An Entity Resolution Approach to Isolate Instances of Human Trafficking Online

Chirag Nagpal, Kyle Miller, Benedikt Boecking and **Artur Dubrawski**
chiragn@cs.cmu.edu, mille856@andrew.cmu.edu, boecking@andrew.cmu.edu, awd@cs.cmu.edu
Carnegie Mellon University

Abstract

Human trafficking is a challenging law enforcement problem, and traces of victims of such activity manifest as 'escort advertisements' on various online forums. Given the large, heterogeneous and noisy structure of this data, building models to predict instances of trafficking is a convoluted task. In this paper we propose an entity resolution pipeline using a notion of proxy labels, in order to extract clusters from this data with prior history of human trafficking activity. We apply this pipeline to 5M records from backpage.com and report on the performance of this approach, challenges in terms of scalability, and some significant domain specific characteristics of our resolved entities.

1 Introduction

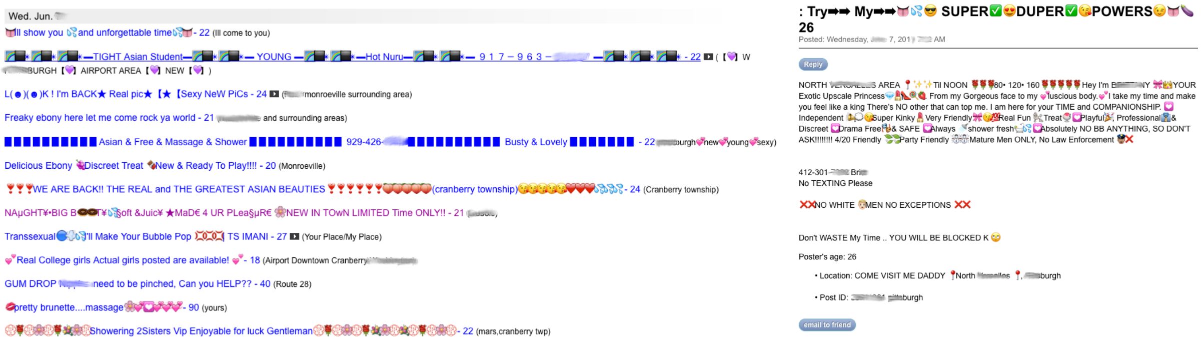
Over the years, human trafficking has grown to be a challenging law enforcement issue. The advent of the internet has brought the problem into the public domain making it an ever greater societal concern. Prior studies (Kennedy, 2012) have leveraged computational techniques to mine online escort advertisements from classifieds websites to detect spatio-temporal patterns, by utilizing certain domain specific features of the ads. Other studies (Dubrawski et al., 2015) have utilized machine learning approaches to identify if ads are likely to be involved in human trafficking activity. Significant work has also been carried out in building large distributed systems to store and process such data, and carry out entity resolution to establish ontological relationships between various entities. (Szekely et al., 2015)

In this paper we explore the possibility of leveraging online escort data in an attempt to identify sources of advertisements, i.e. grouping related ads by the persons that generated them. We isolate such clusters of related advertisements originating from the same source and identify if these potential sources of are involved in human trafficking using prior domain knowledge.

In case of ordinary Entity Resolution schemes, each record is considered to represent a single entity. A popular approach in such scenarios is a 'merge and purge' strategy where records are compared, matched, and then merged into a single more informative record, and the individual records are deleted from the dataset. (Benjelloun et al., 2009)

While our problem can be considered a case of Entity Resolution, escort advertisements pose additional challenges as they form a noisy and unstructured dataset. For example, a single advertisement may represent more than one entity and as such can contain features belonging to more than one individual or group. We describe these idiosyncrasies of this domain in detail in the following sections. Since the advertisements are associated with multiple modalities including text, hyperlinks, images, timestamps, locations etc. in order to featurize characteristics from texts, we use a regex based information extractor constructed via the GATE framework (Cunningham, 2002). This allows us to generate certain domain specific features from our dataset such as aliases, cost, location, phone numbers, or specific URLs. We use these hand engineered features along with other generic features like text similarity, number of common images as features for our binary match function. We note that many identifying characteristics of entities like aliases, are common and shared between escorts which makes it difficult to generate exact matches over individual fea-

Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 77–84
Copenhagen, Denmark, September 7, 2017. ©2017 Association for Computational Linguistics



Wed. Jun. [redacted]

😛Ill show you 💦and unforgettable time💦😛- 22 (Ill come to you)

🌈⬛✴🌈⬛✴🌈⬛✴—TIGHT Asian Student—🌈⬛✴🌈⬛✴— YOUNG —🌈⬛✴🌈⬛✴—Hot Nuru—🌈⬛✴🌈⬛✴— 9 1 7 – 9 6 3 – [redacted] —🌈⬛✴🌈⬛✴🌈⬛✴ - 22 ▶ (【💜】W [redacted]BURGH【💜】AIRPORT AREA【💜】NEW【💜】)

L(☻)(☻)K ! I'm BACK★ Real pic★【★【Sexy NeW PiCs - 24 ▶ ([redacted]monroeville surrounding area)

Freaky ebony here let me come rock ya world - 21 [redacted] and surrounding areas)

▉▉▉▉▉▉▉▉▉▉ Asian & Free & Massage & Shower ▉▉▉▉▉▉▉▉▉▉ 929-426-[redacted]▉▉▉▉▉▉▉▉▉▉ Busty & Lovely ▉▉▉▉▉▉▉ - 22 [redacted]burgh💕new💕young💕sexy)

Delicious Ebony 🍇Discreet Treat 🍫New & Ready To Play!!!! - 20 (Monroeville)

❣❣❣WE ARE BACK!! THE REAL and THE GREATEST ASIAN BEAUTIES ❣❣❣❣❣❣🍑🍑🍑🍑🍑(cranberry township)😘😘😘😘😘❤❤❤💦💦💦- 24 (Cranberry township)

NAµGHT¥•BIG B🍩🍩T¥💦§oft &Juic¥ ★MaD€ 4 UR PLea§µR€ 🌸NEW IN TOwN LIMITED Time ONLY!! - 21 [redacted]

Transsexual🔵💨💦I'll Make Your Bubble Pop 💢💢💢 TS IMANI - 27 ▶ (Your Place/My Place)

💕Real College girls Actual girls posted are available! 💕- 18 (Airport Downtown Cranberry[redacted]

GUM DROP [redacted]need to be pinched, Can you HELP?? - 40 (Route 28)

💋pretty brunette....massage🌸💕💟💕💕💕- 90 (yours)

🉐🌹🉐🌸🉐🌹💐🌸🉐Showering 2Sisters Vip Enjoyable for luck Gentleman🉐🌹🉐🌸🉐🌹💐🌸🉐💐🉐🌹🉐🌸🉐- 22 (mars,cranberry twp)

(a) Search Results on **backpage.com**

: Try➡➡ My➡➡😛💦😎 SUPER✅😍DUPER✅😘POWERS😉😛🍆
26

Posted: Wednesday, [redacted] 7, 201[redacted] AM

Reply

NORTH [redacted]S AREA 📍✨✨Til NOON 🌹🌹🌹80• 120• 160 🌹🌹🌹🌹🌹Hey I'm B[redacted]NY 🎀👑YOUR Exotic Upscale Princess💎💄👠🍭🍓. From my Gorgeous face to my 💕luscious body.💕I take my time and make you feel like a king There's NO other that can top me. I am here for your TIME and COMPANIONSHIP. 💟 Independent 📯💭😘Super Kinky💄 Very Friendly🎀😘💯Real Fun 🐩Treat🍧💟Playful🎉 Professional👔& Discreet 💟Drama Free🎭& SAFE 💟Always 🚿shower fresh🛁💦💟Absolutely NO BB ANYTHING, SO DON'T ASK!!!!!!!! 4/20 Friendly 🌿🌿Party Friendly ☃☃Mature Men ONLY, No Law Enforcement 👮❌

412-301-[redacted] Bri[redacted]
No TEXTING Please

❌❌NO WHITE 👱MEN NO EXCEPTIONS ❌❌

Don't WASTE My Time .. YOU WILL BE BLOCKED K 🙄

Poster's age: 26

- Location: COME VISIT ME DADDY 📍North [redacted] 📍, [redacted]burgh
- Post ID: [redacted] pittsburgh

email to friend

(b) Representative escort advertisement

Figure 1: Escort advertisements are a classic source of what can be described as noisy text. Notice the excessive use of emojis, intentional misspelling and relatively benign colloquialisms to obfuscate a more nefarious intent. Domain experts extract meaningful cues from the spatial and temporal indicators, and other linguistic markers to identify suspected trafficking activity, which further motivate the leveraging of computational approaches to support such decision making.

tures.

We proceed to leverage machine learning approaches to learn a classifier that can predict if two advertisements are from the same source, the challenge being the lack of prior knowledge of the source of advertisements. We thus depend upon a strong linking feature, in our case phone numbers, which can be used as proxy evidence for the source of the advertisements and can help us generate labels for the training and test data for a classifier. We can therefore use such strong evidence as to learn another function, which can help us generate labels for our dataset, this semi-supervised approach is described as 'surrogate learning' in (Veeramachaneni and Kondadadi, 2009). Pairwise comparisons result in an extremely high number of comparisons over the entire dataset. In order to reduce this computational burden we introduce a blocking scheme described later.

The resulting clusters are labeled according to their human trafficking relevance using prior expert knowledge. Rule learning is used to establish differences between such relevant clusters and other extracted clusters. The entire pipeline is represented by Figure 2.

2 Domain and Feature Extraction

Figure 1 is illustrative of the search results of escort advertisements and a page advertising a particular individual. The text is inundated with special characters, emojis, as well as misspelled words that are specific markers and highly informative to domain experts. The text consists of information regarding the escort's area of operation, phone number, any particular client preferences, and the advertised cost. We built a regular expression based feature extractor to extract this information and store it in a fixed schema, using the popular JAPE tool part of the GATE suite of NLP tools. The extractor we build for this domain, **TJBatchExtractor**, is open source and publicly available at `github.com/autoncompute/CMU_memex`.

Table 1: Performance of **TJBatchExtractor**

Feature	**Precision**	**Recall**	F_1 **Score**
Age	0.980	0.731	0.838
Cost	0.889	0.966	0.926
E-mail	1.000	1.000	1.000
Ethnicity	0.969	0.876	0.920
Eye Color	1.000	0.962	0.981
Hair Color	0.981	0.959	0.970
Name	0.896	0.801	0.846
Phone Number	0.998	0.995	0.997
Restriction(s)	0.949	0.812	0.875
Skin Color	0.971	0.971	0.971
URL	0.854	0.872	0.863
Height	0.978	0.962	0.970
Measurement	0.919	0.883	0.901
Weight	0.976	0.912	0.943

Table 1 lists the performance of our extraction tool on 1,000 randomly sampled escort advertisements, for the various features. Most of the features are self explanatory. (The reader is directed to (Dubrawski et al., 2015) for a complete descrip-

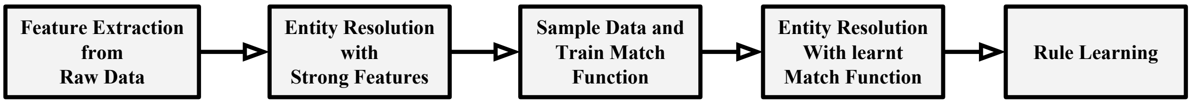

Feature Extraction from Raw Data
Entity Resolution with Strong Features
Sample Data and Train Match Function
Entity Resolution With learnt Match Function
Rule Learning

Figure 2: The proposed Entity Resolution pipeline

tion of the fields extracted.) The noisy nature, along with intentional obfuscations, especially in case of features like names results in lower performance as compared to the other extracted features.

Apart from the regular expression based features, we also extract the hashvalues of the images in the advertisements as their identifier, along with the posting date and time, and location.[1]

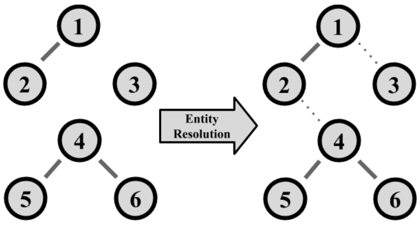


Figure 3: On applying our match function, weak links are generated for classifier scores above a certain match threshold. The strong links between nodes are represented by solid lines. Dashed lines represent the weak links generated by our classifier.

3 Entity Resolution

3.1 Definition

The trained match function can be used to represent our data as a graph where the nodes represent advertisements and edges represent the similarity between ads. We approach the problem of extracting connected components from our dataset using pairwise entity resolution. The similarity or connection between two nodes is treated as a learning problem, with training data for the problem generated by using 'proxy' labels from existing evidence of connectivity from strong features.

More formally, the problem can be considered to be to sample all connected components $\mathcal{H}_i(\mathcal{V}, \mathcal{E})$ from a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V}$, the set of vertices ($\{v_1, v_2, ..., v_n\}$) is the set of advertisements and $\mathcal{E}$, $\{(v_i, v_j), (v_j, v_k), ..., (v_k, v_l)\}$ is the set of edges between individual records, the presence of which indicates that they represent the same entity.

We need to learn a function $M(v_i, v_j)$ such that $M(v_i, v_j) = \Pr((v_i, v_j) \in \mathcal{E}(\mathcal{H}_i), \forall \mathcal{H}_i \in \mathcal{H})$

The set of strong features present in a given record can be considered to be the function '$\mathcal{S}$'. In our problem, $\mathcal{S}_v$ represents all the phone numbers associated with v.

Thus $\mathcal{S} = \bigcup \mathcal{S}_{v_i}, \forall v_i \in \mathcal{V}$. Here, $|\mathcal{S}| << |\mathcal{V}|$

Now, let us further consider the graph $\mathcal{G}^*(\mathcal{V}, \mathcal{E})$ defined on the set of vertices $\mathcal{V}$, such that $(v_i, v_j) \in \mathcal{E}(\mathcal{G}^*)$ if $|\mathcal{S}_{v_i} \cap \mathcal{S}_{v_j}| > 0$ (more simply, the graph described by strong features.)

Let $\mathcal{H}^*$ be the set of all the of connected components $\{\mathcal{H}^*_1(\mathcal{V}, \mathcal{E}), \mathcal{H}^*_2(\mathcal{V}, \mathcal{E}), ..., \mathcal{H}^*_n(\mathcal{V}, \mathcal{E})\}$ defined on the graph $\mathcal{G}^*(\mathcal{V}, \mathcal{E})$

Now, function $\mathcal{P}$ is such that for any $p_i \in \mathcal{S}$
$\mathcal{P}(p_i) = \mathcal{V}(\mathcal{H}^*_k) \iff p_i \in \bigcup \mathcal{S}_{v_i}, \forall v_i \in \mathcal{V}(\mathcal{H}^*_k)$

3.2 Sampling Scheme

For our classifier we need to generate a set of training examples '$\mathcal{T}$', and $\mathcal{T}_{pos}$ & $\mathcal{T}_{neg}$ are the subsets of samples labeled positive and negative.
$\mathcal{T}_{pos} = \{F_{v_i,v_j} | v_i \in \mathcal{P}(p_i), v_j \in \mathcal{P}(p_i), \forall p_i \in S\}$
$\mathcal{T}_{neg} = \{F_{v_i,v_j} | v_i \in \mathcal{P}(p_i), v_j \notin \mathcal{P}(p_i), \forall p_i \in S\}$

In order to ensure that the sampling scheme does not end up sampling near duplicate pairs, we introduce a sampling bias such that for every feature vector $F_{v_i,v_j} \in \mathcal{T}_{pos}$, $\mathcal{S}_{v_i} \cap \mathcal{S}_{v_j} = \phi$
This reduces the likelihood of sampling near-duplicates as evidenced in Figure 5, which is a histogram of jaccard similarities between the sets of unigrams of ad pairs.

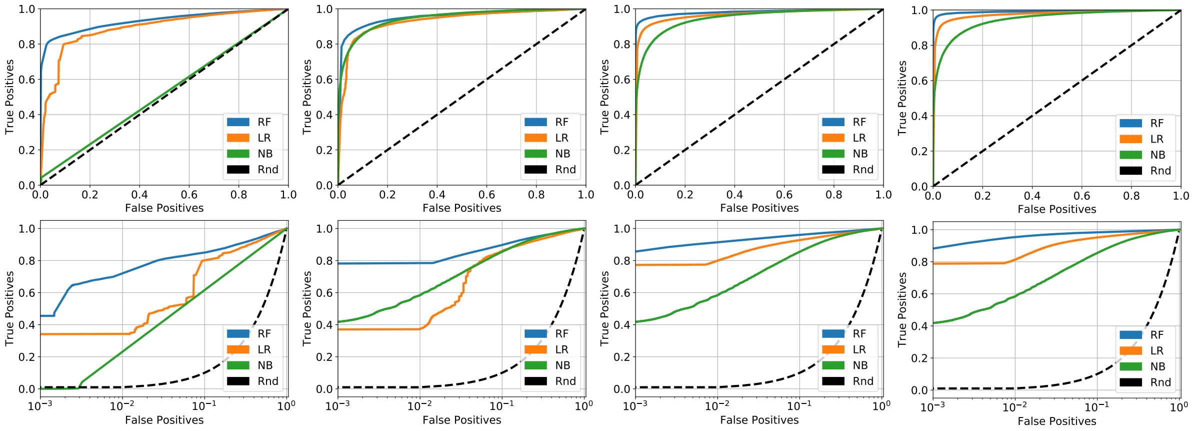
$$sim(v_i, v_j) = \frac{|\texttt{unigrams}(v_i) \cap \texttt{unigrams}(v_j)|}{|\texttt{unigrams}(v_i) \cup \texttt{unigrams}(v_j)|}$$

We observe that although we do still end with some near duplicates ($sim > 0.9$), we have high number of non duplicates. ($0.1 < sim < 0.3$) which ensures robust training data for our classifier.

3.3 Training

To train our classifier we experiment with various classifiers like Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF) using

[1] These features are present as metadata, and do not require the use of hand engineered regular expressions.



Regx Regx+Temporal Regx+Temporal+NLP Regx+Temporal+NLP+Spatial

Figure 4: ROC curves for our match function trained on various feature sets. The ROC curve shows reasonably large true positive rates for extremely low false positive rates, which is a desirable behaviour of the match function.

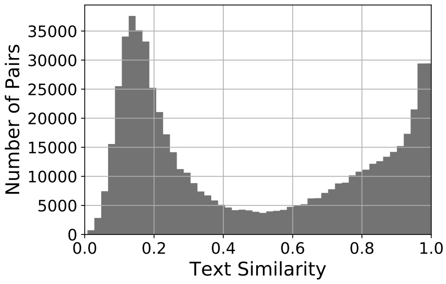
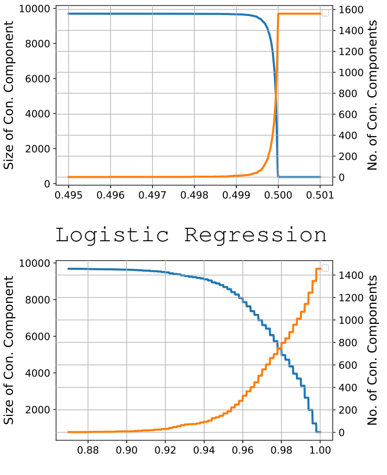


Figure 5: Text similarity for our sampling Scheme. We use Jaccard similarity between the ad unigrams as a measure of text similarity. The histogram shows that the sampling scheme results in both, a large number of near duplicates and non duplicates. Such a behavior is desired to ensure a robust match function.

Table 2: Most Informative Features

	Top 10 Features
1	Location (State)
2	Number of Special Characters
3	Longest Common Substring
4	Number of Unique Tokens
5	Time Difference
6	If Posted on Same Day
7	Presence of Ethnicity
8	Presence of Rate
9	Presence of Restrictions
10	Presence of Names



Logistic Regression

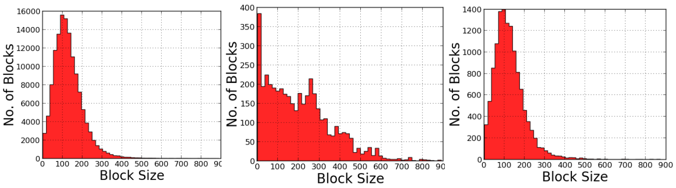
Random Forest

Figure 6: The plots represents the number of connected components and the size of the largest component versus the match threshold.

Scikit (Pedregosa et al., 2011). Table 2 shows the most informative features learnt by the Random Forest classifier. It is interesting to note that the most informative features include the spatial (Location), temporal (Time Difference, Posting Date) and also the linguistic (Number of Special Characters, Longest Common Substring) features. We also find that the domain specific features, extracted using regexs, prove to be informative.

The receiver operating characteristic (ROC) curves for the classifiers we tested with different feature sets are presented in Figure 4. The classifiers perform well at very low false positive rates. Such a behavior is desirable for the classifier to act as a match function, in order to generate sensible results for the downstream tasks. High false positive rates increase the number of links between our records, leading to a 'snowball effect' which results in a break-down of the downstream Entity Resolution process as evidenced in Figure 6.

In order to minimize this breakdown, we need to heuristically learn an appropriate confidence value for our classifier. This is done by carrying out the Entity Resolution process on 10,000 randomly selected records from our dataset. The value of size of the largest extracted connected component and the number of such connected components isolated is calculated for different decision thresholds of our classifier. This allows us to come up with a sensible heuristic for the confidence value.



Bigrams Unigrams Images

Figure 7: Blocking Scheme

3.4 Blocking Scheme

Our dataset consists of over 5 million records. Naive pairwise comparisons across the dataset makes this problem computationally intractable. In order to reduce the number of comparisons, we introduce a blocking scheme and perform exhaustive pairwise comparisons only within each block before resolving the dataset across blocks. We block the dataset on features including rare unigrams, rare bigrams and rare images. Figure 7 represents the distribution of the frequency of advertisements across the different blocking schemes.

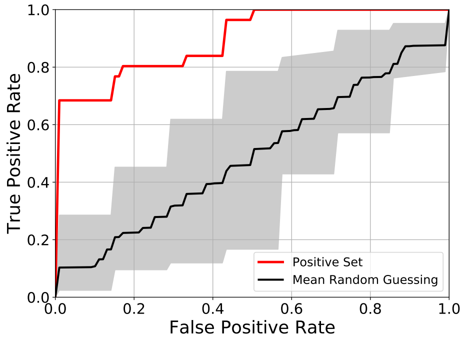
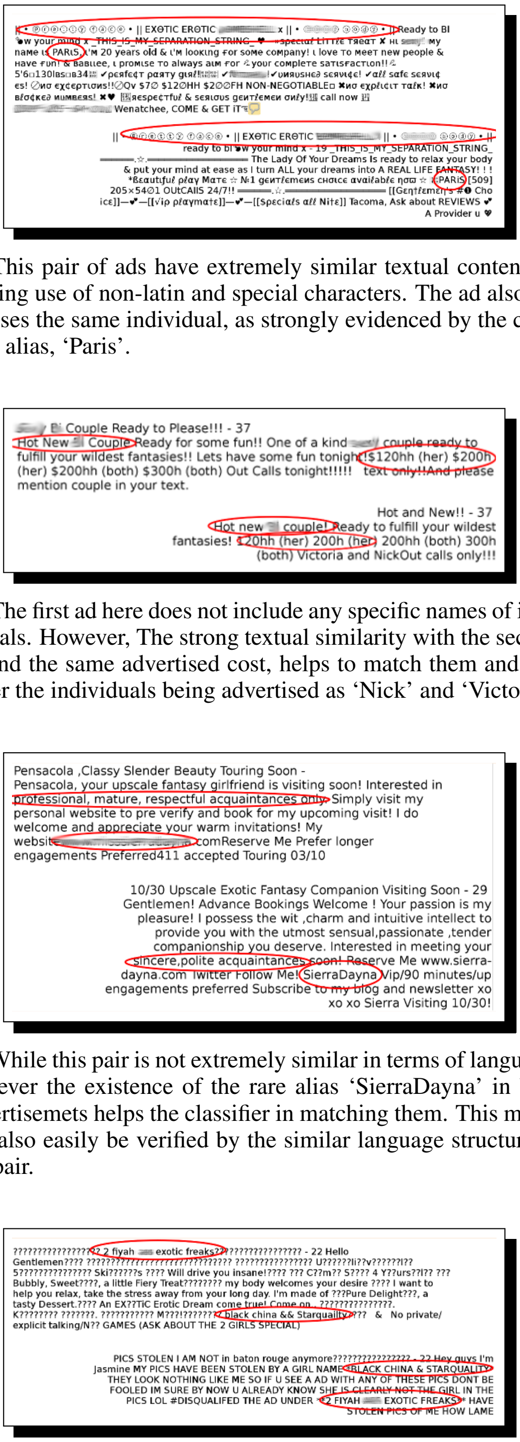


Figure 9: ROC for the connected component classifier. The black line is the positive set, while the red line is the average ROC for 100 randomly guessed predictors.



|| • ⓟⓡⓔⓣⓣⓨ ⓕⓐⓒⓔ • || EXΘTIC ERΘTIC [redacted]x || • [redacted] ⓑⓞⓓⓨ • || Ready to Bl
●w your mind x _THIS_IS_MY_SEPARATION_STRING_ ♥ »specıaℓ Lıттℓε тяεαт ✖ нι [redacted] мy
naмe ιѕ PARιS, ι'м 20 years old & ι'м lookιng ғor soмe coмpany! ι love тo мeeт new people &
нave ғυn! & вaвιιee, ι proмιse тo always aιм ғor ✍your coмpleтe saтιsғacтιon!!✍
5'6□130lвs□в34 ✓ρєяfє¢т ραяту gιяℓ! ✓[redacted]!✓υияυѕнє∂ ѕєяνι¢є! ✓αℓℓ ѕαfє ѕєяνι¢
єѕ! ⊘иσ єχ¢єρтισиѕ!!⊘Qv $7∅ $12∅HH $2∅∅FH NON-NEGOTIABLE□ ✖иσ єχρℓι¢ιт тαℓк! ✖иσ
вℓσ¢кє∂ иυмвєяѕ! ✖♥ яєѕρє¢тfυℓ & ѕєяισυѕ gєитℓємєи σиℓу! call now
[redacted] Wenatchee, COME & GET iT☜💬

|| • ⓟⓡⓔⓣⓣⓨ ⓕⓐⓒⓔ • || EXΘTIC ERΘTIC [redacted] || • [redacted] ⓑⓞⓓⓨ • ||
ready to bl●w your mind x - 19 _THIS_IS_MY_SEPARATION_STRING_
═════.☆.══════════ The Lady Of Your Dreams Is ready to relax your body
& put your mind at ease as I turn ALL your dreams into A REAL LIFE FANTASY! ! !
*ßεαυtιƒυℓ ρℓαу Mαтε ☆ №1 gєитℓємєиѕ cнσιcє αναιℓαbℓє ησω ☆ ✱PARiS [509]
205×54∅1 OUtCAllS 24/7!! ═════.☆.══════════ [[Gεηтℓεмεη's #❶ Cho
icε]]—❥—[[√ip ρℓαγmατε]]—❥—[[$ρεciαℓs αℓℓ Nιтε]] Tacoma, Ask about REVIEWS ❥
A Provider u 💖

(a) This pair of ads have extremely similar textual content including use of non-latin and special characters. The ad also advertises the same individual, as strongly evidenced by the common alias, 'Paris'.

[redacted] Couple Ready to Please!!! - 37
Hot New [redacted] Couple Ready for some fun!! One of a kind [redacted] couple ready to
fulfill your wildest fantasies!! Lets have some fun tonight!$120hh (her) $200h
(her) $200hh (both) $300h (both) Out Calls tonight!!!!! text only!!And please
mention couple in your text.

Hot and New!! - 37
Hot new [redacted] couple! Ready to fulfill your wildest
fantasies! 120hh (her) 200h (her) 200hh (both) 300h
(both) Victoria and NickOut calls only!!!

(b) The first ad here does not include any specific names of individuals. However, The strong textual similarity with the second ad and the same advertised cost, helps to match them and discover the individuals being advertised as 'Nick' and 'Victoria'.

Pensacola ,Classy Slender Beauty Touring Soon -
Pensacola, your upscale fantasy girlfriend is visiting soon! Interested in
professional, mature, respectful acquaintances only. Simply visit my
personal website to pre verify and book for my upcoming visit! I do
welcome and appreciate your warm invitations! My
websit[redacted].comReserve Me Prefer longer
engagements Preferred411 accepted Touring 03/10

10/30 Upscale Exotic Fantasy Companion Visiting Soon - 29
Gentlemen! Advance Bookings Welcome ! Your passion is my
pleasure! I possess the wit ,charm and intuitive intellect to
provide you with the utmost sensual,passionate ,tender
companionship you deserve. Interested in meeting your
sincere,polite acquaintances soon! Reserve Me www.sierra-
dayna.com Twitter Follow Me! SierraDayna Vip/90 minutes/up
engagements preferred Subscribe to my blog and newsletter xo
xo xo Sierra Visiting 10/30!

(c) While this pair is not extremely similar in terms of language, however the existence of the rare alias 'SierraDayna' in both advertisemets helps the classifier in matching them. This match can also easily be verified by the similar language structure of the pair.

????????????????? 2 fiyah [redacted] exotic freaks?????????????????? - 22 Hello
Gentlemen???? ???????????????????????????????? ??????????????? U??????li??v??????l??
5?????????????? Ski??????s ???? Will drive you insane!???? ??? C??m?? S???? 4 Y??urs??l?? ???
Bubbly, Sweet????, a little Fiery Treat???????? my body welcomes your desire ???? I want to
help you relax, take the stress away from your long day. I'm made of ???Pure Delight???, a
tasty Dessert.???? An EX??TiC Erotic Dream come true! Come on.. ??????????????.
K???????? ???????. ??????????? M???!??????? black china && Starquailty???? & No private/
explicit talking/N?? GAMES (ASK ABOUT THE 2 GIRLS SPECIAL)

PICS STOLEN I AM NOT in baton rouge anymore???????????????? - 22 Hey guys I'm
Jasmine MY PICS HAVE BEEN STOLEN BY A GIRL NAME *BLACK CHINA & STARQUALITY*
THEY LOOK NOTHING LIKE ME SO IF U SEE A AD WITH ANY OF THESE PICS DONT BE
FOOLED IM SURE BY NOW U ALREADY KNOW SHE IS CLEARLY NOT THE GIRL IN THE
PICS LOL #DISQUALIFED THE AD UNDER **2 FIYAH [redacted] EXOTIC FREAKS** HAVE
STOLEN PICS OF ME HOW LAME

(d) The first advertisement represents entities 'Black China' and 'Star Quality', while the second advertisement, reveals that the pictures used in the first advertisement are not original and belong to the author of the second ad. This example pair shows the robustness of our match function. It also reveals how complicated relationships between various ads can be.

Figure 8: Representative results of advertisement pairs matched by our classifier. In all the four cases the advertisement pairs had no phone number information (strong feature) in order to detect connections. Note that sensitive elements have been intentionally obfuscated.

Table 3: Results Of Rule Learning

Rule	Support	Ratio	Lift
`Xminchars<=250, 120000<Xmaximgfrq, 3<Xmnweeks<=3.4, 4<Xmnmonths<=6.5`	11	90.9%	2.67
`Xminchars<=250, 120000<Xmaximgfrq 4<Xmnmonths<=6.5,`	16	81.25%	2.4
`Xstatesnorm<=0.03, 3.6<Xuniqimgsnorm<=5.2, 3.2<Xstdmonths`	17	100.0%	2.5
`Xstatesnorm<=0.03, 1.95<Xstdweeks<=2.2, 3.2<Xstdmonths`	19	94.74%	2.37

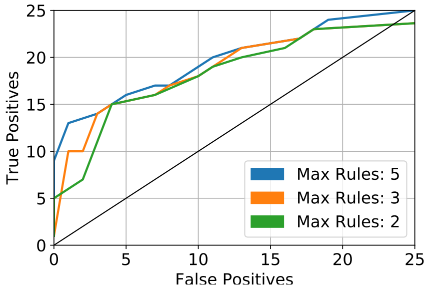


Figure 10: The figure presents PN curves for various values of the maximum rules extracted by the rule learner.

4 Rule Learning

We extract clusters and identify records that are associated with human trafficking using domain knowledge from experts. We featurize the extracted components using features like size of the cluster, the spatio-temporal characteristics, and the connectivity of the clusters. For our analysis, we consider only components with more than 300 advertisements. We then train a random forest classifier to predict if a cluster contains indicators of human trafficking. In order to establish statistical significance, we compare the ROC results of our classifier using four fold cross-validation for 100 random connected components versus the positive set. Figure 9 & Table 4 lists the performance of the classifier in terms of false positive and true positive Rate while Table 5 lists the most informative features for this classifier.

Additionally we extract rules from our feature set that help establish differences between the trafficking relevant clusters. Some of the rules with corresponding ratios and lift are given in Table 3. PN curves for Rule Learning are analogous to ROC Curves in Classification (Fürnkranz and Flach, 2005) and PN curves corresponding to various rules learnt are presented in the Figure 10. It can be observed that the features used by the rule learning to learn rules with maximum support and ratios, correspond to the ones labeled by the random forest as informative. This also serves as validation for the use of rule learning.

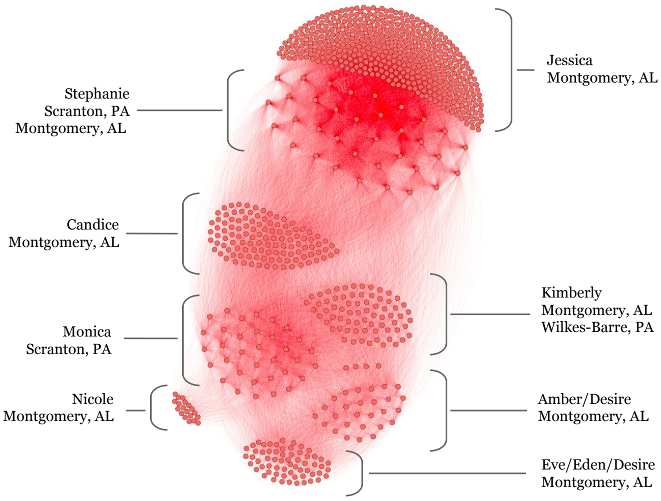


Figure 11: Representative entity isolated by our pipeline, believed to be involved in human trafficking. The nodes represent advertisements, while the edges represent links between advertisements. This entity has 802 nodes and 39,383 edges. This visualization is generated using Gephi. (Bastian et al., 2009). This entity operated in cities, across states and advertised multiple different individuals along with multiple phone numbers. This suggests a more complicated and organised activity and serves as an example of how complicated certain entities can be in this trade.

.

Table 4: Metrics for the Connected Component classifier

AUC	TPR@FPR=1%	FPR@TPR=50%
90.38%	66.6%	0.6%

Table 5: Most Informative Features

	Top 5 Features
1	`Posting Months`
2	`Posting Weeks`
3	`Std-Dev. of Image Frequency`
4	`Norm. No. of Names`
5	`Norm. No. of Unique Images`

5 Conclusion

In this paper we approached the problem of isolating sources of human trafficking from online escort advertisements with a pairwise Entity Resolution approach. We trained a classifier able to predict if two advertisements are from the same source using phone numbers as a strong feature which we exploit as proxy ground truth to generate training data. The resulting classifier proved to be robust, as evidenced from extremely low false positive rates. Other approaches like (Szekely et al., 2015) aim at building similar knowledge graphs using similarity score between each feature. This has some limitations. Firstly, we need labelled training data in order to train match functions to detect ontological relations. The challenge is aggravated since this approach considers each feature independently making generation of enough labelled training data for training multiple match functions an extremely complicated task.

Since we utilise existing features as proxy evidence, our approach can generate a large amount of training data without the need of any human annotation. Our approach requires just learning a single function over the entire featureset. Hence, our classifier can learn multiple complicated relations between features to predict a match, instead of the naive feature independence assumption.

We then proceeded to use this classifier in order to perform entity resolution using a heuristically learned match threshold. The resultant connected components were again featurised, and a classifier model was fit before subjecting to rule learning. On comparison with (Dubrawski et al., 2015), the connected component classifier performs a little better with higher values of the area under the ROC curve and the TPR@FPR=1% indicating a steeper ROC curve. We hypothesize that due to the entity resolution process, we are able to generate larger, more robust amount of training data which is immune to the noise in labelling and results in a stronger classifier. The learnt rules show high ratios and lift for reasonably high support as shown in Table 3. Rule learning also adds an element of interpretability to the models we built and as compared to more complex ensemble methods like Random Forests, having hard rules as classification models are preferred by domain experts to build evidence for incrimination.

6 Future Work

While our blocking scheme performs well to reduce the number of comparisons, scalability is still a significant challenge since our approach involves naive pairwise comparisons. One solution to this issue may be to design such a pipeline in a distributed environment. Another approach could be to use a computationally inexpensive technique to de-duplicate the dataset first, which would greatly help with regard to scalability.

In our approach, the ER process depends upon the heuristically learnt match threshold. Lower threshold values can significantly degrade the performance, producing extremely large connected components as a result. The possibility of treating this attribute as a learning task, would help making this approach more generic, and non domain specific.

Hashcodes of the images associated with the ads were also utilized as a feature for the match function. However, simple features like number of unique and common images etc., did not prove to be very informative. Further research is required in order to make better use of such visual data.

Acknowledgments

The authors would like to thank all staff, faculty and students who made the Robotics Institute Summer Scholars program 2015 at Carnegie Mellon University possible. This work has been partially supported by the National Institute of Justice (2013-IJ-CX-K007) and the Defense Advanced Research Projects Agency (FA8750-14-2-0244).

References

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. 2009. Swoosh: a generic approach to entity resolution. *The VLDB JournalThe International Journal on Very Large Data Bases* 18(1):255–276.

Hamish Cunningham. 2002. Gate, a general architecture for text engineering. *Computers and the Humanities* 36(2):223–254.

Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. 2015.

Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking* 1(1):65–85.

Johannes Fürnkranz and Peter A Flach. 2005. Roc nrule learningtowards a better understanding of covering algorithms. *Machine Learning* 58(1):39–77.

Emily Kennedy. 2012. Predictive patterns of sex trafficking online .

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12:2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

Pedro Szekely, Craig A. Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Subessware S. Karunamoorthy, Rajagopal Bojanapalli, Steven Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo, David Flynt, Rachel Artiss, Shih-Fu Chang, Tao Chen, Gerald Hiebel, and Lidia Ferreira. 2015. Building and using a knowledge graph to combat human trafficking. In *Proceedings of the 14th International Semantic Web Conference (ISWC 2015)*.

Sriharsha Veeramachaneni and Ravi Kumar Kondadadi. 2009. Surrogate learning: from feature independence to semi-supervised classification. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. Association for Computational Linguistics, pages 10–18.