Canonical Autocorrelation Analysis and Graphical Modeling for Human Trafficking Characterization

Qicong Chen
Carnegie Mellon University
Pittsburgh, PA 15213
qicongc@cs.cmu.edu

Maria De Arteaga
Carnegie Mellon University
Pittsburgh, PA 15213
mdeartea@andrew.cmu.edu

William Herlands
Carnegie Mellon University
Pittsburgh, PA 15213
herlands@cmu.edu

Abstract

The present research characterizes online prostitution advertisements by human trafficking rings to extract and quantify patterns that describe their online operations. We approach this descriptive analytics task from two perspectives. One, we develop an extension to Sparse Canonical Correlation Analysis that identifies autocorrelations within a single set of variables. This technique, which we call Canonical Autocorrelation Analysis, detects features of the human trafficking advertisements which are highly correlated within a particular trafficking ring. Two, we use a variant of supervised latent Dirichlet allocation to learn topic models over the trafficking rings. The relationship of the topics over multiple rings characterizes general behaviors of human traffickers as well as behaviours of individual rings.

1 Introduction

Currently, the United Nations Office on Drugs and Crime estimates there are 2.5 million victims of human trafficking in the world, with 79% of them suffering sexual exploitation. Increasingly, traffickers use the Internet to advertise their victims' services and establish contact with clients.

The present reserach chacterizes online advertisements by particular human traffickers (or trafficking rings) to develop a quantitative descriptions of their online operations. As such, prediction is not the main objective. This is a task of descriptive analytics, where the objective is to extract and quantify patterns that are difficult for humans to find.

Such characterizations are important to law enforcement agents who track and apprehend human traffickers. We hope this research will deepen their understanding of how each network operates including who are their victims, who are their clients, and where they operate. Additionally, this reserach seeks to broaden the general knowledge on human trafficking since there is scant quantitative information regarding the market and its victims[8].

We focus on two machine learning techniques: Canonical Correlation Analysis (CCA) and Supervised Latent Dirichlet Allocation (sLDA). The output of both methods are intuitive to interpret for non-machine learning experts, making them ideal for projects involving law enforcement. While the variant of CCA we developed, Canonical Autocorrelation Analysis (CAA), characterizes relationships amongst features in each trafficking ring, sLDA discovers topics which describe how features relate across groups of trafficking rings.

Below we provide a brief description of the data followed by a review of related works, descriptions of our proposed methodological innovations, results and conclusions.

2 Data

The Auton Lab has scraped 13,907,048 ads from a large advertising website for prostitution. Each ad has 34 features including physical descriptions and location information of those advertised, contact details, and meta-information about the ad itself. These features were extracted using image and natural language processing.

We clustered the data into non-overlapping clusters by identifying phone numbers in the ads which occurr over multiple ads. Additionally, using a list of 1,700 phone numbers linked to human traffickers provided by a non-profit organization, we partioned the data into clusters of ads linked to human trafficking (”positive” ads) and clusters of ads not linked to human trafficking (”negative” ads).

2.1 Preprocessing

One challenging aspect of the dataset is that features may have multiple values assigned to them. For example, there may be multiple women advertised in a single ad. To address this possibility for categorical features, we assign one bit to each possible value in order to record the number of times this value appears in the current ad. For example, for *hair color* we assign one bit to each possible hair color and count the number of times each of these appears in an ad. For numeric features, we assign three bits to record minimum, mean and range of all values in each feature. For example, if there are n values for *height*, we calculate the minimum, mean and range of these n values.

Since CAA only operates on one cluster at a time (see section 4.1) we remove features that are null over all ads within that cluster. This reduces the dimentionality of the sparse feature matrix.

For sLDA, we assign one bit to each categorical feature rather than each possible value of it. In this case the feature takes the dominant value, i.e. the value of most counts. Additionally, we mapped all numeric features to categorical features. Instead of calculating the exact minimum, mean, and range of numeric features, we recorded the level (none, below 25%, 25% to 50%, 50% to 75% and above 75%) to which the minimum, mean, and range belongs.

3 Literature Review

3.1 Canonical Correlation Analysis

Canonical Correlation Analysis is a statistical method useful for exploring relationships between two sets of variables. It is used in machine learning, with appplications to multiple domains. Previous applications to medicine, biology and finance include [4], [12] and [13]. In [7] the algorithm is explained in detail, together with some applications to learning methods. In particular, they analyze Kernel CCA, which provides an alternative in cases where the linearity of CCA does not suit a problem. Other relevant references for both the theoretical developments and the applications of CCA include [11], [14] and [5].

A modified version of the algorithm that will be particularly useful for this research was proposed by Witten et al. [13]. Sparse CCA, an L_1 variant of the original CCA, adds constraints to guarantee sparse solutions. This limits the number of features being correlated. Their formulation of the maximization problem also differs from the traditional CCA algorithm. We will use this version since it is more suitable for our needs.

The work that most resembles our research is [6]. Using the notion of *autocorrelation*, they attempt to find underlying components of fMRI data that have maximum autocorrelation, and to do so they use CCA. The type of data they work with differs from ours in that their features are ordered (both temporally and spatially). For autocorrelation, they take X as the original data matrix and construct Y as a translated version of X, such that $Y_t = Y_{t+1}$. Since our data is not ordered we cannot follow the same procedure, and must instead develop a new autocorrelation technique.

3.2 Supervised Latent Dirichlet Allocation

Topic models provide upsupervised methods to learn themes in a collection of documents. These themes, or topics, are defined by words which occurr with some probability in the documents[1]. Latent Dirichelt Allocation (LDA) is a common technique which employs a heirarchical Bayesian model to analyze documents represented by a bag of words[2]. While words are observed variables, all other parameters are infered from data. In addition to discovering topics, LDA determines the distribution of topics over each document. Given its intuitive appraoch to topic modeling, LDA has spawned a range of extensions which are beyond the scope of the present research[1].

Supervised LDA (sLDA) is one variation of LDA which leverages labels, or response variables, associated with each document in the topic modeling. The response variable is incorporated as an observed variable for each document, drawn from a Bayesian prior[9]. Topics can be learned through variational expectation maximization as detailed in [9]. Alternatively, Gibbs sampling can be used to learn topics[3], a technique we employ in section 4.2. Additionally, extensions of sLDA have incorporated different priors, such as Dirichlet-Multinomial priors over the response varaible[10].

4 Methodology

4.1 Canonical Autocorrelation Analysis

Given two matrices X and Y, CCA aims to find linear combinations of its columns that maximize the correlation between them. Usually, X and Y are two matrix representations for one set of datapoints, each matrix using a different set of variables to describe the same datapoints.

We use the formulation of CCA given by [13]. Assuming X and Y have been standarized and centered, the constrained optimization problem is:

$$max_{u,v} u^T X^T Y v \quad ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad ||u||_1 \leq c_1, ||v||_1 \leq c_2 \tag{1}$$

When c_1 and c_2 are small, solutions will be sparse, and thus only a few number of features are correlated.

Our goal is to find correlations within the same set of variables. Therefore, both our matrices X and Y are identical. Currently, none of the variants of CCA we were able to find is suited to do this, and applying Sparse CCA when $X = Y$ results in solutions $u = v$. We develop a modified version of the algorithm capable of finding such autocorrelations by imposing an additional constraint on equation 1 to prevent the model from correlating each variable with itself. Using the Lagrangian form, the problem can be written as follows:

$$max_{u,v} u^T X^T X v - \lambda u^T v \quad ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad ||u||_1 \leq c_1, ||v||_1 \leq c_2 \tag{2}$$

This will penalize vectors u and v for having high values for the same entry, which is precisely what we are trying to avoid. With proper factorization, this can be turned into a Sparse CCA maximization problem. First, notice that $u^T X^T X v - \lambda u^T v = u^T (X^T X - \lambda I) v$. Therefore, the Canonical Autocorrelation Analysis problem can be written as:

$$max_{u,v} u^T (X^T X - \lambda I) v \quad ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad ||u||_1 \leq c_1, ||v||_1 \leq c_2 \tag{3}$$

Finding the singular value decomposition (SVD) of X, we have:

$$\begin{array}{rrcl} & X & = & USV^T \\ \Rightarrow & X^T X & = & VSV^T \\ \Rightarrow & X^T X - \lambda I & = & VSV^T - \lambda VV^T = \quad V(S - \lambda I)V^T \end{array}$$

Now, setting $X_{new} = [V(S - \lambda I)]^T$ and $Y_{new} = V^T$, the problem becomes:

$$max_{u,v} u^T X_{new}^T Y_{new} v \quad ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad ||u||_1 \leq c_1, ||v||_1 \leq c_2 \tag{4}$$

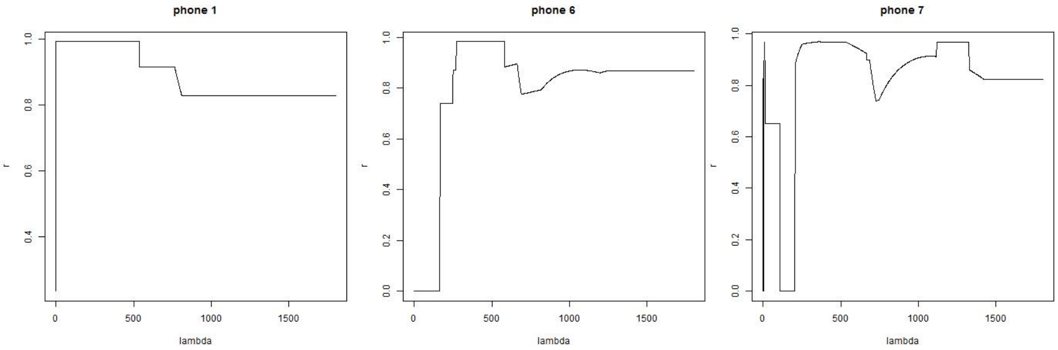

Figure 1: Relative sparseness vs. penalty parameter

This problem can be solved using Sparse CCA, and since the solutions obtained with this method are independent of the factorization of $X_{new}^T Y_{new}$, solving this is equivalent to solving the CAA maximization problem.

This is not convex on λ and is very sensitive to variations of this parameter. For choosing the proper λ, an evaluation metric for relative sparseness was developed. Vectors u and v are considered to have relative sparseness if entries with high values in u do not coincide with entries with high values in v. This can be measured with $r(u, v) = 1 - \sum_i |u_i v_i|$.

Figure 1 shows some examples of how the relative sparseness varies with respect to λ for the three largest clusters in the dataset. The practitioner can then determine the minimum acceptable relative sparseness and the optimal λ is found through a grid search.

Once the desired relative sparseness, r_0, is determined, the grid search will return the smallest λ for which the solutions u and v satisfy $r(u, v) \geq r_0$. The smallest λ is considered the best option because as λ increases the algorithm gives up correlation in order to have a lower penalty cost, and we are interested in the maximum correlation within our relative sparseness constraint.

4.2 sLDA

As described in Section 3.2, topic modeling provides another means of characterizing the human trafficking clusters[2]. We focus on an sLDA model illustrated in Figure 2 [9]. The generative process of this graphical model is specified by:

1. Draw topic proportions $\theta_d | \alpha \sim$ Dirichlet(α)
2. Draw word proportions $\phi_k | \eta \sim$ Dirichlet(η)
3. For each word w in document d:
   (a) Draw topic assignment $z | \theta_d \sim$ Multinomial(θ_d)
   (b) Draw word $w | z, \phi_z \sim$ Multinomial(ϕ_z).
4. Draw cluster label $y_d | \bar{z}_d \sim$ Normal($b^T \bar{z}_d + a, \sigma^2$)

The sLDA model is modified for the context of the present reserach by considering each advertisement to be a document. Each feature extracted from the ads is considered a word. As described in section 2, we encode the numeric features discretely. This dramatically reduces the number of words, making the learning process feasible, allowing documents to share words indicating intuitive feature ranges. In sLDA each document has an associated observed response variable which is sampled from a gaussian linear distribution with its center decided by the current topic distribution of the document. Here, the response variable is the phone number, or equivilently, the cluster to which the ad belongs. Topics retain the equivilent interpretation as under standard sLDA applications: they represent ordered collections of features which describe documents.

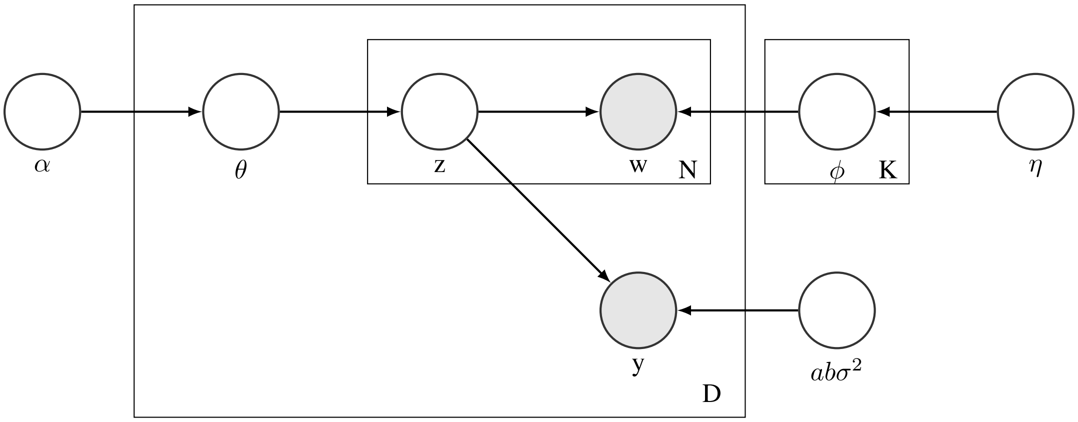

Figure 2: sLDA graphic model

Gibbs Sampling is used to learn the ad-topic distribution θ and topic-word distribution ϕ. The conditional distribution used to sample topic is defined by Equation 5

$$\begin{aligned} p\left(z_{d,n} = k|\alpha, \eta, w, y, b, a, z_{(d,n)}\right) \propto \left(n_{d,k}^{\neg d,n} + \alpha\right) \frac{n_{w_{d,k},k}^{\neg d,n} + \eta_{w_{d,k}}}{N_k^{\neg d,n} + w\eta} \\ exp\left(\frac{2b_k}{N_d}\left(y_d - a - b^T z_d^{\neg n}\right) - \left(\frac{b_k}{N_d}\right)^2\right) \end{aligned} \tag{5}$$

An obvious disadvantage of this model inference approach is that the origninal dense implementation has a high order of complexity up to $O(\#Rounds \times \#Docs \times \#Words \times \#Topics)$. However, to guarantee precision, we beleive this implementation is warranted.

5 Experimental Results

5.1 Canonical Autocorrelation Analysis

Canonical Autocorrelation Analysis was applied to the three largest clusters in the data. Three important conclusions can be drawn from the results: *i.* The method succesfully finds solutions with relative sparseness. *ii.* Good linear correlations between the projections are obtained. *iii.* The results contain information which is potentially useful for law enforcement agents.

The threshold was set to $r = 0.7$. Figures 3 and 4 show scatterplots with $u^T X^T$ in the x-axis and Xv in the y-axis. For phone number 1, which has 429 ads related to it, a strong positive correlation ($r^2 = 0.99$) between restriction of chinese men and restriction based on age was found. The correlation coefficient vectors are: $u = [0, 0.99, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.08, 0, 0]$, $v = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0.09, 0.99, 0, 0]$, where the third entry corresponds to restriction of chinese men and the two later ones correspond to restrictions by age. For phone number 6, which has 1815 associated ads, a weaker correlation was found ($r^2 = 0.56$), this time between phone numbers with Florida Area Code, and the posts being in Illinois. In this case, the correlation coefficient vectors are $u = [-0.13, 0, 0, 0, 0, -0.99, 0, 0, 0, 0, 0, 0, 0, 0]$, $v = [-0.99, 0, 0, 0, 0, -0.13, 0, 0, 0, 0, 0, 0, 0, 0]$. The third cluster analyzed, phone number 7, showed a strong correlation ($r^2 = 0.94$), in which three features characterize the cluster: phone numbers with Florida Area Code, those with Rhode Island Area Code, and the fact that posts are/are not in Illinois, with $u = [-0.18, 0, -0.98, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$, $v = [-0.98, 0, 0, 0, -0.18, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$. In this cluster, the first coefficient corresponds to FL Area Code, the third one to RI Area Code, and the fifth one to ads being posted in Illinois (remember from the preprocessing, features vary from one cluster to the other).

The similarity between the characterization of the last two clusters motivated a manual inspection to determine whether the phone numbers were linked. We found that indeed the phone numbers

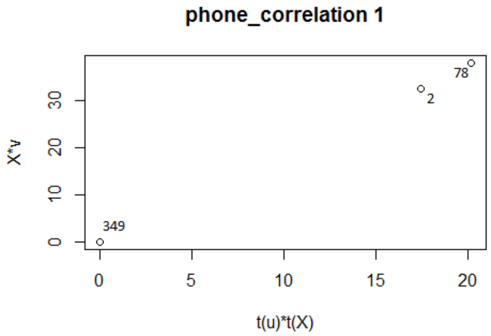

Figure 3: Correlation between projections for phone 1. Numbers next to the points indicate the number of ads being mapped to that point.

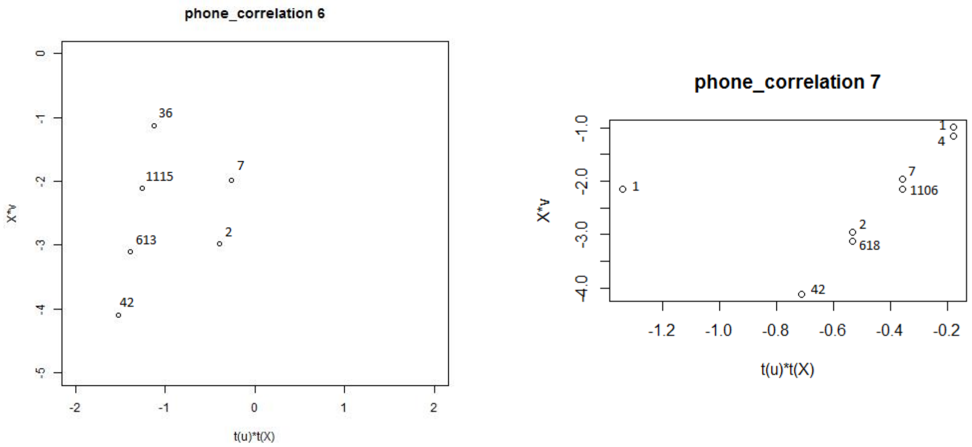

Figure 4: Correlation between projections for phone 6 and phone 7, respectively. Numbers next to the points indicate the number of ads being mapped to that point.

co-occur in an ad. Since such explicit links are not always available to law enforcement agents, establishing relations between phone numbers through data analysis allows law enforcement to connect different cases and understand the structure of a trafficking ring. This represents a potential use case for CAA that we plan to study further.

Ideally, it would have been desirable to obtain information on correlations between a larger subset of features. Since each cluster only has approximately 15 non-null features and relative sparseness is enforced, correlations are only obtained between two or three features. Were clusters to have more features, this would not be a problem. However, it begs the question whether parameters can be tuned to gain more insight with fewer features, or if there is a minimum (or range) number of features for which the method works best.

5.2 sLDA

An sLDA model was learned via Gibbs Sampling using both positive and negative clusters as input data. We used 188 positive clusters containing $9,595$ positive ads in total. We used the 20 largest negative clusters containing $23,829$ negative ads. This provides a sufficient mix of positive and negative ads within the limitations of our computing power. As the words in the sLDA model, we used 264 features, extracted as described in Section 2.

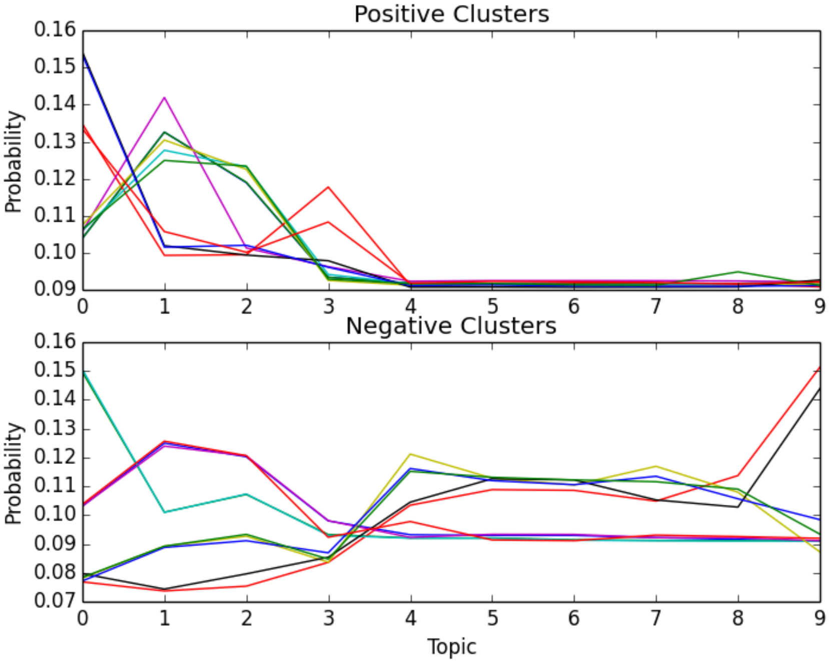

Figure 5: Distribution of clusters for each topic

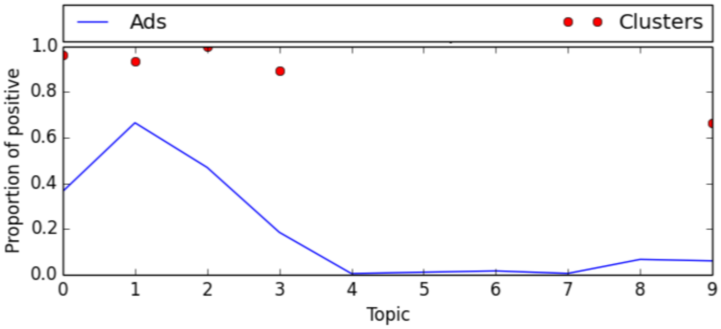

Figure 6: Proportion of positive ads and clusters for each topic

We set the number of topics, K, to 10, and set the hyperparameters $\alpha = 50.0/K, \beta = 0.001, a = 1, b = linspace(0, 1, K)$. 250 iterations were used to ensure mixing. For the cluster labels, y, we arbitrarily mapped the phone numbers to the $[0, 1]$ interval.

5.2.1 Topic Distributions

Figure 5 shows the topic distributions of the ten largest positive clusters and ten largest negative clusters. We calculate topic distribution of each cluster by summing over the individual topic distributions of each ad in the cluster and then normalizing by dividing by the size of the cluster. From Figure 5 it is clear that the positive clusters are dominated by topics 0, 1, and 2 while negative clusters are more uniformly distributed. Note that topic numbers are an arbitrary convention. Additionally, note that the two clusters previously linked by CAA have identical topic distributions (blue and green lines).

Another perspective on this distribution can be seen in Figure 6. The blue line represents the number of positive ads dominated by each topic divided by the total number of ads dominated by the topic. An ad, d, is dominated by a topic, T, if $\theta_{d,T} = max_t(\theta_{d,t})$. The red dots represent the number of

HT Topic 1	HT Topic 2	Negative Topic 1	Negative Topic 2
(Ethnicity,latin)	**(Ethnicity,asian)**	**(Ethnicity,asian)**	**(Ethnicity,korean)**
(Ethnicity,asian)	**(Ethnicity,latina)**	**(Ethnicity,korean)**	**(Ethnicity,taiwanese)**
(Ethnicity,latina)	**(Ethnicity,colombian)**	**(Ethnicity,latino)**	**(Ethnicity,latino)**
(Ethnicity,thai)	**(Ethnicity,dominican)**	**(Ethnicity,foreign)**	**(Ethnicity,african american)**
(Ethnicity,persian)	**(Ethnicity,persian)**	**(Ethnicity,japanese)**	**(Ethnicity,asian)**
(Ethnicity,cuban)	**(Ethnicity,thai)**	***(Age_range,x<8)***	**(Ethnicity,chinese)**
(Ethnicity,colombian)	**(Ethnicity,latin)**	***(Age_mean,x<22.6)***	***(Age_min,x<22)***
(Ethnicity,mexican)	**(Ethnicity,brazilian)**	***(Age_min,x<22)***	***(Age_min,22<x<39)***
(Ethnicity,puerto rican)	**(Ethnicity,cuban)**	***(Age_min,22<x<39)***	***(Age_range,x<8)***
(Ethnicity,american)	**(Ethnicity,czech)**	***(Age_mean,x>45)***	***(Age_mean,x<22)***
(Ethnicity,welsh)	**(Ethnicity,welsh)**	***(Age_mean,22.6<x<30)***	**(Chest_mean,33<x<39)**
(Ethnicity,peruvian)	**(Ethnicity,african)**	**(Chest_mean,33<x<39)**	**(Chest_min,32<x<44)**
(Chest_min,x>56)	**(Ethnicity,african american)**	**(Chest_min,32<x<44)**	(Weight_mean,x<115)
(Chest_mean,x>48)	**(Ethnicity,chinese)**	(Weight_mean,x<115)	(Height_mean,4.6<x<5.2)
(HairColor,auburn)	***(Age_min,x<22)***	(Hip_range,x<5)	(Chest_range,x<8)
(Restriction,african american)	***(Age_min,39<x<55)***	(Weight_range,x<27)	(Hip_mean,33.3<x<40)
(SkinColor,brown)	(Height_ft_min,5<x<6)	(EyeColor,dark)	(Waist_min,x<28)
(Perspective_1st_min,x<14)	(Perspective_1st,14<x<28)	(Chest_range,x<8)	(HairColor,brown)
(Perspective_3rd_min,x<5)	(Waist_range < 6)	(Hip_min,32.0<x<44)	(SkinColor,dark)
(Chest_mean,x<33.25)	(HairColor,strawberry blonde)	(Waist_range,x<6)	(SkinColor,caucasian)
(Height_ft_min,6<x<7)	(Height_ft_mean,4.6<x<5.2)	(Waist_mean,x<25.5)	(EyeColor,brown)
(Height_in_min,3<x<6)	(HairColor,auburn)	(HairColor,dark brown)	(EyeColor,honey)

Figure 7: Top features for 4 representative topics (not ordered)

positive clusters with ads dominated by each topic divided by the total number of clusters with ads dominated by the topic. Note that topics 4 through 8 dominated too few positive or negative clusters to yield an accurate calculation. As before, Figure 6 illustrates that topics 0, 1, and 2 dominate positive ads and clusters.

5.2.2 Significant Features

Here we qualitatively analyze the hidden topics to show how they may provide descriptive characterizations of human traffficking. We extracted the top 22 non-State features with highest probabilities in two topics dominated by positive ads and two topics dominated by negative ads. Figure7 shows these features grouped by feature type to facilitate interpretation.

The features marked in pink show that human trafficking ads focus more on ethnicities than the negative topics, especially exotic ethnicities such as Thai, Persian, and Cuban. Features marked in red show that human trafficking ads are likely to exaggerate the body features of women advertised. Finally, features marked in green indicate that human trafficking ads are unlikely to disclose the ages of women in advertisements. These characterizations are natural for human traffickers who tend to traffick women from foreign countries, many of whom are minors.

6 Conclusion

We analyzed a large corpus of online human trafficking and prostitution ads in order to obtain descriptive characterizations of how human trafficking rings and individuals operate online. Two methods were employed in order to provide a variety of perspectives on the data. CAA characterized individual trafficking clusters and provided results which enabled us to identify certain clusters that were particularly closely related. sLDA found topics which described behaviors across various human trafficking clusters. Analysis of the hidden topics revealed sets of features that were particularly relevant for understanding human trafficking ads. Together, we hope these methods can help law enforcement better understand online human trafficking behaviors and how machine learning techniques can aid in their daily operations.

Acknowledgments

We would like to acknowledge Dr. Artur Dubrawski, Dr. Daniel Neill, Kirthevasan Kandasamy, Willie Neiswanger and Benedikt Boecking for their advice and much appreciated assistance.

References

[1] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Jonathan Chang. Derivation of gibbs sampling equations.

[4] Wim De Clercq, Anneleen Vergult, Bart Vanrumste, Wim Van Paesschen, and Sabine Van Huffel. Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. *Biomedical Engineering, IEEE Transactions on*, 53(12):2583–2587, 2006.

[5] Rene Donner, Michael Reiter, Georg Langs, Philipp Peloschek, and Horst Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1690, 2006.

[6] Ola Friman, Magnus Borga, Peter Lundberg, and Hans Knutsson. Exploratory fmri analysis by autocorrelation maximization. *NeuroImage*, 16(2):454–464, 2002.

[7] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[8] Frank Laczko. Data and research on human trafficking. *International Migration*, 43(1-2):5–16, 2005.

[9] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[10] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

[11] Liang Sun, Shuiwang Ji, and Jieping Ye. Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):194–200, 2011.

[12] Koby Todros and AO Hero. Measure transformed canonical correlation analysis with application to financial data. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, pages 361–364. IEEE, 2012.

[13] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[14] Yi Zhang and Jeff G Schneider. Multi-label output codes using canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 873–882, 2011.