algorithmic racial bias in this setting, specifically in NER models and in documenting different language use. Nevertheless, algorithmic unfairness is only one metric in this context, and more research is needed to uncover sociotechnical forces involved in deploying NLP models, including deeper engagement with stakeholders like affected families, caseworkers, and community advocates.

Limitations and Ethical Considerations. As our work does not contribute to the development of deployable systems, misuse potential of this work is low. However, there is significant risk of our results being taken out of context or misinterpreted. We emphasize that we study specific models in specific contexts, and our results cannot be assumed to generalize to other scenarios without appropriate evaluation. There are also inherent risks associated with working with such high-stakes data. We take numerous steps to mitigate these risks, including abiding by IRB and data sharing protocols, viewing anonymized versions of contact notes for data analyses as much as possible, and not providing any specific examples from notes in this paper. Given the interest in deploying NLP in high stakes settings, we believe the importance of providing visibility into model performance outweighs these risks.

ACKNOWLEDGMENTS

We gratefully thank the anonymous county Department of Human Services for providing feedback and data for this work and Lucille Njoo for providing feedback. This work was supported by seed funding from the Block Center for Technology and Society at Carnegie Mellon University, Pittsburgh, PA. A. Chouldechova has received research funding from the Allegheny County Department of Human services to support studies of racial bias in algorithm-assisted decision making. A.F. acknowledges support from a Google PhD fellowship. A. Coston acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745016 and from a Meta PhD fellowship. Y.T. acknowledges support from NSF CAREER Grant No. IIS2142739, the Alfred P. Sloan Foundation Fellowship, and NSF grants No. IIS2125201, IIS2203097, and IIS2040926.

REFERENCES

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[2] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81

[3] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300271

[4] Children's Bureau. 2017. Making and Screening Reports of Child Abuse and Neglect. https://www.childwelfare.gov/pubPDFs/repproc.pdf

[5] Children's Bureau. 2021. Child Maltreatment 2021. https://www.acf.hhs.gov/cb/report/child-maltreatment-2021

[6] Children's Bureau. 2021. State-Specific Foster Care Data 2021. https://www.acf.hhs.gov/cb/report/state-foster-care-data-2021

[7] Lindsay Cattell, Julie Bruch, et al. 2021. *Identifying Students At Risk Using Prior Performance Versus a Machine Learning Algorithm*. Technical Report. Mathematica Policy Research.

[8] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 162, 22 pages. https://doi.org/10.1145/3491102.3501831

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 134–148. http://proceedings.mlr.press/v81/chouldechova18a.html

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[12] Jude Mary Cénat, Sara-Emilie McIntee, Joana N. Mukunzi, and Pari-Gole Noorishad. 2021. Overrepresentation of Black children in the child welfare system: A systematic review to understand and better act. *Children and Youth Services Review* 120 (2021), 105714. https://doi.org/10.1016/j.childyouth.2020.105714

[13] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[14] Diane DePanfilis. 2003. Child protective services: A guide for caseworkers.

[15] Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. 2011. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review* 33, 9 (2011), 1630–1637.

[16] DHS. 2019. Developing predictive risk models to support child maltreatment hotline screening decisions. https://www.alleghenycountyanalytics.us/index.php/2019/05/01/developing-predictive-risk-models-support-child-maltreatment-hotline-screening-decisions/

[17] Jennifer L Eberhardt. 2020. *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. Penguin Books, USA.

[18] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group, USA. https://books.google.com/books?id=pn4pDwAAQBAJ

[19] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1905–1925. https://doi.org/10.18653/v1/2021.acl-long.149

[20] Nupoor Gandhi, Anjalie Field, and Emma Strubell. 2023. Mention Annotations Alone Enable Efficient Domain Adaptation for Coreference Resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, CA, 12 pages.

[21] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[22] Robert B Hill. 2004. Institutional racism in child welfare. *Race and Society* 7, 1 (2004), 17–33.

[23] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, New York City, USA, 57–60. https://aclanthology.org/N06-2015

[24] Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the Value of Information in Medical Notes. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2062–2072. https://doi.org/10.18653/v1/2020.findings-emnlp.187

[25] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

[26] Shaoxiong Ji, Shirui Pan, and Pekka Marttinen. 2021. Medical Code Assignment with Gated Convolution and Note-Code Interaction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1034–1043. https://doi.org/10.18653/v1/2021.findings-acl.89

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[28] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[29] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* 8 (2020), 64–77.

[30] Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference Resolution without Span Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 14–19. https://doi.org/10.18653/v1/2021.acl-short.3

[31] Wendy G Lane, David M Rubin, Ragin Monteith, and Cindy W Christian. 2002. Racial differences in the evaluation of pediatric fractures for physical abuse. *Jama* 288, 13 (2002), 1603–1609.

[32] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 687–692. https://doi.org/10.18653/v1/N18-2108

[33] Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered Mental Health Stigma in Masked Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2152–2170. https://aclanthology.org/2022.emnlp-main.139

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[35] Claire R McNellan, Daniel J Gibbs, Ann S Knobel, and Emily Putnam-Hornstein. 2022. The evidence base for risk assessment tools used in US child protection investigations: a systematic scoping review. *Child Abuse & Neglect* 134 (2022), 105887.

[36] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. Proceedings of the 2020 KBC Workshop on Bias in Automatic Knowledge Graph Construction.

[37] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16, 4 (2017), 372–403. https://doi.org/10.1093/pan/mpn018

[38] Michael Nash. 2017. Examination of Using Structured Decision Making and Predictive Analytics in Assessing Safety and Risk in Child Welfare. Los Angeles: County of Los Angeles Office of Child Protection.

[39] US Department of Health and Human Services. 2017. Child Maltreatment 2017.

[40] Allegheny County Department of Human Services. 2018. Request for Proposals: Unstructured Data Analytics Solutions. https://www.alleghenycounty.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=6442462821

[41] Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, USA, 838–844. https://doi.org/10.1109/ASRU46091.2019.9003958

[42] Brian E Perron, Bryan G Victor, Gregory Bushman, Andrew Moore, Joseph P Ryan, Alex Jiahong Lu, and Emily K Piellusch. 2019. Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child Abuse & Neglect* 98 (2019), 104180.

[43] Brian E Perron, Bryan G Victor, Joseph P Ryan, Emily K Piellusch, and Rebeccah L Sokol. 2022. A text-based approach to measuring opioid-related risk among families involved in the child welfare system. *Child Abuse & Neglect* 131 (2022), 105688.

[44] Emily Putnam-Hornstein, Rhema Vaithianathan, Jacquelyn McCroskey, and Daniel Webster. 2022. Los Angeles County risk stratification model: Methodology & implementation report. https://dcfs.lacounty.gov/wp-content/uploads/2022/08/Risk-Stratification-Methodology-Report_8.29.22.pdf

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language models are unsupervised multitask learners*. Technical Report. OpenAI.

[46] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[47] Dorothy Roberts. 2009. *Shattered bonds: The color of child welfare*. Civitas Books, USA.

[48] Dorothy E Roberts. 2019. Digitizing the carceral state. *Harvard Law Review* 132 (2019), 1695–1729.

[49] Wendy D Roth. 2016. The multiple dimensions of race. *Ethnic and Racial Studies* 39, 8 (2016), 1310–1338. https://doi.org/10.1080/01419870.2016.1140793

[50] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 8–14. https://doi.org/10.18653/v1/N18-2002

[51] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[52] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376229

[53] Devansh Saxena, Erina Seh-Young Moon, Aryan Chaurasia, Yixin Guan, and Shion Guha. 2023. Rethinking "Risk" in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 873, 19 pages. https://doi.org/10.1145/3544548.3581308

[54] Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. arXiv:2011.06993 [cs.CL]

[55] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1715–1725. https://doi.org/10.18653/v1/P16-1162

[56] Jessica Strolin-Goltzman. 2008. Should I stay or should I go? A comparison study of intention to leave among public child welfare systems with high and low turnover rates. *Child Welfare* 87, 4 (2008), 125–143.

[57] Jessica Strolin-Goltzman, Sharon Kollar, and Joanne Trinkle. 2010. Listening to the voices of children in foster care: Youths speak out about child welfare workforce turnover and selection. *Social work* 55, 1 (2010), 47–53.

[58] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[59] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18, 5 (2011), 552–556.

[60] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. *Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation*. Technical Report. Center for Social data Analytics. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V1-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL.pdf

[61] Bryan G Victor, Brian E Perron, Rebeccah L Sokol, Lisa Fedina, and Joseph P Ryan. 2021. Automated identification of domestic violence in written child welfare records: Leveraging text mining and machine learning to enhance social work research and evaluation. *Journal of the Society for Social Work and Research* 12, 4 (2021), 631–655.

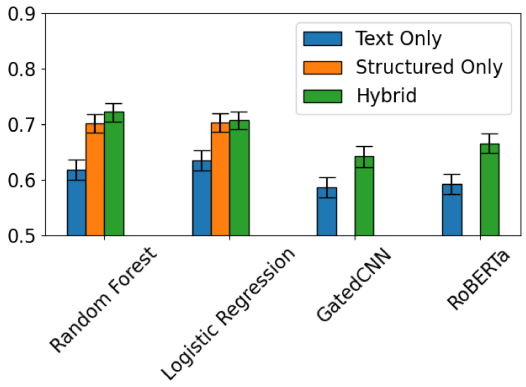
[62] Susan J Wells, Lani M Merritt, and Harold E Briggs. 2009. Bias, racism and evidence-based practice: The case for more focused development of the child welfare evidence base. *Children and Youth Services Review* 31, 11 (2009), 1160–1171.

[63] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018*. Technical Report. AI Now Institute at New York University, New York.

[64] Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference Resolution Model Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5241–5256. https://doi.org/10.18653/v1/2021.emnlp-main.425

[65] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18-2003

[66] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5218–5230. https://doi.org/10.18653/v1/2020.acl-main.466



Common Words		Highly Weighted Words	
Caseworker	Cas, ework, er	twin	tw, in
asked	ask, ed	lice	l, ice
Grandmother	Grand, mother	decision	dec, ision
Maternal	M, aternal	concern	con, cern
visit	vis, it	custody	c, ust, ody
concerns	con, cern, s	probation	pro, b, ation
Youth	Y, outh	unexcused	un, exc, used
Intake	Int, ake	paramour	param, our
Families	F, am, ilies	spank	sp, ank
Paternal	P, aternal	furniture	f, urn, iture
denied	den, ied		

Figure 5: Top: AUC scores of model variants. Metrics only include test data with text. 90% confidence intervals are computed using bootstrap sampling over the test set. Bottom: Words that RoBERTa tokenizer splits into to word pieces, includes the most common words in the training corpus that are split and the words assigned the highest or lowest weights by text-only logistic regression classifiers.

A ALTERNATIVE NLP MODELS FOR RISK ASSESSMENT

In the main paper, we focus on performance of a random forest model. Here, we additionally describe implementations and results for three other types of classification models: logistic regression, a convoluted neural network (CNN)-based neural classifier [26], and a RoBERTa-based neural classifier [34]. In general, we compare training models with only structured features (structured) to training with only text features (text) to training with both structured features and text (hybrid). For the GatedCNN and RoBERTa models, we do not train structured-only versions, as these models were specifically designed for text processing.

GatedCNN. This model is a state-of-the-art classifier developed for assigning codes to medical notes [26]. It uses a CNN-based architecture that involves injecting word embedding between layers and an LSTM-style gating mechanism. The original model additionally computed dot-product interactions between medical notes and codes, using word embeddings derived from code descriptions. In our hybrid model, we incorporated structured features by replacing the medical code representations with the 818-dimensional structured feature vectors. For the GatedCNN model, associated notes were truncated to the last (most recent) 3,000 tokens. The model embeddings were initialized with 100-dimensional word embeddings trained over the full data set of 3.1M contact notes (excluding the test set) using skip-gram Word2Vec with a context window of 5. We used the same kernel, filter sizes, and hidden layer sizes as the original model [26]. The model was trained with a learning rate of 1e-03 for up to 20 epochs, with early stopping if development set performance did not increase for 3 epochs. Hyper-parameters were selected based on development set performance after 5 epochs of training.

RoBERTa. We trained a RoBERTa-based classifier [34], where classification decisions were made using the final CLS representation. Popular variants of the high-performing transformer architecture (including RoBERTa) only support inputs up to 512 tokens [34]. Thus, we concatenated all associated notes and truncated them to the last (most recent) 512 tokens. In early experiments, we found that truncation outperformed alternative approaches to handling long inputs to a transformer, such as selecting input sentences using scoring functions or hierarchical models [41], which likely require larger training corpora. To incorporate structured features, we concatenated them to the CLS representation and passed the concatenated vector through a fully-connected linear layer, followed by a soft-max layer.

Prior to training the model, we conducted additional masked-language-model pretraining over the full data set of 3.1M contact notes (excluding test data) for 1 epoch, which prior work has shown improves performance on domain-specific data [21]. We fine-tuned the model for classification using a learning rate of 1e-05 and weight decay of 0.01 for up to 30 epochs, where training was stopped early if development set performance did not increase for 3 epochs. As above, hyper-parameters were selected based on development set performance after 5 epochs of training.

RoBERTa uses Byte-Pair Encoding (BPE) to enable handling large vocabulary, which divides out-of-vocabulary words into sub-pieces in order to derive components that are part of the vocabulary [34, 45, 55]. In Figure 5, we show some of the most common words in our corpora that were not included in the model vocabulary and were sub-divided by the tokenizer.

Random Forest and Logistic Regression. For the text models, we extracted TF-IDF-weighted bag-of-words features using a 10,000 word vocabulary. For the hybrid models, early experiments showed that concatenating the full 10,000 text features with the 818 structured features caused the model to ignore the structured features. Instead, we first trained a text model using a logistic regression classifier. We then took the 500 words with the highest learned coefficients and the 500 words with the lowest (most negative) coefficients and constructed TF-IDF features from this 1,000 word vocabulary. We then concatenated these features with the 818 structured features, constructing 1,818-dimensional feature vectors. All random forest classifiers used 500 trees and all logistic regression classifiers used L2 regularization.

Results. Figure 5 reports AUC scores for the different models. Across all model variants, the text models performed worse (random forest: 61.89, logistic regression: 63.50, GatedCNN: 58.66, RoBERTa: 59.19) than the structured (random forest: 70.19, logistic regression:

70.29) or hybrid (random forest: 72.21, logistic regression: 70.74, GatedCNN: 64.24, RoBERTa: 66.60) models. We also observed that the GatedCNN and RoBERTa models performed worse than the bag-of-words statistical classifiers. Based on the model performance results in Figure 1, in the main paper, we analyzed the performance of the random forest structured and hybrid models.

B ADDITIONAL MODEL METRICS

	All (14,417)		Black (6,841)		White (5,763)	
	Struct.	Hybrid	Struct.	Hybrid	Struct.	Hybrid
AUC	75.77	76.27*	74.70	75.69*	74.96	74.96
TPR	56.04	56.40*	56.81	57.46*	53.94*	53.46
FPR	19.59	19.53*	20.79	20.36*	19.69*	20.01
FNR	43.96	43.60*	43.19	42.54*	46.06*	46.54
Precision	32.47	32.67*	37.02	37.77*	28.18*	27.67
F1	41.12	41.38*	44.83	45.58*	37.02*	36.47
Accuracy	76.91	77.01*	75.25	75.71*	77.00*	76.66

Table 8: Metrics for risk prediction task with different variants of the random forest model over all test data. Where the difference between the hybrid and structured models is significant ($p < 0.05$) the better-performing value is starred.

	All (4,133)		Black (1,880)		White (1,894)	
	Struct.	Hybrid	Struct.	Hybrid	Struct.	Hybrid
AUC	69.83	71.88*	68.26	70.72*	70.09	71.77*
TPR	59.49	70.20*	59.45	71.76*	58.53	67.32*
FPR	31.78*	40.34	33.42*	43.55	30.70*	37.37
FNR	40.51	29.80*	40.55	28.24*	41.47	32.68*
Precision	31.37*	29.82	36.42*	34.68	26.56*	25.47
F1	41.07	41.85*	45.17	46.75*	36.54	36.95*
Accuracy	66.51*	61.72	64.84*	60.18	67.58*	63.38

Table 9: Metrics for risk prediction task with different variants of the random forest model over test data that contains text. Where the difference between the hybrid and structured models is significant ($p < 0.05$) the better-performing value is starred.

C DECISION DOCUMENTATION EXPERIMENTAL RESULTS

We investigated whether text data is liable to increase overfitting to human decisions by examining how predictive contact notes are of near-term decisions. Specifically, given a referral-child observation that was screened in, we trained a model to predict if the referral will be accepted for services, and we separately trained a model to predict if the child will be placed (removed from home) within 2 years of the referral. For each task, we trained structured and hybrid models, where we incorporated notes associated with the current referral into the hybrid model. Thus, while our primary hybrid model incorporated notes that precede screening decisions, these models incorporated notes typically generated during the investigation phase, after a screening decision has already been made. As most screened-in referrals have associated notes generated during investigation, in this data set, 28,340/28,769 training observations had notes and all 14,417 test observations had notes.

Table 10 reports the AUC scores of the random forest structured and hybrid models. The hybrid models showed strong improvements over the structured models. Table 10 also reports the top-weighted words by text logistic regression models over the same data. The model assigns the highest weights to "decis" (lemmatized decision), "hear," "servic" (lemmatized service), and "crisi" (lemmatized crisis). These results provide evidence that notes often contain documentation and justification of decisions.

Model	Service	Placement
Structured AUC	77.2	75.9
Hybrid AUC	86.1	80.5
Highest-weighted Words	crisi, servic, hear, decis, group	foster, placement, hear, author, physic

Table 10: *Top:* AUC for random forest models with and without text features from notes directly associated with referrals over two nearer-term decisions: whether or not the referral will be accepted for services and out-of-home placement. *Bottom:* words assigned the highest weights by a text logistic regression classifier. Notes often document and explicitly justify decisions.