Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Detecting Human Trafficking: Automated Classification of Online Customer Reviews of Massage Businesses

Ruoting Li, Margaret Tobey, Maria E. Mayorga, Sherrie Caltagirone, Osman Y. Özaltın

To cite this article:
Ruoting Li, Margaret Tobey, Maria E. Mayorga, Sherrie Caltagirone, Osman Y. Özaltın (2023) Detecting Human Trafficking: Automated Classification of Online Customer Reviews of Massage Businesses. Manufacturing & Service Operations Management 25(3):1051-1065. https://doi.org/10.1287/msom.2023.1196

https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 25, No. 3, May–June 2023, pp. 1051–1065
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Detecting Human Trafficking: Automated Classification of Online Customer Reviews of Massage Businesses

Ruoting Li,[a] Margaret Tobey,[b] Maria E. Mayorga,[a] Sherrie Caltagirone,[c] Osman Y. Özaltın[a,*]

[a] Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695; [b] Operations Research Graduate Program, North Carolina State University, Raleigh, North Carolina 27695; [c] Global Emancipation Network, Clermont, Florida 34715

*Corresponding author

Contact: rli25@ncsu.edu, https://orcid.org/0000-0002-9138-4606 (RL); mgtobey@ncsu.edu, https://orcid.org/0000-0001-7539-767X (MT); memayorg@ncsu.edu, https://orcid.org/0000-0002-6399-2153 (MEM); sherrie@globalemancipation.ngo, https://orcid.org/0000-0002-5451-650X (SC); oyozalti@ncsu.edu, https://orcid.org/0000-0002-0093-5645 (OYÖ)

Received: November 23, 2021
Revised: September 3, 2022; December 13, 2022
Accepted: January 9, 2023
Published Online in Articles in Advance: Febrruary 22, 2023

https://doi.org/10.1287/msom.2023.1196

Abstract. ***Problem definition*:** Approximately 11,000 alleged illicit massage businesses (IMBs) exist across the United States hidden in plain sight among legitimate businesses. These illicit businesses frequently exploit workers, many of whom are victims of human trafficking, forced or coerced to provide commercial sex. ***Academic/practical relevance:*** Although IMB review boards like Rubmaps.ch can provide first-hand information to identify IMBs, these sites are likely to be closed by law enforcement. Open websites like Yelp.com provide more accessible and detailed information about a larger set of massage businesses. Reviews from these sites can be screened for risk factors of trafficking. ***Methodology*:** We develop a natural language processing approach to detect online customer reviews that indicate a massage business is likely engaged in human trafficking. We label data sets of Yelp reviews using knowledge of known IMBs. We develop a lexicon of key words/phrases related to human trafficking and commercial sex acts. We then build two classification models based on this lexicon. We also train two classification models using embeddings from the bidirectional encoder representations from transformers (BERT) model and the Doc2Vec model. ***Results:*** We evaluate the performance of these classification models and various ensemble models. The lexicon-based models achieve high precision, whereas the embedding-based models have relatively high recall. The ensemble models provide a compromise and achieve the best performance on the out-of-sample test. Our results verify the usefulness of ensemble methods for building robust models to detect risk factors of human trafficking in reviews on open websites like Yelp. ***Managerial implications*:** The proposed models can save countless hours in IMB investigations by automatically sorting through large quantities of data to flag potential illicit activity, eliminating the need for manual screening of these reviews by law enforcement and other stakeholders.

Funding: This work was supported by the National Science Foundation [Grant 1936331].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2023.1196.

Keywords: human trafficking • massage businesses • online customer reviews • Natural Language Processing • ensemble learning

1. Introduction and Motivation

Human trafficking is a form of modern-day slavery in which traffickers use force, fraud, or coercion to exploit victims for commercial sex or labor against their will (U.S. Department of State 2021). Identifying victims of trafficking and prosecuting the offenders is difficult. In most cases, there is a lack of evidence other than witness testimonies, which are difficult to obtain. Although the U.S. Department of State (2021) reported the identification of more than 100,000 trafficking victims globally in 2020, this resulted in barely 5,000 convictions.

We focus on trafficking in the massage industry. Illicit massage businesses (IMBs) commit a unique form of hybrid sex and labor trafficking. Current licensing and regulation of massage businesses is an "easily exploitable patchwork of state and local laws and ordinances" (Polaris Project 2019a), allowing approximately 11,000 alleged IMBs to thrive across the United States (Heyrick Research 2021), hidden among legitimate businesses. Federation of State Massage Therapy Boards (2017, p. 5) explains that: "Massage therapy is a profession that is commonly associated with human trafficking. […] This is in part a perception problem, but it is largely based on reality due to the fact that illicit businesses use massage therapy as a front for their illegal activity."

IMBs widely use the Internet to advertise in the massage or therapeutic sections of classified sites and to lure victims by deceptive job advertisements. IMB-specific review websites like Rubmaps.ch and MPReviews.com

also provide a platform for sex buyers to rate their experience and share details about the commercial sex acts purchased. These online activities can be screened by the counter-trafficking community. However, similar websites such as Backpage.com was seized by federal authorities (U.S. Department of Justice 2018). Hence, diversifying the data sources by considering open websites like Yelp.com is important. Although most massage businesses on Yelp are legitimate, IMBs can also be on Yelp as they use mainstream websites to appear legitimate in online search results (Polaris Project 2019a), and worried users can create Yelp business pages to report suspected activities. Customer reviews on Yelp can provide a unique perspective about a massage business. A survey of sex buyers in the United States found that active sex buyers are more likely to believe that people in prostitution "choose it as a profession" and "enjoy the act of prostitution" (Demand Abolition 2018), which may make them ignore or fail to recognize signs of trafficking. Yelp reviewers, however, can be concerned customers complaining of suspicious activities, hence providing more information about risk factors for trafficking in a massage business (e.g., workers living in the business). Example Yelp reviews are presented in Table A1 in the online appendix.

We define an *illicit review* as any customer review that explicitly states or hints at activities related to commercial sex or other risk factors of human trafficking at a business. It is important to note that commercial sex does not equate to human trafficking unless it is induced by force, fraud, or coercion (U.S. Department of State 2021). However, evidence shows that a nonnegligible proportion of massage business workers who are engaged in commercial sex are victims of trafficking. One study that interviewed 116 massage business employees in New York City and Los Angeles County who provided commercial sex found that 17% of these employees were forced or coerced to do so (Chin et al. 2019). In a counter-trafficking initiative that helped more than 1,200 women who had been arrested at massage businesses, one of five women said they had been trafficked or had experienced coercion (Yakowicz 2021). These numbers are likely drastically under-reported as trafficking victims usually do not self-report their status for fear of retribution by their traffickers or distrust of authorities (U.S. Department of Justice 2017).

Identifying signs of trafficking requires domain expertise. Manual screening is extremely time and resource intensive due to the high volume of online activity. The problem is made more difficult by the dearth of ground truth. Methods like text mining, natural language processing, and machine learning can enable better informed counter-trafficking measures. We aim to fill the research gap for automatically detecting illicit reviews on open websites that might potentially be associated with human trafficking in IMBs. We create labeled data sets of Yelp reviews that are crucial for learning such tasks. We analyze reviews via two approaches based on lexicon terms and embeddings. We use data augmentation and ensemble learning techniques to build classification models. Law enforcement can use these models to build human trafficking cases and supply evidence to justify warrants against suspected IMBs. Furthermore, victim service organizations can use our models to identify risky massage businesses and reach out to vulnerable people working in such places.

2. Literature Review

We first review previous efforts that use text analysis to combat sex trafficking. Then we discuss the studies that use text analysis to uncover IMBs and where our work fits in with these efforts.

2.1. Text Analysis to Combat Sex Trafficking

The analysis of sex trafficking advertisements from websites like Craigslist and Backpage can be traced back to Kennedy (2012) and Wang et al. (2012). Under guidance from law enforcement, Kennedy (2012) identified keywords and other features in a set of Backpage advertisements that may indicate underage victims or shared management of victims in a larger network, both signs of sex trafficking. Wang et al. (2012) created TrafficBot, a tool that integrated classified advertisements for escort and massage services with reviews from online bulletin boards.

Recent efforts proposed methods to identify sex trafficking from the text and other extracted features of advertisements such as locations and phone numbers. These studies linked advertisements to uncover sex trafficking networks (Keskin et al. 2021, Ramchandani et al. 2021), predicted whether advertisements involve trafficking (Wang et al. 2012, 2020; Alvari et al. 2017; Tong et al. 2017; Esfahani et al. 2019; Zhu et al. 2019a), or both (Dubrawski et al. 2015, Nagpal et al. 2017). Most of these works created text features from sex trafficking keywords provided by law enforcement (Dubrawski et al. 2015, Tong et al. 2017), provided by the counter-human trafficking nonprofit organization Global Emancipation Network (Wang et al. 2020) or collected from anecdotal sources (Alvari et al. 2017). However, these keywords do not carry the same meaning in reviews of massage businesses on open websites. For example, "fresh," which may indicate a minor victim in a classified ad, is frequently used to describe legitimate spa treatments in Yelp reviews. Such difference in language suggests the need for a specific lexicon to detect illicit massage business reviews on open websites.

A group of studies in the literature extracted features through natural language processing. Wang et al. (2012) and Dubrawski et al. (2015) analyzed the classified advertisements from multiple sites, including Backpage,

using the bag-of-words method. Alvari et al. (2017) and Zhu et al. (2019a) used the term frequency-inverse document frequency (TF-IDF) method to extract feature vectors. TF-IDF identifies the relative importance of a word in a document that is part of a larger collection (Wu et al. 2008). Zhu et al. (2019a) identified a list of keywords related to sex trafficking by applying feature selection on the TF-IDF vectors obtained from advertisements.

The keyword search, bag-of-words, and TF-IDF methods are based on word counts. In contrast, embedding methods can capture syntactic and semantic relationships of words (Li and Yang 2018). Word2Vec (Mikolov et al. 2013a) is one of the most popular methods to generate word embeddings. Skip-gram and continuous bag of words (CBOW) are two model architectures of Word2Vec. Wang et al. (2020) and Tong et al. (2017) trained skip-gram neural network models (Mikolov et al. 2013b) on advertisement text. Ramchandani et al. (2021) used the CBOW model (Mikolov et al. 2013a) to obtain word embeddings on advertisement text. Doc2Vec (Le and Mikolov 2014) is an extension of Word2Vec that extracts embeddings for documents instead of the words. Simonson (2021) used a Doc2Vec model to extract embeddings for social media posts and trained a semisupervised model to identify commercial sex related posts. Bidirectional encoder representations from transformers (BERT) is a deep learning model developed by Google (Devlin et al. 2019). Pretrained BERT models can capture contextualized meaning in a sentence, generate sentence embeddings using pooling strategies, and have proven useful in a variety of applications such as detecting offensive tweets (Zhu et al. 2019b) and fake news (Jwa et al. 2019). Esfahani et al. (2019) presented a method that combines BERT with other language models, and trained a classifier to detect trafficking in online advertisements.

We train a Doc2Vec model and use a pretrained BERT model to extract embeddings for massage business reviews on open websites written by customers instead of advertisements written by traffickers. These two types of texts occur at different stages of the exploitation process (Caltagirone 2017). We aim to identify commercial sex acts or other risk factors for human trafficking. We found one paper with a similar objective of predicting commercial sex acts from an open business review site. Helderop et al. (2019) analyzed hotel reviews from Travelocity.com and trained a random forest classifier to predict whether a hotel had high prostitution activity. The features they considered included embeddings extracted using fastText (Joulin et al. 2016) and hotel price and geographic information. We focus on massage businesses instead of hotels. Furthermore, our models make predictions for each review. In contrast, Helderop et al. (2019) made predictions at the hotel level by combining all reviews from a hotel into one block of text. Because IMBs disguise as legitimate businesses, they have a mixture of illicit and nonillicit reviews. A single illicit review could warrant a closer look from law enforcement or victim service providers. By predicting at the review level, we can prioritize businesses based on the distribution of illicit reviews.

2.2. Detection of IMBs Through Text Analysis

We found three previous studies that analyzed massage business reviews using machine learning models. de Vries and Radford (2021) conducted stakeholder interviews to identify human trafficking "risk markers" in IMBs such as the rotation of victim workers. They obtained seed words from the interviews and trained a skip-gram model to detect similar terms in reviews from an online review board for sexual services. The authors used this approach to create a list of IMB risk markers but did not predict the risk of a given review from the risk markers present.

Through a private business partnership, a member of our team worked to develop *Artemis*, a tool that aggregates large sums of data and uses machine learning to predict massage businesses at risk for human trafficking (Vyas and Caltagirone 2019, Accenture 2020). This tool was created for use by law enforcement and private companies and is not publicly available. We explore different machine learning methods than those used in Artemis and create a new lexicon to recognize the specific language used in customer reviews on open websites such as Yelp. Furthermore, the analysis for Artemis was limited to reviews from massage businesses in Florida. We train and evaluate our models on reviews from multiple states.

To the best of our knowledge, only one previous work considered identifying commercial sex or other risk factors of human trafficking from massage business reviews on an open website. Diaz and Panangadan (2020) proposed an automated review labeling process and trained a random forest classifier on Yelp reviews. Our work differs from Diaz and Panangadan (2020) in three ways. First, we label reviews as illicit or nonillicit with input from domain experts. Diaz and Panangadan (2020) automatically labeled Yelp reviews as illicit only if the business appears on Rubmaps. Second, Diaz and Panangadan (2020) extracted features using the bag-of-words method with TF-IDF weighting, whereas we use lexicon-based and embedding-based models (i.e., BERT and Doc2Vec) to extract features. Last, Diaz and Panangadan (2020) removed infrequent terms appeared in less than 6% of all reviews to reduce the model's computational burden. However, we recognize that most terms related to human trafficking risk factors have low frequency across all reviews. Our lexicon-based approach ensures that these terms are considered regardless of their frequency.

3. Data Sources and Preparation

In this section, we describe our two data sources: Rubmaps and Yelp reviews. We also discuss the labeling of

Yelp reviews and preprocessing of the review text. Both data sources are supplied by our collaborators at the counter-human trafficking nonprofit: Global Emancipation Network. The first data set, obtained from the IMB review site Rubmaps, includes reviews and location information for 10,058 massage businesses across the United States. Review dates range from 2011 to 2019.

Yelp hosts customer reviews of businesses in many industries. We obtained a Yelp data set that contains reviews from massage related businesses in California, Florida, Georgia, Texas, and the Washington, DC metropolitan area. These states were chosen by our collaborators because of existing partnerships with local agencies or because they are known IMB hotspots (Heyrick Research 2021). The data include reviews that were available on Yelp between 2019 and 2020. Although some reviews were posted as early as 2005, around 85% of all reviews across the five regions were posted since 2015. The Yelp data set includes business information such as the address, phone number, and services offered in addition to review details such as review text and rating (one to five). We then filtered the Yelp data to keep only reviews from businesses listing at least one form of massage or spa treatment as one of their business services. After applying this filter, 430,682 reviews remained from 64,676 businesses across the five states.

3.1. Labeling Yelp Reviews

Obtaining ground truth for human trafficking is difficult. Some previous literature have relied on contributions from law enforcement or victim survivors to manually label classified advertisements (Alvari et al. 2017, Nagpal et al. 2017, Tong et al. 2017). There have also been some efforts to automate the labeling process; however, avoiding a labor-intensive labeling process requires making assumptions that can reduce the accuracy of the labels. For example, Dubrawski et al. (2015) used phone numbers of known traffickers to label instances of suspected trafficking in advertisements, but traffickers can change phone numbers frequently. Diaz and Panangadan (2020) used Rubmaps data to label the Yelp reviews. They assumed that all Yelp reviews from a massage business on Rubmaps are illicit, and all Yelp reviews from massage businesses that are not on Rubmaps are not illicit. However, Bouché and Crotty (2018) showed that some massage businesses on Rubmaps were likely not illicit. Based on our summary of Yelp businesses in Florida (Table A2 in the online appendix), we identified 108 businesses that were on Rubmaps but had no illicit labeled Yelp reviews and 42 businesses that were not on Rubmaps but had at least one illicit labeled Yelp review.

Ramchandani et al. (2021) used an active learning approach that combine manual efforts to label online commercial sex advertisements as either recruitment or sales posts. We followed a similar approach and began by manually labeling a subset of Yelp reviews as illicit and nonillicit. Because IMBs make up a small proportion of all massage businesses on Yelp, most Yelp reviews describe legitimate massage experiences. Additionally, sex buyers usually try to conceal illicit activities from the public, further reducing the number of illicit Yelp reviews. Therefore, we developed a targeted search process to identify reviews that are most likely to be illicit to account for the low representation of illicit Yelp reviews. We then used a voting process that required two of the three reviewers to agree on the label.

The following steps, L1 through L6, describe the search criteria we used for identifying reviews to label. We considered all Yelp reviews that met one of these criteria. Steps L1 through L5 were applied to Florida Yelp reviews, and step L6 was applied to Yelp reviews from Washington, DC, Texas, and Georgia. The steps were conducted in order, as some steps required information about previously labeled reviews. Figure 1 illustrates the process in more detail. The number of reviews labeled as nonillicit (zero) and illicit (one) in each step are displayed in the rectangular boxes on the arrows.

L1: Yelp/Rubmaps Location Intersection. We geocoded business addresses using the Google Maps Platform. We labeled all Yelp reviews for each business whose location matches a Rubmaps business location by latitude, longitude, and address suite number.

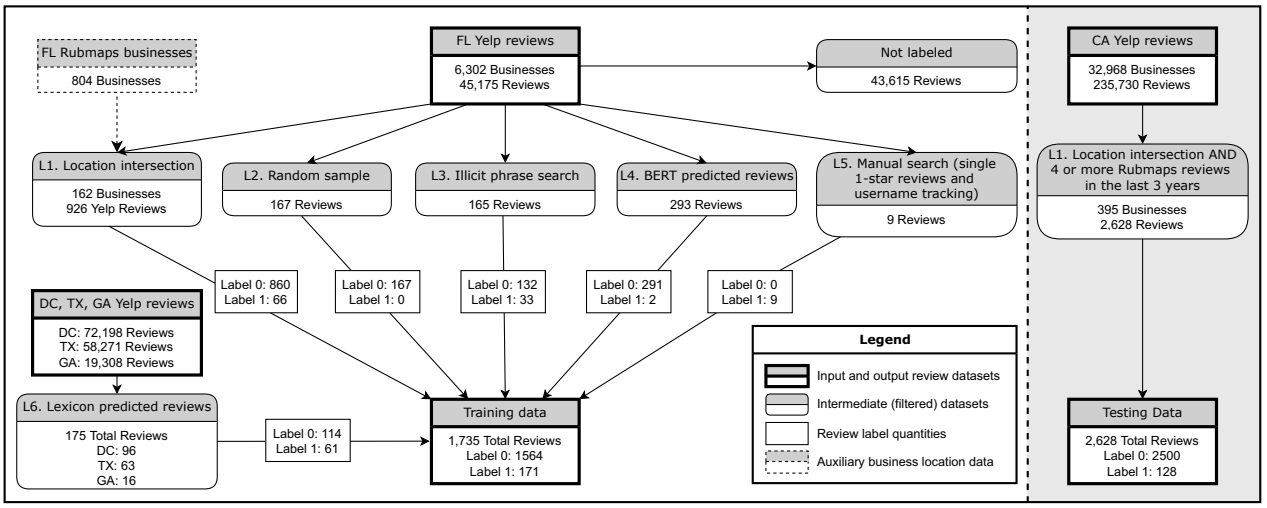
L2: Random Sample. To balance the types of businesses represented, we took a random sample of 500 reviews from businesses that did not match a Rubmaps business location. After filtering reviews from businesses that didn't list massage or spa services, 167 reviews remained.

L3: Illicit Phrase Search. We labeled reviews containing a word or phrase that was significantly more common in illicit reviews than nonillicit reviews, for example, "happy ending," "prostitution," and "investigation." We will elaborate on this process and how it contributed to the Yelp lexicon-based model in Section 4.1.1.

L4: BERT-predicted Reviews. We used BERT to extract embeddings for all Florida Yelp reviews in our data set. We trained a logistic regression model using the embeddings of labeled reviews up to this point to classify reviews. The model was applied to all unlabeled reviews after step L3, and we labeled the reviews that were classified as illicit.

L5: Manual Search for Illicit Reviews. We searched for two criteria to identify and label additional reviews: (i) A massage business had only one review and that review gave a rating of one star (out of five) or (ii) a review that was written by an author of another review previously determined to be illicit. We did not include all labeled reviews that met these criteria in the labeled data set, only the ones labeled as illicit.

Figure 1. Steps Used to Create the Labeled Review Data



Notes. The input data sets were filtered on the specified criteria in the order presented. A 0/1 label indicates a nonillicit or illicit review, respectively.

L6: Lexicon-predicted Reviews. From the labeled reviews up to this point, we created an initial lexicon based on the most frequent terms in illicit reviews. Section 4.1.1 provides more details of this process. We labeled the reviews from Washington, DC, Texas, and Georgia that were classified as illicit by the first lexicon-based classification model presented in Section 4.1.2.

Steps L1 through L6 created a labeled data set of 1,735 reviews, 171 of which were labeled as illicit. In this data set, 1,560 reviews came from Florida, 96 from Washington, DC, 63 from Texas, and 16 from Georgia. Step L1 was also applied to the California reviews to create a labeled data set of 2,628 reviews that were withheld for testing, 128 of which were labeled as illicit (Figure 1). Table A1 in the online appendix shows example Yelp reviews and their labels. There are generally two types of illicit Yelp reviews. One type is rated more favorably and written by customers looking for an IMB to buy sex. The other type tends to be more negative and is written by concerned customers complaining or warning other customers about suspected illegal activities. Table A3 in the online appendix shows example Rubmaps reviews (*Warning!* The content of these reviews is explicit). Compared with Rubmaps, Yelp reviews give a more holistic view of a business' services, staff, and facility, among other features. However, the context of illicit reviews on Yelp is usually more nuanced and expressed in plain English with less explicit phrases. Another challenge with the Yelp data are the small number of illicit reviews, resulting in an unbalanced labeled data set. We present classification models that address these two challenges in Section 4.

3.2. Text Preprocessing

The text of each review was preprocessed using standard Natural Language Processing (NLP) techniques and custom steps designed for IMB-specific language. The following steps were performed on each review text.

P1: Custom Contractions. We created custom contractions for IMB specific terms like "happyending" from "happy ending," "tableshower" from "table shower," and "handjob" from "hand job." This step prevents the loss of important IMB context. Without this step, the phrases "happy ending" and "happy with the end" would both be interpreted as "happy end," due to the stopword removal in step P3 and the lemmatization in step P4.

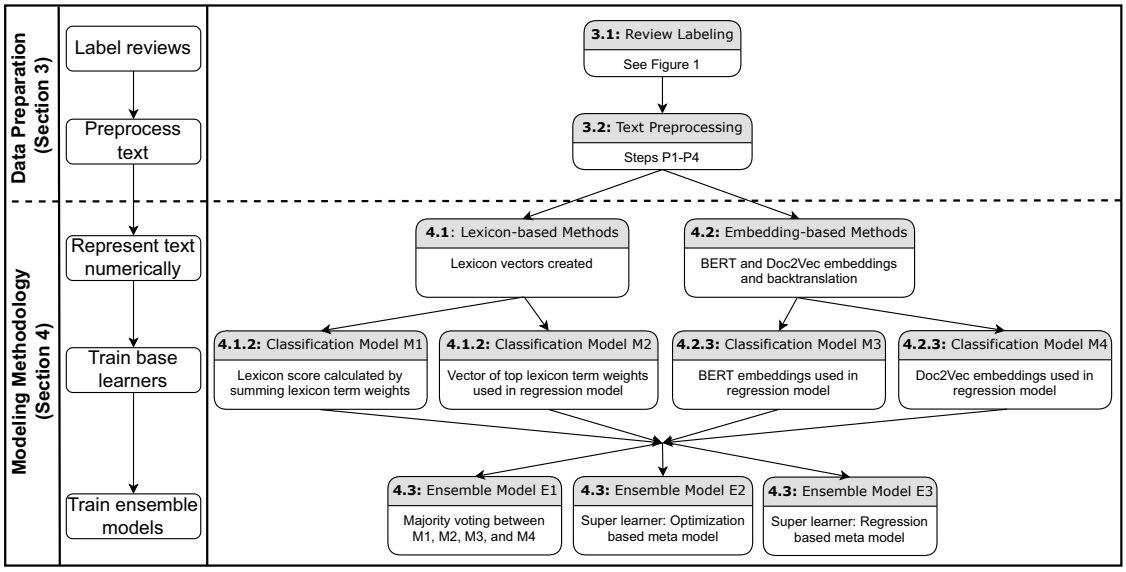
P2: Spelling Corrections. Misspellings are sometimes deliberate to conceal the meaning of sensitive information and illicit activities (Wang et al. 2012). We used the *pyspellchecker* package (Barrus 2021) to identify common misspellings of the lexicon terms (see Section 4.1) in all Florida Yelp reviews by adding the lexicon terms to the package's dictionary. We then corrected the misspellings to the properly spelled lexicon terms in all reviews.

P3: Stopword Removal. We removed the stopwords listed in the *NLTK* package (Bird et al. 2009). We customized this list to keep words that provide important context like "no," "not," and "only." For example, "men only" can imply commercial sex.

P4: Standard NLP Techniques. We applied lowercase conversion, decontraction, punctuation removal, tokenization using the RegexpTokenizer function, and lemmatization using the WordNetLemmatizer function in the *NLTK* package.

After text preprocessing, we represented reviews numerically through the methods discussed in Section 4. Figure 2 provides a flowchart of the steps discussed in this section and the next section.

Figure 2. Overview of Data Preparation and Modeling Steps



Notes. A simple linear flowchart of the five major steps is depicted on the left and the corresponding detailed steps are depicted on the right. The bold numbers indicate the section number where the step is described.

4. Methods

We first describe the lexicon-based and embedding-based classification approaches. We then develop ensemble models that combine both model types. Figure 2 outlines the proposed modeling methodology. A linear flowchart of the major steps is shown on the left side of Figure 2 with the corresponding detailed steps depicted in the branching flowchart on the right side. Two classification models are trained using the lexicon-based approach (Section 4.1) and two additional models are trained using the embedding-based approach (Section 4.2). All four models are combined in the ensemble models.

4.1. Lexicon Text Analysis

In this section, we first develop a lexicon from the terms used in labeled illicit reviews. We then propose two lexicon-based classification models.

4.1.1. Lexicon Development. We calculate the occurrence rate of each single word and two-word phrase (bigram) in illicit and nonillicit reviews. We then compute the ratio of illicit occurrence rate to nonillicit occurrence rate. For this analysis, the illicit reviews include all reviews labeled as illicit in labeling step L1, and the nonillicit reviews include all reviews that are not labeled as illicit at this point (including unlabeled reviews). Table 1 shows the 10 terms with the highest ratios, meaning an illicit review is associated with the highest odds of having these terms.

We use the top illicit terms to find more reviews to label in step L3 (see Section 3.1). For example, we label every Florida review that mentions "happyending," "prostitution," or "investigation." After labeling steps L4 and L5, we recompute the ratios and use the list of terms, sorted by ratio, as a starting point to identify terms to include in the Yelp-specific lexicon for IMBs. Based on

Table 1. Top 10 Terms with Highest Ratio of Occurrence Rate in Illicit Reviews to Occurrence Rate in Nonillicit Reviews (from Florida Yelp Reviews)

Term	Illicit rate (per 100,000 terms)	Nonillicit rate (per 100,000 terms)	Illicit:nonillicit ratio
Pimp	62.66	0.00	Na
Prostitution	112.80	0.04	2,664.58
Tableshower	225.59	1.06	213.17
Arrest	62.66	0.30	211.47
Sexual	100.26	0.51	197.38
Investigation	62.66	0.34	185.04
Happyending	200.53	1.19	169.18
Sex	87.73	1.31	66.85
Extra service	63.58	1.29	49.11
Asian massage	127.16	2.85	44.65

Notes. For this analysis, all unlabeled reviews are assumed to be nonillicit. This table only shows terms that occur at least five times.

input from counter-human trafficking experts at Global Emancipation Network, we consider the top 1,000 terms that occur at least twice to identify those related to commercial sex or other risk factors of trafficking. We identify 38 relevant terms, 24 of which are in the top 100. We include different forms of these terms in the lexicon and identify additional terms using domain knowledge of IMB characteristics such as locked doors and covered windows. We assign each term a weight of one or two based on expert opinion. Terms with weight 2 are strong signs of an illicit review such as "prostitution" or "happyending," and terms with weight 1 are potential signs of an illicit review such as "extra service" or "men only." This initial lexicon is used to create a classification model as described in the next section. We use this model to identify more reviews to label in step L6 (see Section 3.1). After this step, we further expand the lexicon by identifying synonyms of the existing terms. We review the synonyms and identify terms to add to the lexicon, for example "intercourse" as a synonym for "sex." The final lexicon includes 169 terms and is available upon request.

4.1.2. Lexicon-Based Classification Models. We develop two classification models based on the lexicon. The first model, referred to as M1, calculates a total score by summing the weights of the lexicon terms in a review. Because some reviews are lengthy, we set a limit on the number of terms counted in each review. If a review contains more terms than the maximum number counted, only the lexicon terms with the highest weights are considered. After summing the weights of the terms in each review, we normalize the total scores to give all reviews a score between zero and one. We then classify a review as illicit or nonillicit by selecting a decision threshold. We perform parameter tuning on the decision threshold and the number of counted terms to maximize the F1 score. We consider thresholds in 0.05 increments from 0 to 1 and three to eight counted terms. Results of the experiments with M1 are presented in Section 5.1. Table A4 in the online appendix shows how to score an example review using the lexicon.

The simplicity of model M1 is desirable for its interpretability and ease of application. However, the single score assigned to a review leads to loss of information regarding the individual weights of the lexicon terms. For example, when the top four terms are counted, a review with two terms of weight 2 will score the same as a review with four terms of weight 1. To address this issue, we design a second model, M2, which considers the individual scores of the highest scoring lexicon terms in each review. Model M2 also controls the number of terms counted. The model input for each review is a vector of the highest scores of lexicon terms in the review, in descending order, followed by zeroes if the review has less than the maximum counted lexicon terms. For example, when the top four terms are counted, the vector [2, 1, 1, 0] means the review contains one term of weight 2 and two terms of weight 1. We then train a logistic regression model to predict the review label from these vectors. Results for model M2 are presented in Section 5.1.

4.2. Embedding-Based Models and Data Augmentation

This section presents another text analysis approach where we generate numerical vector representations (i.e., embeddings) of reviews through the pretrained BERT model (Devlin et al. 2019) and the Doc2Vec model (Le and Mikolov 2014). We then train logistic regression models using embeddings to classify labeled illicit and nonillicit reviews. We also apply a data augmentation technique to account for the small number of illicit reviews.

4.2.1. BERT-Based Classification Model. We use the *bert-as-service* package (Xiao 2018) to extract 1,024 dimensional embeddings for labeled reviews from the pretrained BERT-Large-Uncased model. This package implements a pooling method on the second-to-last layer of BERT to generate embeddings that are less biased to the pretraining tasks (Xiao 2019). We keep all input values as default except setting the maximum word sequence length to 150, that is, the first 150 words of each review are considered when generating the embedding. Setting this parameter to a small value improves the speed of extracting the embeddings (Xiao 2018). In our Yelp review data set, only 11% of reviews have more than 150 words after preprocessing. We perform 10 replications of fivefold stratified cross-validation. In each fold, we train a logistic regression on the embeddings to classify reviews as illicit and nonillicit. The performance of this approach is reported in Section 5.2.

4.2.2. Doc2Vec-Based Classification Model. We train a Doc2Vec model using the distributed bag-of-words algorithm on all reviews from Washington, DC, Georgia, Florida, and Texas through the *gensim* package (Řehůřek and Sojka 2010). We set the minimum frequency of words to two. All other parameters are set to the default values. The Doc2Vec model is used to obtain an embedding for each review. We perform 10 replications of fivefold stratified cross-validation. In each fold, we train a logistic regression on the embeddings to classify reviews as illicit and nonillicit. We tune the epoch number and embedding dimension in the Doc2Vec model by considering average recall, F1 score, and area under the receiver operating characteristic curve (AUC) over 10 replications. We ultimately choose to use 150 epochs and a vector dimension of 600. The performance of this approach is reported in Section 5.2.

4.2.3. Data Augmentation. Illicit reviews are rare on Yelp. The number of nonillicit reviews is approximately

nine times larger than the number of illicit reviews in our labeled data set. This class imbalance is a challenge for training an accurate classifier. To address this problem, we perform data augmentation (DA) on the illicit reviews. DA aims to increase the size of a data set without collecting new data. Paraphrasing is an augmentation technique that modifies an original sentence to generate a new one by changing the sentence structure or word choices (Chen et al. 2021). We apply paraphrasing through back-translation. This method translates a text to another language and then back to the original language (Sennrich et al. 2016). In particular, we translate each illicit review from English to five other languages: Spanish, French, Chinese, German, and Russian using the GOOGLETRANSLATE function in Google Sheets. An additional illicit review is generated in the training set when the review in each language is translated back to English. We then train the previously discussed BERT and Doc2Vec classification models using the original and augmented reviews as the training set. Applying paraphrasing through transformation functions such as synonym replacement, random insertion, random swap, and random deletion (Wei and Zou 2019) generated results similar to back-translation in our preliminary experiments.

Back-translation can also be applied to the test set through a method called test time augmentation (TTA). With TTA, predictions for different versions of the test data are combined into one prediction for the original test data (Shorten and Khoshgoftaar 2019). Previous works have shown that TTA can improve accuracy (Wang et al. 2019) and robustness (Moshkov et al. 2020). We apply TTA on the leave-out test set. Each review in the test set is back-translated from five languages, generating five new versions of the test set. The final predicted labels are assigned according to the average predicted probabilities across all six versions of the reviews. We report classification results with no back-translation (BERT-No-DA; D2V-No-DA), with back-translation on training data only (BERT-DA-Training; D2V-DA-Training), and with back-translation on training and test data (BERT-DA-Training&Testing; D2V-DA-Training&Testing) in Section 5.2. We ultimately select BERT-DA-Training&Testing (referred to as M3) and D2V-DA-Training (referred to as M4).

4.3. Ensemble of Lexicon-Based and Embedding-Based Models

We propose two approaches to identify illicit reviews on Yelp. The first approach consists of models M1 and M2 based on the lexicon. The second approach consists of models M3 and M4 based on embeddings. We refer to models M1–M4 as *base learners* (Figure 2). The lexicon-based models can reliably identify evident illicit reviews with high precision, but they generate more false negatives than the embedding-based models, thus having lower recall. The embedding-based models, on the other hand, can recognize subtle semantic meaning and contextual language of the illicit reviews on Yelp, and thus they have higher recall but relatively low precision. We can achieve better results by combining the two approaches through ensemble learning. Bagging (Breiman 1996), boosting (Freund and Schapire 1996), and stacking (Wolpert 1992) are three main ensemble learning approaches. The bagging method helps reduce overfitting through bootstrap aggregation. The boosting method corrects prediction errors through iterations. We use the stacking method that combines multiple base learners into a single model.

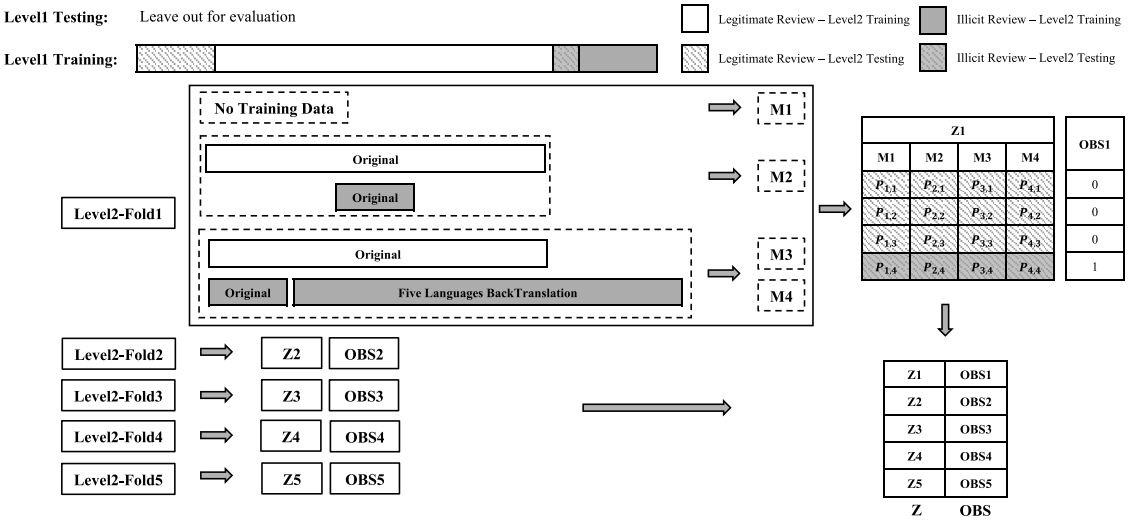
We consider two stacking methods. First, we implement the voting ensemble that is referred to as E1. This method returns the majority label among all base learners as the predicted label. The limitation of the majority voting ensemble is that each learning algorithm has the same weight. Second, we use the super learner ensemble method (van der Laan et al. 2007, Polley and van der Laan 2011). van der Laan et al. (2007) first proposed the super learner ensemble method and showed that it performs at least as well as the best base learner asymptotically. The super learner method is applied to classification problems in many fields such as precision medicine and image classification (Luedtke and van der Laan 2016, Ju et al. 2018). We implement the super learner ensemble through two meta-models: a linear optimization model and a logistic regression model. We refer to these two ensemble models as E2 and E3, respectively. In all three ensemble models E1, E2, and E3 (Figure 2), we first perform fivefold stratified data split. We train base learners on the training set of each fold and then make predictions for the reviews in the corresponding test set. We test two and three as majority vote thresholds for E1. We report results with threshold 2 in Section 5.3 due to its better performance. For the super learner ensemble models E2 and E3, we execute the following steps in each fold:

1. We store the predicted illicit probabilities from the base learners (M1–M4) for each review in the test set in a matrix $\boldsymbol{Y}$ of size $N_{test} \times 4$, where N_{test} is the number of test set reviews.

2. We split the training set into level 2 training and test sets through stratified fivefold cross-validation (Figure 3). For each level 2 fold, we retrain the base learners using the level 2 training set. Then, we predict the reviews in the corresponding level 2 test set with all four base learners and store the predicted illicit probabilities in a matrix. We denote the final matrix as $\boldsymbol{Z}$, which is the concatenation of five matrices, one for each level 2 fold. The size of matrix $\boldsymbol{Z}$ is $N_{train} \times 4$, where N_{train} is the size of the training set.

3. We train meta-models for ensembles E2 and E3 using the $\boldsymbol{Z}$ matrix. There is a true observation (label) for each row in the $\boldsymbol{Z}$ matrix (each row corresponds to a review).

Figure 3. Processes to Generate Input Matrix (i.e., $\boldsymbol{Z}$) for Training the Super Learner Ensemble Models in One Level 1 Fold



(a) For E2, we consider a linear model and optimize the weights of four base learners (M1–M4) to maximize the F1 score. We choose the F1 score as the objective because of class imbalance (Sun et al. 2009). The formulation of the F1 score is not convex. Hence, traditional gradient descent methods cannot find the exact optimal weights. Instead, we implement a meta-heuristic called differential evolution (Storn and Price 1997) in the *SciPy* package (Virtanen et al. 2020).

(b) For E3, we train a logistic regression meta-model using the $\boldsymbol{Z}$ matrix and the labels.

4. We input the $\boldsymbol{Y}$ matrix, that is, the predicted probabilities from the four base learners for the reviews in the test set, to ensemble models E2 and E3 to classify each review in the test set.

We repeat these four steps for five folds and compute the average performance. The results for each base learner and ensemble model are presented in Section 5.3.

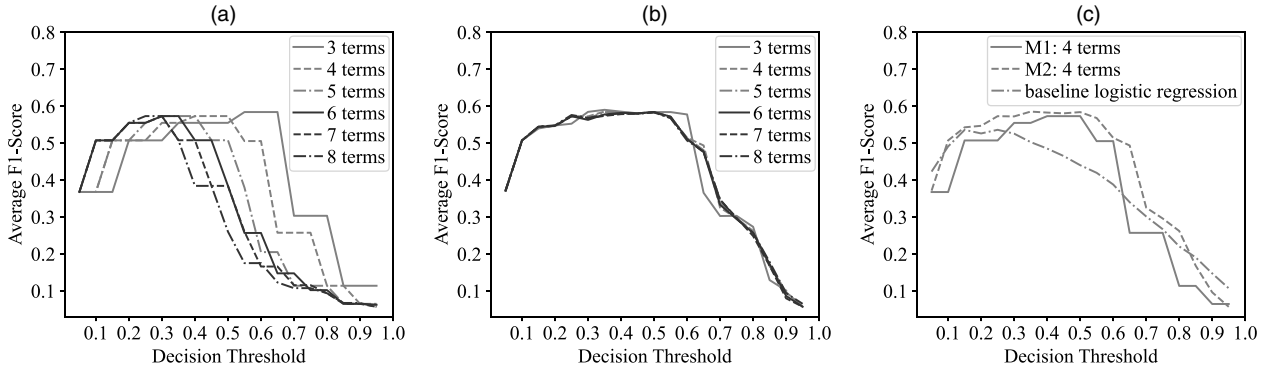
5. Computational Results and Discussion

This section presents our results for detecting illicit reviews in two separate Yelp data sets. The first data set is used for model training and testing, and it includes 1,564 nonillicit reviews and 171 illicit reviews (Figure 1). The second data set contains Yelp reviews from California and is withheld for out-of-sample testing. We first report the results from the lexicon-based classification models. We then evaluate the embedding-based models with and without DA. Finally, we report the performance of the ensemble models and base learners. The code developed in this work is available to relevant researchers and practitioners upon request at https://zenodo.org/record/7407511#.Y5ipmOzMLKI. The data can be requested from the Global Emancipation Network for approved uses established by a data use agreement.

5.1. Performance of the Lexicon-Based Models

We present the average F1 score for models M1 and M2 with different decision threshold values and different

Figure 4. Parameter Tuning Experiments for Lexicon-Based Models



Notes. This figure displays the average F1-score over 10 replications of fivefold cross-validation. (a) M1 for three to eight counted terms. (b) M2 for three to eight counted terms. (c) M1 and M2 (four terms) and baseline.

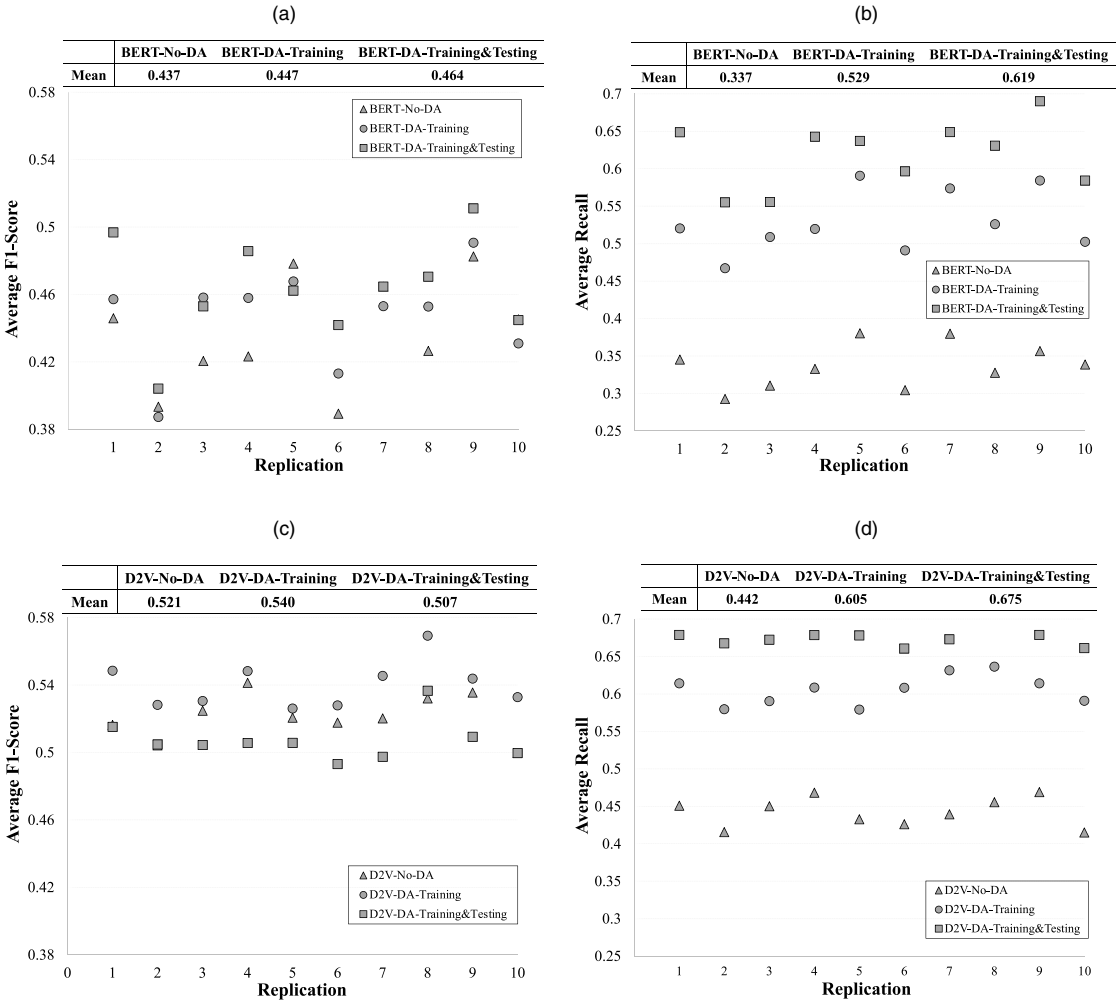
number of counted terms in Figure 4(a) and (b), respectively. For model M1, the optimal threshold decreases as the number of lexicon terms counted increases. This is because most reviews contain a small number of lexicon terms while a small portion contain many lexicon terms. As more terms are counted, the maximum possible lexicon score increases, but a review's score only increases if the review contains more lexicon terms. After normalizing each review's score by the maximum score, there are more reviews with lower scores as the terms counted increases, resulting in a lower optimal threshold. We choose the version that counts the top four lexicon terms in a review because it has one of the highest F1 scores and uses a natural decision threshold of 0.5 for classifying reviews.

All versions of model M2 with a different number of counted lexicon terms perform similarly at the 0.5 threshold. In our experiments, counting more than four terms does not greatly impact the predictions. Hence, we choose to use the M2 model that counts the top four terms for consistency with model M1. In Figure 4(c), we compare the results of M1 and M2 with a baseline logistic regression model which is trained on the binary vectors that indicate the existence of each lexicon terms for each review. Both M1 and M2 achieve higher F1 scores than the baseline. Model M2 performs slightly better than M1; however, M1 is a simpler model.

5.2. Performance of the Embedding-Based Models

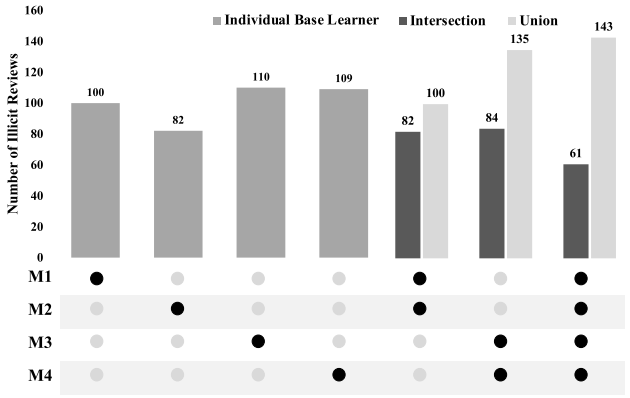
We evaluate three different versions of the BERT-based and Doc2Vec-based classification models: with no back-translation (BERT-No-DA; D2V-No-DA), with back-translation on training data only (BERT-DA-Training; D2V-DA-Training), and with back-translation on training and test data (BERT-DA-Training&Testing; D2V-DA-Training&Testing). Figure 5(a) and (b), displays the average F1 score and recall for BERT-based models,

Figure 5. Effect of DA on F1 Score and Recall in Each Replication for BERT-Based and Doc2Vec-Based Classification Models



Notes. (a) BERT: The effect of DA on F1 score. (b) BERT: The effect of DA on recall. (c) Doc2Vec: The effect of DA on F1 score. (d) Doc2Vec: The effect of DA on recall.

Figure 6. Number of Illicit Reviews Identified by Base Learners



Notes. The black circles indicate which models are considered. For example, the rightmost bar (four black circles) means that 61 illicit reviews are found by all models, and 143 are found by at least one of the models.

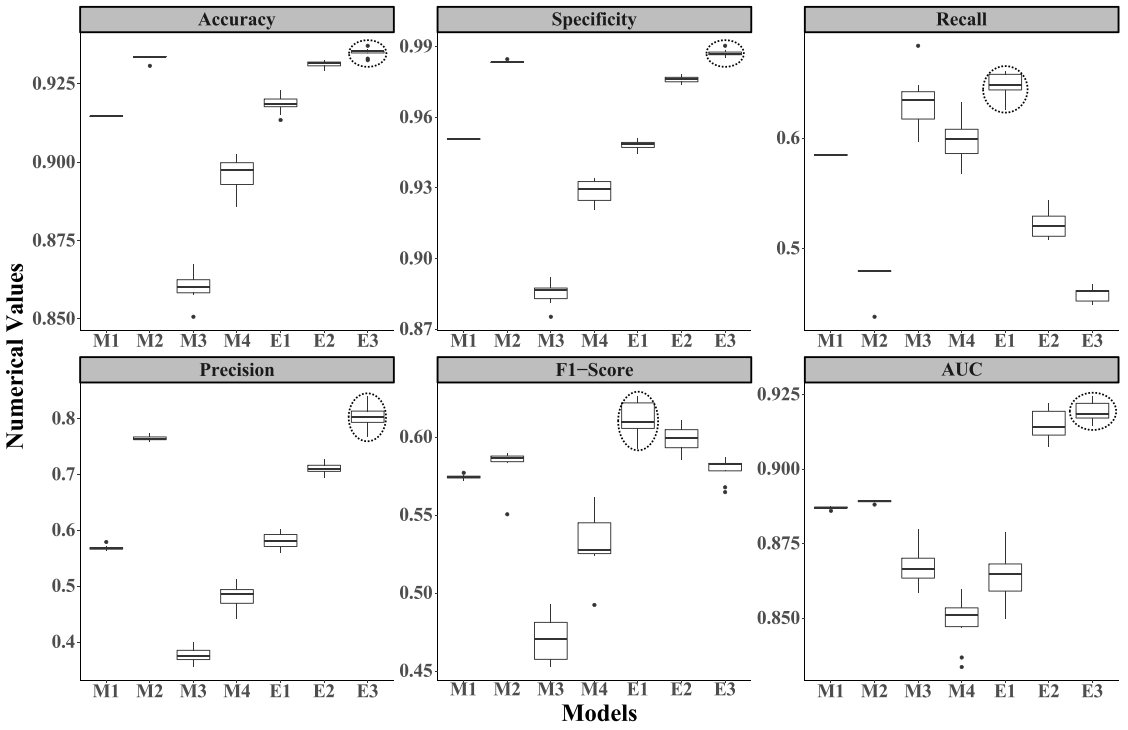
and Figure 5(c) and (d), displays the average F1 score and recall for Doc2Vec-based models in 10 replications of the fivefold data split. Each marker in the figures indicates the average performance over five folds. In Figure 5(a), using back-translation on the training and test data yields the highest F1 score in 8 of 10 replications. In Figure 5(c), using back-translation only on the training data provides the highest F1 score in all replications. Furthermore, there is a consistent improvement in recall with DA in all experiments as shown in Figure 5(b) and (d). We also display the average performance over all 10 replications in the table above each figure. These results indicate that using back-translation can improve the classification performance. Thus, we select BERT-DA-Training&Testing (referred to as M3) and D2V-DA-Training (referred to as M4).

5.3. Performance of the Base Learners and Ensemble Models

Figure 6 shows the number of illicit reviews found by each base learner and their intersections and unions. M1, M2, M3, and M4 detect 100, 82, 110, and 109 illicit reviews, respectively. Overall, four base learners identify 143 unique illicit reviews, 84% of all illicit reviews in the labeled data set. There are 61 illicit reviews detected by all models and 28 illicit reviews detected by none of them. Eight illicit reviews detected by the lexicon-based models are not detected by the embedding-based models, and 43 illicit reviews detected by the embedding-based models are not detected by the lexicon-based models.

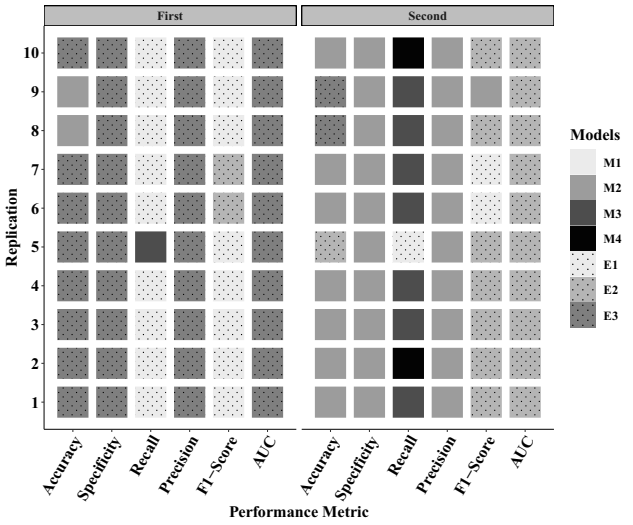
We evaluate the classification performance of base learners and ensemble models through 10 replications of stratified fivefold cross-validation. Figure 7 shows the distribution of six metrics for each model. For all performance metrics, lexicon-based models M1 and M2 have less variation across replications. They are

Figure 7. Performance of Base Learners and Ensemble Models on the Labeled Data Set over 10 Replications



Notes. Boxplots show percentile values. Circles indicate best performing model for that metric.

Figure 8. First and Second Ranking Models for Various Metrics on the Labeled Data Set in Each of 10 Replications



more stable and competitive than the embedding-based models M3 and M4 across all metrics except recall. The illicit reviews detected by M2 is a subset of those detected by M1 (Figure 6). Thus, M1 has higher recall. However, M2 has higher accuracy, specificity, precision, F1 score, and AUC than M1. Among two embedding-based models, M3 has higher recall and AUC but performs worse than M4 in other metrics. The ensemble models yield the best performance across all six metrics. E1 has the highest recall and F1 score, whereas E3 has the highest accuracy, specificity, precision, and AUC. E2 performs between E1 and E3 in all metrics.

We also consider the performance in each replication. Figure 8 displays the first and second best models in each replication (y-axis) for six metrics (x-axis). We make the following observations:

- M2 has the best accuracy in 2 of 10 replications, and E3 has the best accuracy in the remaining 8 replications. Furthermore, the second best model for accuracy is E3 in each replication where M2 has the best accuracy.
- E3 has the highest specificity and M2 has the second highest specificity in all replications.
- E1 has the highest recall in all but one of the replications, and M3 has the second highest recall in most replications.
- E3 has the best precision in all replications, and M2 is a close second choice.
- E1 has the best F1 score in most replications, and E2 is a close second choice.
- E3 has the best AUC in all replications, and E2 is a close second choice.

Recall that we use the lexicon to expand the set of illicit reviews in the labeling process. This approach may create bias helping the lexicon-based models M1 and M2 and the ensemble model E1 perform better. To address this issue, we collect and label a second data set of 2,628 California Yelp reviews, which has 128 labeled illicit ones, from business locations listed on Rubmaps (see Section 3.1). The proposed base learners and ensemble models are trained on the original data set and then applied to classify reviews in this new data set.

Table 2 shows the results of this experiment. In each row, the bold entries are the highest score for a metric, and italic entries are the next highest score. The results indicate that the performance of M1, M2, and E1 are worse than their performance on the data set used for training, especially for recall and F1 score. Conversely, embedding-based models M3 and M4 yield the highest recall, and the overall performance of ensemble models is better than base learners. The results of this experiment highlight the strength of ensemble models, especially the super learner ensemble models E2 and E3, for improving performance of the base learners.

5.4. Summary of the Numerical Experiments

The lexicon-based models have high precision because they use a list of specific terms that are reliable indicators of illicit activity. However, the predictions are based on a limited dictionary that does not capture nuanced context expressed in plain English. This results in more conservative predictions and a lower recall of the illicit reviews. Meanwhile, the embedding-based models exhibit higher recall because they can identify illicit reviews with subtle language elements. To improve the

Table 2. Model Performance on California Yelp Reviews

	M1	M2	M3	M4	E1	E2	E3
Confusion Matrix	2,486 14 93 35	2,493 7 101 27	2,289 211 46 82	2,321 179 52 76	2,445 55 62 66	2,490 10 68 60	2,491 9 81 47
Accuracy	0.9593	0.9589	0.9022	0.9121	0.9554	**0.9703**	*0.9658*
Specificity	0.9944	**0.9972**	0.9156	0.9284	0.9780	0.9960	*0.9964*
Recall	0.2734	0.2109	**0.6406**	*0.5938*	0.5156	0.4688	0.3672
Precision	0.7143	0.7941	0.2799	0.2980	0.5455	**0.8571**	*0.8393*
F1 score	0.3955	0.3333	0.3895	0.3969	*0.5301*	**0.6061**	0.5109

Notes. For the confusion matrix, clockwise from top left: true negative, false positive, true positive, false negative. In each row, the bold entries are the highest score for a metric, and italic entries are the next highest score.

lexicon-based models, we can refine the lexicon using results from the embedding-based models. Specifically, we can analyze the reviews that are only classified as illicit by the embedding-based models and consider adding the relevant terms to the lexicon. Yelp reviews describe illicit activities using plain English in contrast to the slang terms and acronyms frequently used in Rubmaps reviews or other sex buyer forums. Thus, we do not expect large changes in the language of Yelp reviews at least in the short term. Furthermore, our approach is robust to temporal phrase changes because these trends can be monitored over time to update the lexicon as necessary. In addition to expanding and updating the lexicon, we can refine the weights assigned to the lexicon terms. We currently assign weights based on domain expertise but could explore quantitative methods for determining the optimal term weights (Ustun and Rudin 2019).

The proposed ensemble models build on the individual strengths of each base learner. There is not one single model that outperforms in every metric. Thus, the best model to use depends on the user's objective. A framework for model selection based on user preferences is presented in Swan et al. (2021). For example, law enforcement may want to prioritize investigations with the highest precision so as not to waste resources on false positives. Conversely, organizations that help victims may choose a model with higher recall to reach most of the potential victim-workers.

6. Conclusions and Future Work

We propose a text analysis approach for detecting illicit reviews containing potential risk factors for human trafficking. There are limited resources for identifying suspected illicit businesses that exploit victim-workers. Currently, investigators make substantial manual efforts to sort through evidence including business reviews. Our work can save valuable time by prioritizing risky businesses and pointing to specific evidence of illicit activity in business reviews. As the models are implemented in other regions or on reviews from other open websites, we can expand the labeled training data set to improve the classification performance.

We recognize that our models may result in excess focus or pressure on consensual sex workers at massage businesses. However, when used in conjunction with other digital evidence to build human trafficking cases against suspected IMBs, the proposed models can help target traffickers more effectively which would reduce law enforcement interaction with consensual sex workers and trafficking victims. Current approaches to fight trafficking in the illicit massage industry rely on victim testimony and other interactions with massage business workers such as undercover stings, which can harm both consensual sex workers and trafficking victims. Law enforcement can use our models to build cases and supply evidence to justify warrants before interacting with massage business workers. Furthermore, victim service organizations can use our models to identify risky massage businesses and provide assistance to vulnerable people working in those places whether they are currently being trafficked or not to mitigate the unintended consequences.

Our classification models can be applied to reviews from other open websites like Google Maps. Furthermore, review-level classification results can inform a business-level model. Specifically, the risk level of a massage business can be derived from the reviews considering information like the number or percent of illicit reviews, review ratings, and the date of reviews. The review date is important because recent illicit reviews might indicate a current exploitation case that should be prioritized by law enforcement and victim organizations. In addition to customer reviews, several other data sources including massage business and therapist license records, geographic information, and business information like phone numbers, operating hours, website domain, foot traffic, and images can be combined to predict the risk of a massage business. This type of multimodal data integration, however, presents challenges for machine learning. One challenge is to accurately identify unique points of interest (POIs) from business address data. Identifying POIs rather than business locations ensures that we do not attribute data to the wrong business whether it is another nearby business or a business that existed at a different point in time. Another challenge pertains to data sparsity because it is unlikely that each business would be covered by all data sources.

The proposed text analysis methods have potential crime fighting applications in other commonly reviewed business domains that might serve as fronts for human trafficking. Examples include nail salons (Hultquist 2019), hotels (Paraskevas and Brookes 2018, Kragt 2020), housekeeping (Polaris Project 2019b), and home healthcare services (Michelen 2019). Furthermore, a similar approach could be developed to screen massage therapist job recruitment advertisements, truck stop reviews and advertisements, or farm labor job advertisements. Traffickers advertise to recruit victim-workers on various online platforms. Polaris has identified key phrases in massage therapist recruitment ads that may indicate trafficking (Polaris Project 2019a) and phrases that indicate commercial sex at truck stops (Polaris Project 2012). A comprehensive fight against trafficking requires intervention at all stages of the human trafficking kill chain (Caltagirone 2017), from recruitment (job advertisements) to exploitation (business reviews). Combining these methods with multimodal data integration and machine learning techniques can create powerful automated tools.

Acknowledgments

The authors thank the department editor Özlem Ergun, associate editor, and two referees for valuable comments and suggestions during the revision process, which have greatly improved this paper.

References

Accenture (2020) Exposing human trafficking networks with AI. Accessed September 15, 2021, https://www.accenture.com/us-en/case-studies/applied-intelligence/artemis.

Alvari H, Shakarian P, Snyder JEK (2017) Semi-supervised learning for detecting human trafficking. *Security Inform.* 6(1):1–14.

Barrus T (2021) pyspellchecker. Accessed May 17, 2021, https://pypi.org/project/pyspellchecker/.

Bird S, Loper E, Klein E (2009) *Natural Language Processing with Python* (O'Reilly Media, Sebastopol, CA).

Bouché V, Crotty SM (2018) Estimating demand for illicit massage businesses in Houston, Texas. *J. Human Trafficking* 4(4):279–297.

Breiman L (1996) Bagging predictors. *Machine Learn.* 24(2):123–140.

Caltagirone S (2017) The human trafficking kill chain: A guide to systematic disruption. Accessed November 19, 2021, https://www.globalemancipation.ngo/the-human-trafficking-kill-chain/.

Chen J, Tam D, Raffel C, Bansal M, Yang D (2021) An empirical survey of data augmentation for limited data learning in NLP. Preprint, submitted June 14, https://arxiv.org/abs/2106.07499.

Chin J, Takahashi L, Baik Y, Ho C, To S, Radaza A, Wu E, et al. (2019) Illicit massage parlors in Los Angeles county and New York City: Stories from women workers. Accessed July 26, 2022, http://johnchin.net/Article_Files/MP_Study_10.11.19_FINAL.pdf.

de Vries I, Radford J (2021) Identifying online risk markers of hard-to-observe crimes through semi-inductive triangulation: The case of human trafficking in the United States. *British J. Criminology* 62(3):639–658.

Demand Abolition (2018) Who buys sex? Understanding and disrupting illicit market demand. Accessed July 27, 2022, https://www.demandabolition.org/wp-content/uploads/2019/07/Demand-Buyer-Report-July-2019.pdf.

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Burstein J, Doran C, Solorio T, eds. *Proc. Conf. of the North Amer. Chapter of the Assoc. for Computat. Linguistics* (ACL, Stroudsburg, PA), 4171–4186.

Diaz M, Panangadan A (2020) Natural language-based integration of online review datasets for identification of sex trafficking businesses. Ceballos C, ed. *Proc. IEEE 21st Internat. Conf. on Inform. Reuse and Integration for Data Sci.* (IEEE, New York), 259–264.

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

Esfahani SS, Cafarella MJ, Pouyan MB, DeAngelo G, Eneva E, Fano AE (2019) Context-specific language modeling for human trafficking detection from online advertisements. Korhonen A, Traum D, Màrquez L, eds. *Proc. 57th Annual Meeting of the Assoc. for Comput. Linguistics* (ACL, Stroudsburg, PA), 1180–1184.

Federation of State Massage Therapy Boards (2017) Human trafficking task force report. Accessed July 27, 2022, https://www.fsmtb.org/media/1606/httf-report-final-web.pdf.

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Saitta L, ed. *Proc. 13th Internat. Conf. on Machine Learn.* (Morgan Kaufmann, San Francisco), 148–156.

Helderop E, Huff J, Morstatter F, Grubesic A, Wallace D (2019) Hidden in plain sight: A machine learning approach for detecting prostitution activity in Phoenix, Arizona. *Appl. Spatial Anal. Policy* 12:941–963.

Heyrick Research (2021) Snapshot: The illicit massage industry at a glance. Accessed September 9, 2021, https://www.heyrickresearch.org/research/what-is-the-illicit-massage-industry.

Hultquist I (2019) Human trafficking awareness for salon professionals. Accessed December 6, 2022, https://www.floridacosmetologist.com/human-trafficking-awareness-salon/.

Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T (2016) FastText.zip: Compressing text classification models. Preprint, submitted December 12, https://arxiv.org/abs/1612.03651.

Ju C, Bibaut A, van der Laan M (2018) The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Statist.* 45(15):2800–2818.

Jwa H, Oh D, Park K, Kang JM, Lim H (2019) exBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). *Appl. Sci.* 9(19):4062.

Kennedy E (2012) Predictive patterns of sex trafficking online. Senior honors thesis, Carnegie Mellon University, Pittsburgh.

Keskin B, Bott G, Freeman N (2021) Cracking sex trafficking: Data analysis, pattern recognition, and path prediction. *Production Oper. Management* 30(4):1110–1135.

Kragt G (2020) Human trafficking in the hospitality industry in the Netherlands. *Res. Hospitality Management* 10(2):131–136.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. Xing EP, Jebara T, eds. *Proc. 31st Internat. Conf. on Machine Learning* (PMLR, held virtually), 1188–1196.

Li Y, Yang T (2018) Word embedding for understanding natural language: A survey. *Guide to Big Data Applications* (Springer, Berlin), 83–104.

Luedtke AR, van der Laan MJ (2016) Super-learning of an optimal dynamic treatment rule. *Internat. J. Biostatist.* 12(1):305–332.

Michelen O (2019) 200 Filipino nurses win human trafficking lawsuit against SentosaCare and its owners. Accessed December 6, 2022, https://courtroomstrategy.com/2019/10/200-filipino-nurses-win-human-trafficking-lawsuit-against-sentosacare-and-its-owners/.

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. Preprint, submitted January 16; last revised September 7, 2013, https://arxiv.org/abs/1301.3781.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. *Adv. Neural Inform. Processing Systems* 26:3111–3119.

Moshkov N, Mathe B, Kertesz-Farkas A, Hollandi R, Horvath P (2020) Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Sci. Rep.* 10(1):1–7.

Nagpal C, Miller K, Boecking B, Dubrawski A (2017) An entity resolution approach to isolate instances of human trafficking online. Derczynski L, Xu W, Ritter A, Baldwin T, eds. *Proc. 3rd Workshop on Noisy User-Generated Text* (ACL, Stroudsburg, PA), 77–84.

Paraskevas A, Brookes M (2018) Human trafficking in hotels: An "invisible" threat for a vulnerable industry. *Internat. J. Contempory Hospitality Management* 30(3):1996–2014.

Polaris Project (2012) Sex trafficking at truck stops. Accessed October 28, 2021, https://humantraffickinghotline.org/resources/sex-trafficking-truck-stops.

Polaris Project (2019a) Human trafficking in illicit massage businesses. Accessed May 28, 2021, https://massagetherapy.nv.gov/Resources/Resource_Home/.

Polaris Project (2019b) New report spotlights the trafficking of nannies, house cleaners, other domestic workers in the U.S. Accessed December 6, 2022, https://polarisproject.org/press-releases/new-report-spotlights-the-trafficking-of-nannies-house-cleaners-other-domestic-workers-in-the-u-s/.

Polley EC, Rose S, van der Laan MJ (2011) Super learning. *Targeted Learning. Springer Series in Statistics*. (Springer, New York). https://doi.org/10.1007/978-1-4419-9782-1_3.

Ramchandani P, Bastani H, Wyatt E (2021) Unmasking human trafficking risk in commercial sex supply chains with machine learning. Preprint, submitted June 13; last revised August 4, 2022, https://dx.doi.org/10.2139/ssrn.3866259.

Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. Witte R, Cunningham H, Patrick J, Beisswanger E, Buyko E, Hahn U, Verspoor K, Coden AR, eds. *Proc. LREC Workshop on New Challenges for NLP Frameworks* (ELRA, Paris), 45–50.

Sennrich R, Haddow B, Birch A (2016) Improving neural machine translation models with monolingual data. Erk K, Smith NA, eds. *Proc. 54th Annual Meeting of the Assoc. for Comput. Linguistics* (ACL, Stroudsburg, PA), 86–96.

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *J. Big Data* 6(1):1–48.

Simonson E (2021) Semi-supervised classification of social media posts: Identifying sex-industry posts to enable better support for those experiencing sex-trafficking. Preprint, submitted April 7, https://arxiv.org/abs/2104.03233.

Storn R, Price K (1997) Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* 11(4):341–359.

Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: A review. *Internat. J. Pattern Recognition Artificial Intelligence* 23(4):687–719.

Swan BP, Mayorga ME, Ivy J (2021) The SMART framework: Selection of machine learning algorithms with ReplicaTions: A case study on the microvascular complications of diabetes. *IEEE J. Biomedical Health Inform.* 26(2):809–817.

Tong E, Zadeh A, Jones C, Morency LP (2017) Combating human trafficking with deep multimodal models. Barzilay R, Kan M-Y, eds. *Proc. 55th Annual Meeting of the Assoc. for Comput. Linguistics* (ACL, Stroudsburg, PA), 1547–1556.

U.S. Department of Justice (2017) Understanding the perspective of the victim: Recognizing the complexity of sex trafficking situations. Office of Juvenile Justice and Delinquency Prevention. Accessed August 3, 2022, https://ojjdp.ojp.gov/sites/g/files/xyckuh176/files/pubs/252021.pdf.

U.S. Department of Justice (2018) Justice Department leads effort to seize Backpage.com. Accessed October 7, 2021, https://www.justice.gov/opa/pr/justice-department-leads-effort-seize-backpagecom-internet-s-leading-forum-prostitution-ads.

U.S. Department of State (2021) Trafficking in persons report, 2021. Accessed July 21, 2021, https://www.state.gov/reports/2021-trafficking-in-persons-report/.

Ustun B, Rudin C (2019) Learning optimized risk scores. *J. Machine Learn. Res.* 20(150):1–75.

van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. *Statist. Appl. Genetic Molecular Biology* 6(1):25.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, et al. (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17:261–272.

Vyas M, Caltagirone S (2019) Combating human trafficking using analytics. Accessesd July 14, 2021, https://conf.splunk.com/files/2019/slides/BAS2793.pdf?podcast=1577146223.

Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T (2019) Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338:34–45.

Wang H, Cai C, Philpot A, Latonero M, Hovy EH, Metzler D (2012) Data integration from open Internet sources to combat sex trafficking of minors. Bertot JC, Luna-Reyes LF, Mellouli S, eds. *Proc. 13th Annual Internat. Conf. on Digital Government Res.* (ACM, New York), 246–252.

Wang L, Laber E, Saanchi Y, Caltagirone S (2020) Sex trafficking detection with ordinal regression neural networks. Ross D, Sinha A, Staheli D, Streilein B, eds. *Proc. AAAI-20 Workshop on Artificial Intelligence for Cyber Security*. Preprint, submitted August 15, 2019; last revised January 12, 2020. https://arxiv.org/abs/1908.05434.

Wei J, Zou K (2019) EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Preprint, submitted January 31; last revised August 25, 2019, https://arxiv.org/abs/1901.11196.

Wolpert DH (1992) Stacked generalization. *Neural Networks* 5(2):241–259.

Wu HC, Luk RWP, Wong KF, Kwok KL (2008) Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inform. Systems* 26(3):1–37.

Xiao H (2018) bert-as-service. Accessed July 20, 2021, https://bert-as-service.readthedocs.io/en/latest/index.html.

Xiao H (2019) Serving Google BERT in production using Tensorflow and ZeroMQ. Accessed July 25, 2022, https://hanxiao.io/2019/01/02/Serving-Google-BERT-in-Production-using-Tensorflow-and-ZeroMQ/.

Yakowicz W (2021) Inside the $4.5 billion erotic massage parlor economy. Accessed July 26, 2022, https://www.forbes.com/sites/willyakowicz/2021/04/04/inside-the-45-billion-erotic-massage-parlor-economy/?sh=7aef38eb79a8.

Zhu J, Li L, Jones C (2019a) Identification and detection of human trafficking using language models. Brynielsson J, ed. *Proc. Eur. Intelligence and Security Inform. Conf.* (IEEE, New York), 24–31.

Zhu J, Tian Z, Kübler S (2019b) UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM, eds. *Proc. 13th Internat. Workshop on Semantic Evaluation* (ACL, Stroudsburg, PA), 788–795.