Always Lurking: Understanding and Mitigating Bias in Online Human Trafficking Detection

Kyle Hundman
kyle.a.hundman@jpl.nasa.gov
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109 USA

Thamme Gowda *
tg@isi.edu
Information Sciences Institute
University of Southern California
Marina Del Rey, CA 90292, USA

Mayank Kejriwal
kejriwal@isi.edu
Information Sciences Institute
University of Southern California
Marina Del Rey, CA 90292, USA

Benedikt Boecking
boecking@cmu.edu
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

Web-based human trafficking activity has increased in recent years but it remains sparsely dispersed among escort advertisements and difficult to identify due to its often-latent nature. The use of intelligent systems to detect trafficking can thus have a direct impact on investigative resource allocation and decision-making, and, more broadly, help curb a widespread social problem. Trafficking detection involves assigning a normalized score to a set of escort advertisements crawled from the Web – a higher score indicates a greater risk of trafficking-related (involuntary) activities. In this paper, we define and study the problem of trafficking detection and present a trafficking detection pipeline architecture developed over three years of research within the DARPA Memex program. Drawing on multi-institutional data, systems, and experiences collected during this time, we also conduct post hoc bias analyses and present a bias mitigation plan. Our findings show that, while automatic trafficking detection is an important application of AI for social good, it also provides cautionary lessons for deploying predictive machine learning algorithms without appropriate de-biasing. This ultimately led to integration of an interpretable solution into a search system that contains over 100 million advertisements and is used by over 200 law enforcement agencies to investigate leads.

Introduction

Human trafficking has seen increasing media attention and government focus in recent years due to its pervasiveness and insidious nature (Austin and Farrell 2017). It is also characterized by a significant *Web presence*, with traffickers often advertising their victims on public platforms such as `backpage.com` (Szekely et al. 2015). Forums and review sites also contain discussions by 'clients' about (potentially trafficked) escorts, and other aspects of their experiences, such as the youth of, and services provided by, the escort.

The high prevalence of online sex advertisements (ads) and reviews, even on the Open Web, was a motivating factor in the creation of the DARPA Memex[1] program, under which this work was funded and conducted over a period of three years. Memex was designed to advance the state-of-the-art in building domain-specific search systems over massive Web corpora, especially in difficult domains like human trafficking. Various Memex-funded systems can be integrated to build *end-to-end* domain-specific search systems, starting from *domain modeling* and *discovery* (including crawling the Web for relevant pages), knowledge graph construction, machine learning and information retrieval (Szekely et al. 2015), (Kejriwal and Szekely 2017), (Krishnamurthy et al. 2016), (Shin et al. 2015).

An important *inferential* problem that needs to be addressed at scale in this pipeline is to detect potential trafficking activity by assigning a *risk score* to a set of advertisements, usually collected by an investigative expert like a law enforcement official. In the simplest case, the risk score is a binary flag, with 1 indicating that the ads in the set warrant further trafficking-related investigation. Intuitively, a set represents an informal version of a 'case study' that, for reasons grounded in real-world activities like tip-offs from contacts in the field, arrest records or exploratory search, has come to the attention of an official. While a single ad is often not useful by itself, intriguingly, when considered in aggregate, even a small set of ads in the case study can provide *subtle* clues indicating trafficking, rather than voluntary escort activities. For example, there may be evidence in one of the ads that an escort is *underage* or is advertising sex services that are *risky* and unusual relative to the domain. There may also be evidence of *movement* between cities, or in the case of brothels often fronting as Asian massage parlors, *ethnicity*-related clues.

The problem of trafficking detection, even by a human carefully analyzing the case study, is further compounded by the fact that ads in such case studies tend to be related only *latently*, and the relation itself can be subtle. In most cases, the assumption in identifying trafficking-related case studies is that escorts represented in the case study ads are being trafficked in a similar *context*, i.e. by a *single* individual or

*This work was done when the author was at the NASA Jet Propulsion Laboratory, Pasadena, CA, USA.

[1] `https://www.darpa.mil/program/memex`

an organization. Precisely identifying such contexts, and evidence backing the contexts, can be used to inform investigative decision making, alleviate the cognitive burden and significantly preserve the limited resources of both law enforcement and state district attorneys, who are often tasked with prosecuting trafficking-related cases. Given that the Memex program has already scraped many millions of sex advertisements on the Web for investigative purposes (Kejriwal and Szekely 2017), *automatic detection* methods, ideally with *interpretation*, can serve an important, socially beneficial function. At the same time, because the latest state-of-the-art methods in text classification use complex machine learning models like deep neural networks (Zhang, Zhao, and LeCun 2015a), with strong (and not very well understood) dependencies on the input data, it is important to understand the *biases* and limitations of such methods.

Systematically understanding the tradeoff described above, between building a system that can serve as an important example of *AI for Social Good*, and ensuring that the system is *fit* for use in a real world, requires a collaborative, socio-technological exchange. The Memex program provided such an exchange through a series of competitions and *quarterly progress reviews* (QPRs) when participants would regularly gather for myriad purposes such as system evaluation, strategic collaborations, and detailed meetings with potential real-world users of the system such as law enforcement. The QPRs led to a rich trove of insights around trafficking detection, both as a *technical* and an *investigative field* problem. In this paper, we use these insights to present the trafficking detection problem in depth for the broader research community, solutions explored and developed over the course of three years, corresponding lessons learned, eventual (and ongoing) integration into a comprehensive search system, and continuing impact. Specific contributions are listed below.

Contributions. The main contributions of this work are described as follows. We describe the trafficking detection problem and the motivations for devising automatic trafficking detection methods. We present an *architectural overview* of the approach that was first developed (before bias analysis and mitigation), followed by a *bias mitigation plan* that was put in place and continues to be implemented. We describe the lessons learned over three years of research from this exercise, and the changes that were made to the system in response to real-world users. The most important change was that the problem definition itself became finer-grained, allowing a degree of interpretability. Solutions to this finer-grained problem are already in the process of being integrated into a large-scale domain-specific search system that has had considerable impact in the last year on sex trafficking-related prosecutions.

In keeping with the scope of this work, we favor discussions of methodology over technical descriptions, except when necessary. Similarly, while we attempt to provide quantitative data when feasible, we also pay close attention to aspects of this work that cannot be easily controlled for or quantified, but still provide important lessons for deploying similarly complex systems.

Related Work

This work primarily draws on two emerging areas of research that are continuing to increase in significance. First, devising feasible and useful solutions for automatic trafficking detection is a good example of the general field of *AI for social good*, at the levels of both algorithmic development and engineering effort. Second, the bias analysis conducted in this paper directly reflects recent debate on the (often unintentional) biases that creep into AI systems. Rather than attempt comprehensive coverage of these research areas, we focus on work that is closely tied to human trafficking (and where applicable, similar *illicit* domains).

Intelligent Systems for Counter-Human Trafficking. Because of the alarming rise in online sex advertisement activity, a non-trivial portion of which may pertain to trafficking, building intelligent systems for assisting investigators and for enabling counter-human trafficking efforts has emerged as an important agenda. (Dubrawski et al. 2015) presented a host of data mining methods to support sex trafficking investigators in individual cases as well as lawmakers in understanding community-level statistics, amongst them the use of anomaly detection methods for community level statistics and an analysis of different text classifiers to detect sex trafficking-related activity. (Portnoff et al. 2017) present techniques to link related escort advertisements including a stylometry based classifier as well as an approach that exploits data leakage in the payment system of advertisements conducted via bitcoin. (Nagpal et al. 2015) investigate clustering approaches for escort advertisements and rely on blocking schemes to link large amounts of escort advertisement data. Features such as rare n-grams or rare images can be used to create blocks of data within which exact comparisons of advertisements are carried out to generate clusters. The final cluster resolution is then achieved by subsequently resolving the dataset across blocks.

More generally, there is a growing body of work for assisting investigative efforts in *illicit domains*, which includes not just human trafficking (HT), but also domains such as illegal weapons sales and securities fraud, both of which have also been investigated under the Memex program. Often, both text and multi-modal data are involved. (Mattmann et al. 2016) explored metadata of multimedia files (such as Exif Tags in images and videos) retrieved from escort ad websites to assist HT forensics. (Gowda, Hundman, and Mattmann 2017) applied a deep learning-based computer vision framework to assist the detection and classification of ads related to illegal and dangerous weapons. Other relevant examples include building search systems for helping investigators search for, and research, promising leads. The Domain-specific Insight Graph (DIG) system is a good example of the latter and is currently being used by over 200 law enforcement agencies to counter human trafficking (Szekely et al. 2015; Kejriwal and Szekely 2017).

Trust and Bias in AI. A series of recent studies have shown that even *standard* algorithms are not without bias. For example, (Caliskan, Bryson, and Narayanan 2017) show that standard learning algorithms trained on widespread text corpora learn stereotyped biases. (Sandvig et al. 2014) provide an overview of research that criticizes and reverse en-

gineers algorithms to understand consequences of their deployment and to discuss potential discrimination stemming from their use.

The presence of undesired bias in feature representations automatically learned from the training dataset has also been studied in the literature; for instance, (Zhao et al. 2017) analyzed the gender biases in machine learning models.

The studies cited above make it clear that detecting bias in intelligent systems is both important and non-trivial. Thus, in describing our trafficking detection approach, we also present a *bias mitigation plan* that emerged from months of effort, and key elements of which are already being integrated in an in-use counter-trafficking search system.

Problem Definition and Challenges

We assume a data collection process that yields a domain-specific collection C of webpages, almost all of which may be assumed to be either advertising sex (an *escort ad*) or reviewing the services of an escort (a *review ad*), collected via specially-tuned Web crawlers. In the most general case, each ad $c \in C$ is simply an HTML page, but it is convenient to assume that some preprocessing has been done (most importantly, *text scraping* and *information extraction*) and that c is itself a set of *key-value* pairs. More concretely, the *schema* of C, which is a union over all keys in the collection, includes such attributes as phone numbers, locations, dates, and cleaned ad text, to name a few, and were used both for modeling and bias mitigation as well as in search systems that were eventually exposed to law enforcement.

Given a small set $C' \subset C$ of ads from this collection, we define the *automatic trafficking detection problem* as discovering an *assignment* function $f : C' \rightarrow [0, 1]$, where we denote $f(C')$ as the trafficking *risk score* of the *case study* C'. There are reasons for using this terminology. First, the manner in which C' is isolated from C is qualitatively similar to how a case study file is constructed. Ads in C' are not sampled at random, although the precise reason for grouping the ads in C' together may not be known in advance. Second, we note that any outputs by f cannot be validated (even by a human reading the ad) except through a real-world investigation. In this sense, the problem is different from ordinary text or cluster classification problems, and more similar to Information Retrieval problems that seek to assess 'relevance.' For this reason, f is only said to assign a *risk* score.

Given the success and pervasive use of commercial search platforms like Google, why is an automated solution to trafficking detection even necessary? It is not unreasonable to assume a hypothetical *lead generation* workflow whereby an investigator, after a period of browsing or exploration on high-activity portals like the *adult* section of `backpage.com`, could make informed investigative decisions in an online fashion on whether a generated lead warrants a high-priority investigation.

There are several problems that arise with such exploratory lead generation. First, although a sizable portion of trafficking activity takes place on the public Web, it is *sparsely* interspersed among escort ads and reviews. In general, the connection between human trafficking and online sex advertising is still not well-understood (Latonero 2011) but preliminary studies conducted under Memex provided some evidence that trafficking-related lead generation is a non-trivial problem. For *investigative* purposes, non-trafficking leads do not take priority over those that may exhibit some form of trafficking. The second problem, which aggravates the effects of sparsity, is prohibitive *scale*[2]. Although potential case study clusters can be isolated from this collection through a conservative process of rule- or keyword-based application, the many thousands of clusters that emerge as a result are far too many for a human to classify or study. Even more importantly, the domain is notoriously *dynamic* and *adversarial*, since potentially indicative signals (such as tattoos on specific body parts, and language use in advertisements) are constantly evolving, and are not well-understood sociologically.

Beyond addressing general problems of scale and sparsity (particularly, class skew), timely and automatic detection of trafficking-related activity has two other domain-specific motivations, both of which have the potential for widespread social impact. First, identifying cases at high risk can lead to early investigations, which may stop others from being trafficked. In this case, the motivation has a preventive aspect to it. Second, the ads, if flagged in time and retained *offline*, have proven to be potent exhibits in actual trafficking-related criminal cases that were recently prosecuted in the US in New York and California (Greenemeier 2016). By supporting prosecutions and evidence gathering, automatic detection systems have real potential for accelerating social justice and for raising the barrier to entry for traffickers that are using the Web to advertise illicit activity.

Approach

Human trafficking detection, from the algorithmic perspective, is a binary classification task on retrieved escort ad and review data. Several tasks involved in this pipeline were performed as a collaborative effort among multiple research groups funded by the Memex program. Figure 1 illustrates the stages involved in an end-to-end pipeline designed for training and evaluating automatic trafficking detection:

Step 1: Crawling and Data Collection: Web crawlers designed under Memex were tuned to search for, and scrape, sex-related advertisements, particularly from online marketplaces with significant *escort*- and *escort review*-related activity. Data analysis showed the majority of ads and reviews to be respectively coming from the *adult* sections of *backpage.com* and *craigslist.com*, and *eroticreview.com*. Data collected from crawlers was made available by indexing to the *common data repository (CDR)* system.

Step 2: Extraction: In this stage, the crawled webpages and multimedia files were processed by *Information Extraction* algorithms (Chang et al. 2006) that extracted domain-specific attributes such as phone numbers, dates, locations of services, image links, and plain text decriptions (Szekely et al. 2015; Mattmann and Zitting 2011). All extractions were indexed in the CDR along with the raw data.

Step 3: Labeling: Labeling of escort data by trafficking experts is a costly and time-consuming task, involving sig-

[2]The current Memex repository indexes over 100M documents.

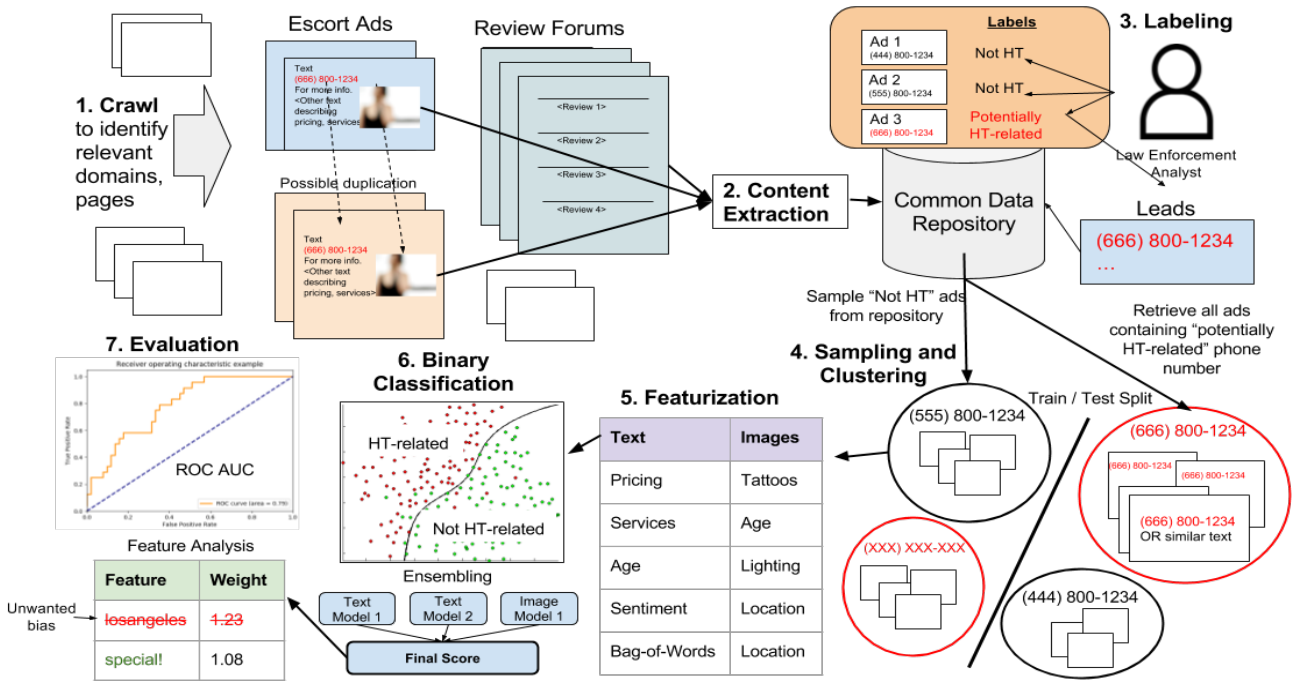

Figure 1: The end-to-end trafficking detection pipeline designed for multi-modal Web-scale corpora containing sex advertisements and reviews. This pipeline includes data collection stages such as *crawling* and *extraction*, data preparation stages such as *sampling* and *clustering* for acquiring labeled data for training and evaluation, and finally, the standard machine learning stages of *featurization, training, ensembling, evaluation* and *post hoc feature analysis*.

nificance cognitive burden. Several law enforcement partners across the U.S. shared information on historical and recent cases of (suspected) human trafficking that allowed us to retrieve relevant positive advertisements using unique identifiers such as phone numbers. Obtaining a reasonable amount of negatively labeled (i.e. not of investigative interest) ad data proved to be more difficult. Athough the Memex program contracted experts to acquire a training set of negatively labeled ads, the data was limited.

Step 4: Sampling and Clustering: *Sampling* attempts to address the scarcity of negatively-labeled ads by relying on the assumption that the *relative* prevalence of positive ads is small. The key idea is to obtain noisy negative labels by selecting groups of related advertisements *at random* from the entire corpus, and labeling them as negative. However, since this process of obtaining negative labels is substantially different from the process of obtaining positive labels, the approach resulted in biased training subsets. This led to the development of a more rigorous negative sampling approach detailed further in the *Bias Mitigation Plan*.

In the real world, as discussed in earlier section, ads are not independently created and there exist *underlying clusters of ads* (denoted earlier as *case studies*) that are generated by *distinct entities* (such as escorts) but tend to contain *similar* text and media by way of a *latent relationship*. Clusters acquired for the purposes of training and evaluating the system need to satisfy downstream independent and identically distributed (i.i.d.) data modeling assumptions. To achieve this, we designed a set of multi-modal similarity functions over extracted attributes such as descriptive text, phone numbers, and images. Intuitively, these similarities can be used to cluster similar ads together and dissimilar items apart.

Correlation clustering is one such clustering approach that was found to be very useful in practice. There exist fast correlation clustering approaches with provable guarantees that scale linearly with the data, such as KWIKCLUSTER (Ailon, Charikar, and Newman 2008), which obtains a 3-approximation ratio. Parallel correlation clustering approaches (Pan et al. 2015) based on KWIKCLUSTER have proven to be both efficient and effective in practice. Building on prior research findings (Elsner and Schudy 2009), we tuned and implemented a combination of KWIKCLUSTER, consensus clustering and local heuristics in our approach.

Steps 5 and 6: Featurization and Binary Classification Supervised text classification generally involves the mapping words and phrases into numerical feature vectors, using both classic bag-of-words approaches, and in more recent work, *word embedding* algorithms like word2vec and fastText (Bojanowski et al. 2016). A variety of approaches for both 'featurization' and model selection were explored throughout the project, but bias issues made systematic comparisons and validation of models a difficult task. We note that 'black box' machine learning models such as deep neural networks provide no assurance that the learned features are not fitting to undesired biases in the labeled dataset (in addition to offering limited performance benefits for smaller training sets like ours (Zhang, Zhao, and LeCun 2015b)). This led us to favor more transparent approaches, allow-

ing for greater visibility into potential sources of bias in our system. Models that generally performed well, and offered (at least marginal) interpretability were linear-kernel SVMs (which are capable of suggesting the relative importance of features per (Guyon et al. 2002)), ensemble models such as random forest classifiers, and penalized logistic regression models trained on bag-of-word vectors. These modeling experiments informed both the *Bias Mitigation Plan*, and the subsequent integration of the approach into end-user tools.

Step 7: Evaluation: The approach was evaluated on an independent set of ad clusters that was mutually exclusive from the training set, but was gathered using a similar protocol. We used the area under the Receiver Operating Characteristic (ROC) curve as the performance metric. Posthoc evaluation studies involved detailed analysis of feature importance, and the origin of important features.

Bias Mitigation Plan

.

Mitigating bias in intelligent systems is a complicated issue, as the sources of bias are not easy to isolate. Bias may arise when certain algorithmic assumptions are violated, or when the training data is biased, either because the *sample size* (compared to the population) is too small, or because the *labeled data acquisition process* is biased.

For example, while positive labels for groups of escort ads may come from a small number of law enforcement contacts that only provide cases for specific regions, noisy negative labels sampled at random will follow the true location distribution in the corpus more closely. Classifiers trained on such data may learn to differentiate classes using locations specific to law enforcement contacts, rather than learn actual and meaningful signals indicating human trafficking.

Diagnosing and Evaluating Bias

The first step in addressing bias is identifying potential sources of bias that often become apparent through thorough data exploration, domain understanding, or the modeling process. Simple statistical significance tests or the computation of correlation coefficients can assist in validating suspected biases. Also, expert knowledge can also help us determine obvious information in the data that should not be indicative of the output label.

For example, we can test the null hypothesis that the distribution of a potentially-biased feature like *Web domain* is independent with respect to label class (i.e. not biased via sampling, labeling, clustering, etc.) with a Pearson's chi-squared test (Pearson 1900). Using $\alpha = 0.05$ as the threshold for rejecting our null hypothesis, the below table shows actual counts of ads for labeled data available for modeling during the program:

Table 2: Ads by Domain Group and Label Class

	Positive	Negative	**Total**
backpage.com	165,686	125,467	**291,153**
other	155,271	154,627	**309,898**
Total	**320,957**	**280,094**	**601,051**

Calculation of the test statistic and subsequent p-value results in $p < 0.00001$ and our assumption of independence is violated at the chosen α. To correct for this bias we can sample additional (presumably) negative ads from the CDR with the aim of aligning the two distributions. Although these are actual numbers, this is a simplified example intended for demonstration – a more appropriate correction would involve more granular groupings of domains and other biased features would need to be considered during the sampling of additional negatives.

Under a more traditional multiple hypothesis testing scenario this would necessitate correcting for the issue of multiple comparisons using, for instance, a Bonferroni correction (Curran-Everett 2000) to set a more rigorous threshold for finding statistically significant relationships. Our motives differ, however, in that we are not using hypothesis tests to identify variables that may help explain a resulting dependent variable, but rather that we are using domain knowledge to determine information that should not strongly correlate with the class labels.

In addition to testing for bias, this approach can be used to evaluate the effectiveness of mitigation efforts described below. It should also be used when constructing folds for cross validation to ensure independence assumptions hold during learning. If mitigation is successful independence assumptions should be satisfied according to the above criteria.

Mitigation of Bias

Information Removal and Conditioned Sampling. Upon identifying biased information in training data, we want to ensure that we correct for the bias before or during model training. One way to achieve this is to try to remove any information that relates to a biased feature either before or after vectorization of the data points. For example, in the context of *location bias*, one could try to remove any tokens in the ad text that refer to locations. Naturally, this requires information extractors with high recall and ideally high precision, which may not always be available.

Another approach towards alleviating this issue is to add or remove data points from the training data with the goal of ensuring that the distribution of the biased feature is similar across the positive and negative data points. Adding additional human-labeled data, especially in such a way as to align distributions is difficult, expensive, and may introduce additional unwanted sampling biases. In this context, a more appropriate method to control for bias is to extract additional negative ads from the CDR such that our sample contains similar distributions of biased features across classes, discouraging classifiers from learning them. As mentioned in *Step 4* of the *Approach* section, the process of sampling of negative ads and clusters from the CDR is conducive to the conditioning, especially when extractions for biased features are readily available. This strategy introduces the step of determining how to measure similarity between distributions, which may be done via distribution divergence measures such as the Rényi divergence (Van Erven and Harremos 2014) or by doing a two-sample Kolmogorov–Smirnov test (Daniel and others 1978).

Clustering. Ignoring *inherent* dependencies between ads

Table 1: Bias Mitigation Overview

Bias Type	Overview	Diagnosis and Evaluation	Mitigation Steps
Labeling	Labeled data biased toward certain locations, Web domains, and positive class due to nature of labeling	1. Use correlation and hypothesis tests to evaluate independence of potentially-biased features relative to class	1. Sample additional negative examples conditioned on biases found in positive class data 2. Remove biased features
Domain-Specific	Cluster sizes vary and escort ad content is often duplicated across accounts and domains	1. Single-feature modeling (e.g. cluster size) 2. Test for duplicate data across classes	1. Sample negative clusters to resemble positive cluster sizes 2. Use multi-objective clustering to prevent duplicated content from appearing across clusters
Estimation	Training data is limited and careless partitioning can cause overfitting to samples and invalid results	1. Classes in cross-validation folds should show homogeneity between distributions of biased features	1. Condition cross-validation folds to have matching distributions of unwanted features 2. Maintain model interpretability

introduces numerous problems when the aim is to build text classifiers that assume i.i.d. data. For example, the implicit weighting of features that occurs in a dataset with many near duplicates will not reflect the importance of the features in relation to the output label one is trying to predict. Rather, a model will be encouraged to memorize specific patterns in training data duplicated across classes. This is especially problematic under widely varying cluster sizes, which would suggest *cluster-level classification* as a sensible scheme.

Recovering the true underlying clusters presents unique challenges. The same person or group may produce ads for multiple individuals. Additionally, ad text and images are sometimes copied across personas and phone numbers are often intentionally obfuscated. As described in *Step 4* of the *Approach*, we used a correlation clustering method and some local heuristics to achieve subjectively good clustering results. To estimate how well the clustering helped in creating a clean training and test split as well as clean cross-validation folds one can estimate the out-of-cluster loss of a model under dependency within latent clusters, i.e. a way to recover the loss in the independent setting, such as shown in (Barnes and Dubrawski 2017).

Indicator Mining and Integration. In Table 1, interpretability of the model is a key mitigation step and involves a social aspect since the utility of interpretability is to the *users* of the system. Detailed conversations with law enforcement officials revealed a strong desire that the systems produce finer-grained 'clues' suggestive of potential, context-dependent trafficking detection, rather than a single score. These clues, called *indicators*, are highly specific and are designed to detect such high-level features as *escort movement*, advertisement of *risky* sex services, and presence of *multiple girls* within a single advertisement and several others. Indicators are defined to be features that are (believed to be) relevant to inferring accurate trafficking risk scores. In the last phase of the program, we significantly extended the principles of the approach with expert-elicited rules and unsupervised text embeddings to *supplement* ads with indicators. At the time of writing, these indicators are actively being integrated into the DIG search system, currently in use by more than 200 U.S. law enforcement agencies.

Impact and Conclusion

The majority of Memex trafficking detection systems are being permanently transitioned to the office of the District Attorney of New York, and generic 'non-trafficking' versions have been released as open-source software in the DARPA Memex catalog[3]. In the last year, DIG, along with other trafficking detection tools from Memex, has led to at least three trafficking prosecutions, including a recently concluded case in San Francisco where a man was sentenced to 97 years to life for human trafficking[4]. More than 25 victims were rescued, and the DA's office in San Francisco publicly acknowledged the Memex tools in making this possible.

This paper presented and defined an important problem called trafficking detection, which has much potential to be aided by recent advances in intelligent systems. We presented a general approach to the problem developed and evaluated over years of research under the DARPA Memex program, and a mitigation plan for addressing biases in the approach. Given Memex's sustained impact, we hope to continue improving our trafficking detection systems.

Acknowledgements. This effort was supported in part by JPL, managed by the California Institute of Technology on behalf of NASA, and additionally in part by the DARPA Memex/XDATA/D3M programs and NSF award numbers ICER-1639753, PLR-1348450 and PLR-144562 funded a portion of the work. The authors particularly thank Dr. Chris Mattmann from JPL, Dr. Pedro Szekely from ISI, and Dr. Artur Dubrawski from Carnegie Mellon University for their support and all of the Memex collaborators for their contributions.

[3] https://opencatalog.darpa.mil/MEMEX.html

[4] http://www.sfgate.com/crime/article/Man-sentenced-to-97-years-in-human-\trafficking-7294727.php

References

Ailon, N.; Charikar, M.; and Newman, A. 2008. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)* 55(5):23.

Austin, R., and Farrell, A. 2017. Human trafficking and the media in the united states.

Barnes, M., and Dubrawski, A. 2017. The binomial block bootstrap estimator for evaluating loss on dependent clusters. In *In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Chang, C.-H.; Kayed, M.; Girgis, M. R.; and Shaalan, K. F. 2006. A survey of web information extraction systems. *IEEE transactions on knowledge and data engineering* 18(10):1411–1428.

Curran-Everett, D. 2000. Multiple comparisons: philosophies and illustrations. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 279(1):R1–R8.

Daniel, W. W., et al. 1978. *Applied nonparametric statistics*. Houghton Mifflin.

Dubrawski, A.; Miller, K.; Barnes, M.; Boecking, B.; and Kennedy, E. 2015. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking* 1(1):65–85.

Elsner, M., and Schudy, W. 2009. Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, 19–27. Association for Computational Linguistics.

Gowda, T.; Hundman, K.; and Mattmann, C. A. 2017. An approach for automatic and large scale image forensics. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*, 16–20. ACM.

Greenemeier, L. 2016. Human traffickers caught on hidden internet.

Guyon, I.; Weston, J.; Barnhill, S.; and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46(1):389–422.

Kejriwal, M., and Szekely, P. 2017. Knowledge graphs for social good: An entity-centric search engine for the human trafficking domain. *IEEE Transactions on Big Data*.

Krishnamurthy, Y.; Pham, K.; Santos, A.; and Freire, J. 2016. Interactive exploration for domain discovery on the web. *Proc. of KDD IDEA*.

Latonero, M. 2011. Human trafficking online: The role of social networking sites and online classifieds.

Mattmann, C., and Zitting, J. 2011. *Tika in action*. Manning Publications Co.

Mattmann, C. A.; Yang, G. H.; Manjunatha, H.; Gowda, T.; Zhou, A. J.; Luo, J.; and McGibbney, L. J. a. 2016. Multimedia metadata-based forensics in human trafficking web data. *Second Workshop on Search and Exploration of X-Rated Information (SEXI'16), Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* 10.

Nagpal, C.; Miller, K.; Boecking, B.; and Dubrawski, A. 2015. An entity resolution approach to isolate instances of human trafficking online. *arXiv preprint arXiv:1509.06659*.

Pan, X.; Papailiopoulos, D.; Oymak, S.; Recht, B.; Ramchandran, K.; and Jordan, M. I. 2015. Parallel correlation clustering on big graphs. In *Advances in Neural Information Processing Systems*, 82–90.

Pearson, K. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50(302):157–175.

Portnoff, R. S.; Huang, D. Y.; Doerfler, P.; Afroz, S.; and McCoy, D. 2017. Backpage and bitcoin: Uncovering human traffickers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1595–1604. ACM.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*.

Shin, J.; Wu, S.; Wang, F.; De Sa, C.; Zhang, C.; and Ré, C. 2015. Incremental knowledge base construction using deepdive. *Proceedings of the VLDB Endowment* 8(11):1310–1321.

Szekely, P.; Knoblock, C. A.; Slepicka, J.; Philpot, A.; Singh, A.; Yin, C.; Kapoor, D.; Natarajan, P.; Marcu, D.; Knight, K.; Stallard, D.; Karunamoorthy, S. S.; Bojanapalli, R.; Minton, S.; Amanatullah, B.; Hughes, T.; Tamayo, M.; Flynt, D.; Artiss, R.; Chang, S.-F.; Chen, T.; Hiebel, G.; and Ferreira, L. 2015. Building and using a knowledge graph to combat human trafficking. In *Proceedings of the 14th International Semantic Web Conference (ISWC 2015)*.

Van Erven, T., and Harremos, P. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory* 60(7):3797–3820.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015a. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, 649–657.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015b. Character-level convolutional networks for text classification. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 649–657.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.