Using Similarity Measures to Detect Organizations in Online Escort Advertisements

Carl Edwards[1], Anthony Wertz[2] and Artur Dubrawski[3]

Abstract— **Sex trafficking is a substantial problem in the world, and millions suffer from forced sexual exploitation. Due to the proliferation of internet, a significant amount of potential information for detecting trafficking behavior is now available online. By using this online data to trace organizations over time, law enforcement can both target specific trafficking organizations as well as better understand and address changes in trafficking activity. Past work has extracted various features from the data and attempted to cluster and classify it. In this paper, we approach this problem with several similarity measures in order to detect and monitor organizations in escort advertisement data. Our framework allows for easy incorporation of new similarities as well. Additionally, we examine multiple modalities (text advertisements and corresponding images) to enhance the detection of these trends by providing multiple perspectives. This work finds that organizations can be detected which show the evolution of advertisements temporally and geographically even as both names and phone numbers change.**

machine learning, escort advertisements, similarity measures, trend detection, multimodal data

I. Introduction

Human trafficking is a pervasive and global problem today. In 2017, the International Labour Organization estimated that 40 million people globally were victims of modern slavery. Among these, “3.8 million adults were victims of forced sexual exploitation and 1.0 million children were victims of commercial sexual exploitation in 2016” [1]. Due to the massive social impact of the internet, advertising for sex trafficking has moved online in the form of social networking sites and online classifieds [2]. In order to avoid detection by law enforcement, trafficking organizations attempt to blend in with non-trafficking related advertisements. They may periodically change phone numbers or other identifying features. Essentially, connecting advertisements becomes a game of whack-a-mole. An organization might pop up with a phone number in one location and then somewhere else with another number. By using multiple similarity measures, we are able to draw connections between advertisements even if these identifying features changes.

In 2014, DARPA launched its Memex program with a specific focus on combating human trafficking; this three year program resulted in a significant amount of research involving information retrieval, feature extraction, and classification and clustering to obtain and utilize data from this source of online information [3]. To best use this information, approaches have included training classifiers [4], [5], entity resolution [6], and graph-based techniques [7], [8] Additionally, some techniques go beyond text and incorporate multiple modalities of data using smaller supervised datasets such as in [4].

Recent work has focused on unsupervised natural language processing techniques such as word and document embeddings in order to extract and match templates for organizations [9] or to indicate sentences likely to be related to human trafficking [10].

Prior work has primarily focused on text or images, however, our technique allows for easy integration of new similarity measures from different modalities. Additionally, past efforts have frequently avoided the use of pairwise similarities since it is computationally intractable to compute them between all data points.

In this paper, we investigate the usage of multiple pairwise similarity measures to find trends and connections between advertisements. We examine text similarity based on word embeddings, similarities based on features extracted from the data such as phone numbers, names, and image hashes, and similarities based on face recognition. Additionally, we incorporate geospatial and temporal information into our framework. Like [9], we define an organization as a singular individual or group of related indivudals posting about escort services on backpage.com. Our approach would allow law enforcement officials to look for and monitor organizations suspected of sex trafficking and then build a case against them as they evolve and change over time. By combining multiple similarity measures, we are able to better characterize these organizations.

II. Methodology

A. Dataset

The dataset, D, consists of roughly 40 million advertisements which were scraped from the escorts section of backpage.com from September 2012 to December 2017. Each advertisement contains text, location, time, and images. There are approximately 20 million unique images present in the data and 562 unique locations, such as New York City. The text portion of the data is very noisy; emojis and mispellings are frequently used, words run together, and grammatical rules are ignored. Additionally, the dataset does not have labels due to its size and the domain expertise required to estimate an advertisement’s likelihood of being posted by traffickers.

[1] Carl Edwards is a senior student at the Department of EECS at the University of Tennessee cedwar45@utk.edu

[2,3] Anthony Wertz, Artur Dubrawski are in the Auton Lab, Robotics Institute, Carnegie Mellon University awertz@cmu.edu; awd@cs.cmu.edu

B. Similarity Measures

Similarities measures take two points, $q, d \in D$, and compute the similarity $f : D \times D \to \mathbb{R}$. In this work, we use similarity techniques which have a range of $[0, 1]$ where q and d are more related if $f(q, d)$ is nearer to 1 and less related if $f(q, d)$ is nearer to 0.

- Text similarity: We determine text similarity between advertisements using unsupervised word embeddings. In recent years, word embeddings which use distributed representations have significantly improved state-of-the-art results on a variety of NLP tasks [11], [12]. In particular, we use fastText embeddings [13], which extend `word2vec` [11] to allow for the incorporation of subword information; this is desirable for representing this noisy text data where characters such as emojis may replace a single letter in a word and mispellings are frequent. Additionally, this model is capable of producing embeddings for unseen words, which could allow new ads to be easily embedded by an already-trained model. A fastText model is initially trained using a skip-gram architecture [11] and default hyperparameters on a corpus consisting of the text body of every ad in the dataset. It creates 100-dimensional embeddings for the vocabulary in the corpus. Following this, paragraph embeddings for each ad are created by taking the average of the word embeddings for each word in the ad. According to [14], "simply averaging word embeddings of all words in a text has proven to be a strong baseline or feature across a multitude of tasks." Cosine similarity is used to compute a value between -1 (least similar) and 1 (most similar) to measure the similarity between two embeddings. This is the only similarity which outputs values below 0. However, empirical results produce only positive values. Cosine similarity [15] between two vectors v_1 and v_2 is defined as follows:

$$similarity(v_1, v_2) = \frac{< v_1, v_2 >}{\|v_1\| \|v_2\|}$$

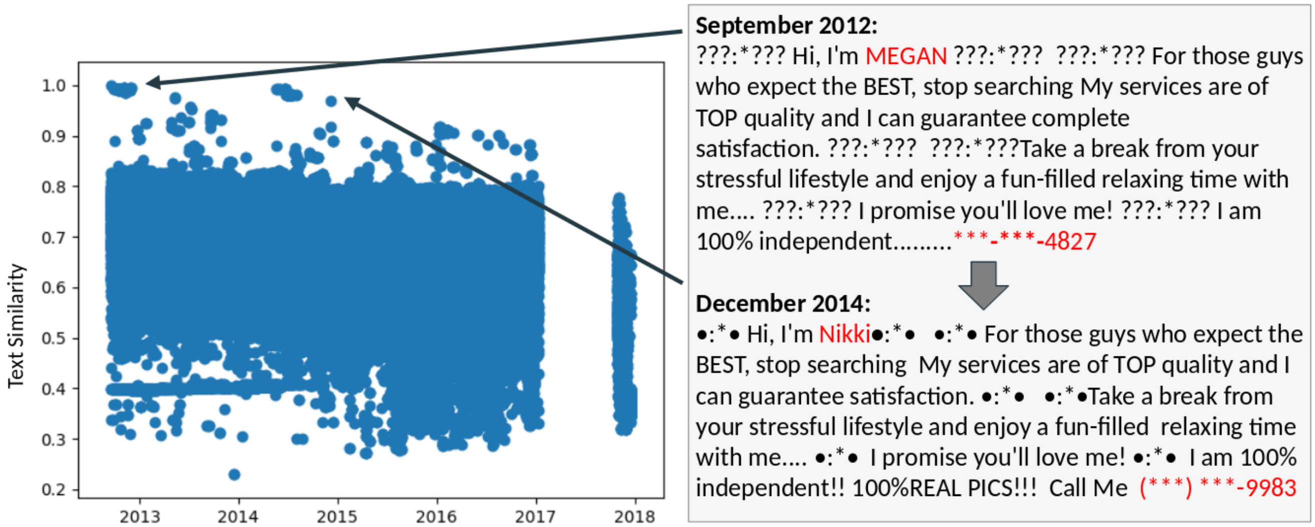
- Common-Feature similarity: This boolean similarity measure is 1 if two advertisements have a feature in common (e.g. a phone number) and 0 otherwise. We use it for the following similarities:
  - Phone Number Similarity: Phone numbers are extracted using regular expressions.
  - Image Hash Similarity: Images are hashed for each advertisement. This allows detection of image reuse between advertisements.
  - Name Similarity: Names are extracted using the **AnonymousExtractor** regex from [6].
- Face similarity: Faces are extracted and processed into 128-dimensional vectors using a pipeline of dlib pretrained models [16]. Face detection is performed using the CNN-based model `cnn_face_detection_model_v1`. This is chosen over the HOG (histogram of oriented gradients) model since it can be accelerated on a GPU. Next, a 5-point landmarking model trained on the dlib 5-point face landmark dataset is used to localize the faces. Finally, `dlib_face_recognition_resnet_model_v1` is used to create a 128-dimensional embedding for each detected face. This model is a ResNet based on [17] with only 29 convolutional layers instead of 34 and with the number of filters reduced by half. Two faces are considered to belong to the same person if the Euclidean distance between them is less than 0.6; this produces a binary output [16]. When using the HOG model instead of the CNN for face detection, this pipeline achieves a reported accuracy of 99.13% on the Labeled Faces in the Wild dataset [18]. The CNN-based model which we use, however, is reported to be more accurate than HOG [16]. Two advertisements are considered similar if they both have images containing the same face.

C. Pipeline

- Initially, the data is processed and textual features such as phone numbers are extracted using regex.
- A fastText model is trained on message body text, and paragraph embeddings are calculated by averaging constituent word embeddings.
- Faces are detected in the associated images and face embeddings are computed. This completes the extraction of features from the data.
- First, we optionally restrict our dataset to a specific location, such as New York City. We use an initial advertisement, or query point, to look for related advertisements. This is necessary since computing the pairwise similarity between all the data points is computationally intractable.
- Following this, we use a similarity function in order to isolate a smaller subset, $S \subseteq D$, of relevant advertisements. In this work, text similarity is used to create this subset. The text similarity is calculated between our query point and all other paragraph embeddings. A subset of advertisements which are above a certain threshold are selected as S.
- Next, other similarity measures, f, such as common-feature similarities and face similarity are used to link advertisements within this subset. For all $d_1, d_2 \in S$, d_1 and d_2 are linked if $f(d_1, d_2)$ is above some threshold for that similarity measure. This produces a temporal trail which links advertisements as they change over time.

D. Thickness of Tail

As a potential indicator of whether a query has produced an interesting trend, we use the thickness of the tail of the distribution of text similarities as a proxy representing signal vs noise in the subset, S, of text related to the query point, q. We calculate this as the ratio of points above two different similarity thresholds, the numerator threshold, t_n, and the denominator threshold, t_d. The number of ads above these two thresholds are counted: n_n and n_d respectively.



September 2012:
???:*??? Hi, I'm MEGAN ???:*??? ???:*??? For those guys who expect the BEST, stop searching My services are of TOP quality and I can guarantee complete satisfaction. ???:*??? ???:*???Take a break from your stressful lifestyle and enjoy a fun-filled relaxing time with me.... ???:*??? I promise you'll love me! ???:*??? I am 100% independent.........***-***-4827

December 2014:
•:*• Hi, I'm Nikki•:*• •:*• For those guys who expect the BEST, stop searching My services are of TOP quality and I can guarantee satisfaction. •:*• •:*•Take a break from your stressful lifestyle and enjoy a fun-filled relaxing time with me.... •:*• I promise you'll love me! •:*• I am 100% independent!! 100%REAL PICS!!! Call Me (***) ***-9983

Fig. 1: Each point represents an advertisement. The y-axis is the text similarity between the query point and each ad. The query point can be seen at 100% similarity in the upper left of the figure. This organization is found within 247,000 advertisements in Pittsburgh. Activity ceases after a couple months and resumes two years later with a different name and phone number.

We then calculate our thickness, $\Theta = \frac{n_n}{n_d}$. A high Θ value indicates that a large portion of the closely related ads are very closely related in text similarity: the tail gets thicker after t_n than it does at t_d.

III. Results

A. Text Similarity

We find that cosine similarity between unsupervised paragraph embeddings can be used to successfully extract advertisements related to the query point shown over time. Most advertisements fall in a range of 0.4 to 0.8. Advertisements over a certain threshold are considered related to the query point. We empirically determine this threshold as 0.95. We find that advertisement text is frequently reused, even over several years and when features like phone numbers and names change. The evolution of advertisements can also be observed, which can be seen in Figure 2 where the same organization modifies an advertisement from 2012 to 2016; the text similarity slowly decreases as the ad is continuously tweaked. Figure 1 shows an example of related ads in Pittsburgh which would not be linked by any of these features alone. Prior work often uses phone numbers as an oracle label for groups. Unfortunately, this approach can fail to successfully capture an entire organization. For example, Figure 3 showcases an organization which periodically changes its phone number. Additionally, many phone numbers are obfuscated, such as in Figure 6, which poses another challenge to using phone numbers to predict organizations.

Although text similarity can successfully be used to find related advertisements, false positives due to high text similarity to the query point are erroneously included in the subset of textually related ads. This creates noise which can make isolating organizations more difficult. Figure 5 shows an organization which is surrounded by this noise. Noise is much more common if a specific location is not selected to reduce the size of the subset. This is due to an increase in the magnitude of advertisements from a few hundred thousand in an individual city to the entire dataset of 40 million ads.

B. Linking

In order to address the noise in the subset of advertisements selected using text similarity, we isolate organizations from the noise by using additional similarity measures to link advertisements together. Since the non-text similarity measures we use are boolean in nature, we link ads if the similarity between them is 1. Additionally, this serves to help connect advertisements into organizations which may change names, phone numbers, and other features at various intervals of time. For example, Figure 3 show an organization which can be linked using text, name, and phone similarity. Figure 4 shows an example where an organization is isolated from noise using other similarity measures.

C. Face Similarity

In order to augment the text similarity and similarities based on features extracted from the text, we incorporate multiple modalities of the data by incorporating visual information in the form of faces. Previous studies, such as [6], only use image hashcodes. While hashcodes are very useful, they don't incorporate visual information which may help tie together organizations. Figure 4 shows how faces connect two groups of ads which image hashcodes do not. The top-left organization completely changes all its images in 2013, but they still share the same face.

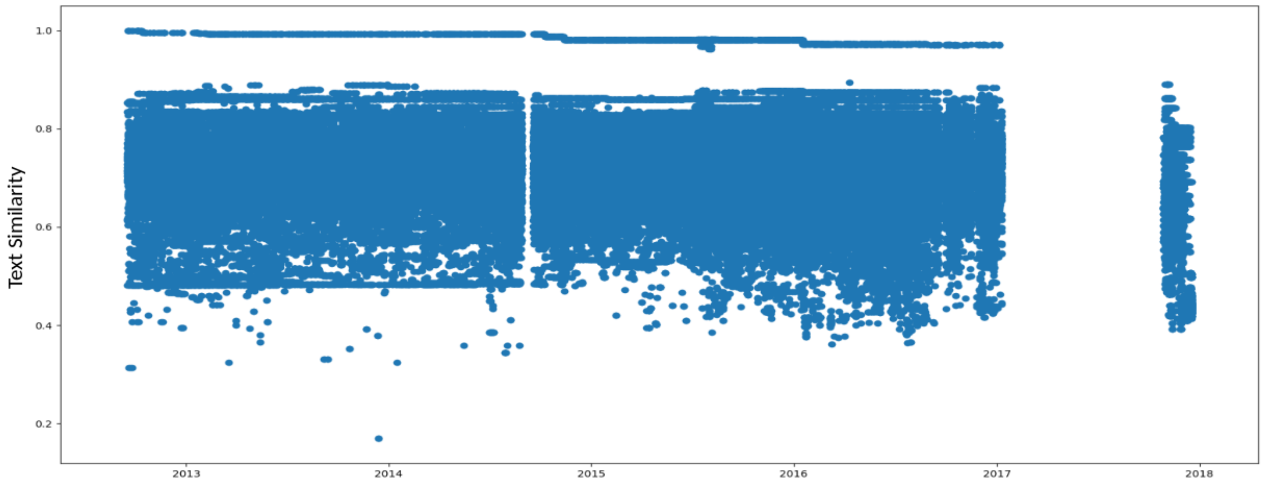
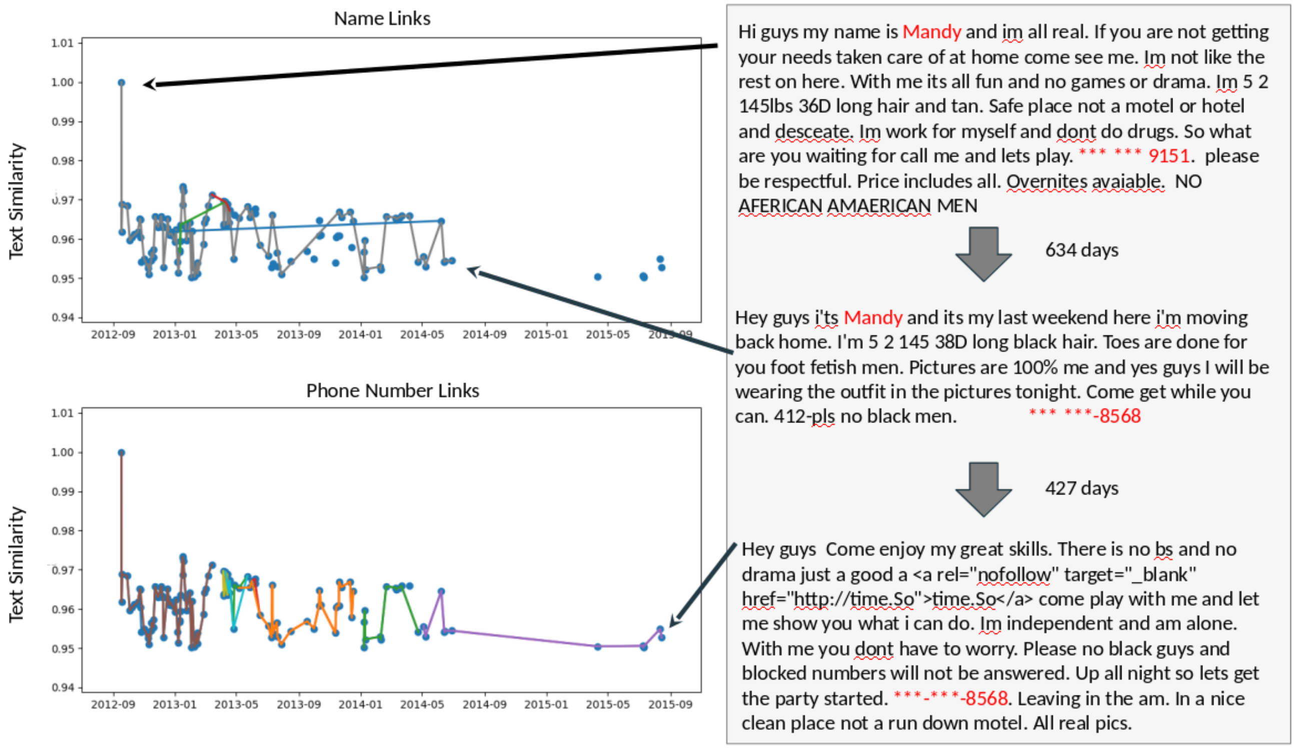


Fig. 2: This figure also shows another organization within Pittsburgh. An ad is posted over 2,000 times and continuously tweaked. Notably, the names of the organization's apparent pimps change. This organization was actually originally identified in [19] during 2012.



Hi guys my name is Mandy and im all real. If you are not getting your needs taken care of at home come see me. Im not like the rest on here. With me its all fun and no games or drama. Im 5 2 145lbs 36D long hair and tan. Safe place not a motel or hotel and desceate. Im work for myself and dont do drugs. So what are you waiting for call me and lets play. *** *** 9151. please be respectful. Price includes all. Overnites avaiable. NO AFERICAN AMAERICAN MEN

Hey guys i'ts Mandy and its my last weekend here i'm moving back home. I'm 5 2 145 38D long black hair. Toes are done for you foot fetish men. Pictures are 100% me and yes guys I will be wearing the outfit in the pictures tonight. Come get while you can. 412-pls no black men. *** ***-8568

Hey guys Come enjoy my great skills. There is no bs and no drama just a good a <a rel="nofollow" target="_blank" href="http://time.So">time.So</a> come play with me and let me show you what i can do. Im independent and am alone. With me you dont have to worry. Please no black guys and blocked numbers will not be answered. Up all night so lets get the party started. ***-***-8568. Leaving in the am. In a nice clean place not a run down motel. All real pics.

Fig. 3: An organization from Pittsburgh is linked together over several years by combining phone number and name similarities. Each colored line represents a shared attribute. For example, in the top plot the name 'Mandy' is connected using a gray line. The bottom plot shows periodic changes in phone number.

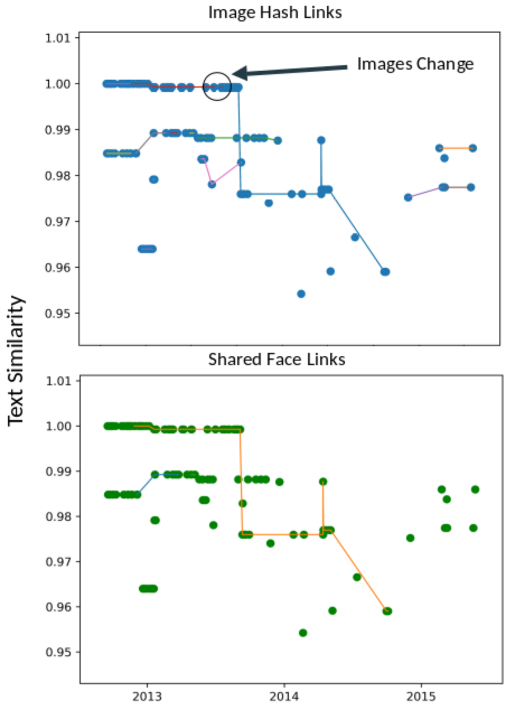


Fig. 4: This figure shows an organization (top left in plot) in LA which changes all its images at once. Face similarity is used to connect these two groups of ads with separate image hashes.

In this work, we extract approximately 635,000 faces from the first 5 million chronologically-posted unique images in the dataset. Due to time constraints and computational limitations, we were unable to process the remaining images. However, images which may have been initially posted in 2012 can often be found during later years, since images are frequently reused. The usage of stock images is also common.

Although using face recognition allows improved connections between ad images, the face recognition model appears to suffer from false positives. This may be due to irregular poses, occlusions, and intentional obfuscation. Additionally, the face recognition model may not perform as well on minorities. Figure 5 shows an example of two organizations which are linked using both text and face similarity, However, inspection shows they are not related. Due to privacy concerns, an example of matching faces which are false positives are not shown.

D. Multiple Locations

In addition to searching for organizations within specific cities, nationwide searches can also be conducted to find multi-city organizations, such as in Figure 6. In fact, the URLs in these ads appear to be subdomains of an organization taken down by the DoJ in early 2019 [20]. Unfortunately, searching all locations results in more noise being included in the textually related subset of advertisements due to the increase in magnitude of data. Since nationwide similarity computations are more expensive, we find it prudent to examine trends occurring within a specific location first before looking for multi-city activity. This is supported by [6]'s finding that location is the most informative feature for their entity resolution classifier.

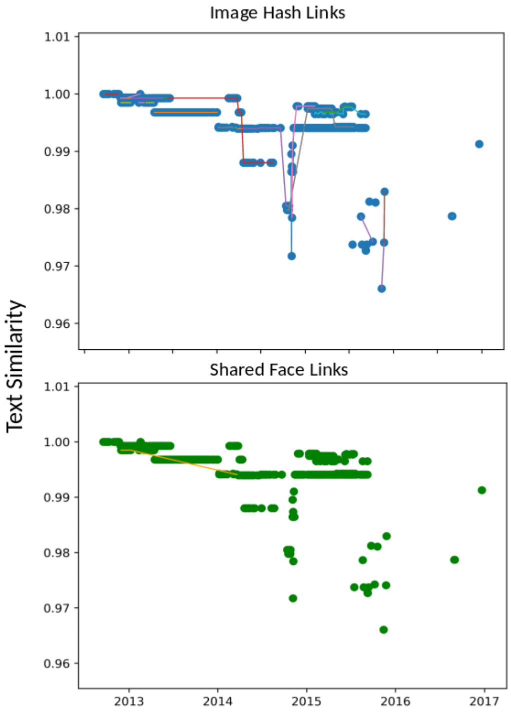
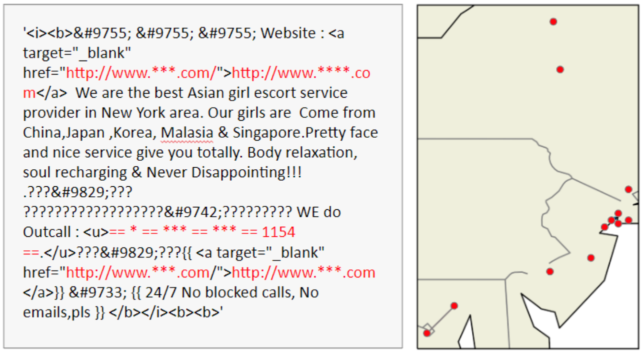


Fig. 5: Although these two organizations in LA are connected by faces, manual inspection shows that this is a false positive.



```
'<i><b>&#9755; &#9755; &#9755; Website : <a
target="_blank"
href="http://www.***.com/">http://www.****.co
m</a>  We are the best Asian girl escort service
provider in New York area. Our girls are  Come from
China,Japan ,Korea, Malasia & Singapore.Pretty face
and nice service give you totally. Body relaxation,
soul recharging & Never Disappointing!!!
.???&#9829;???
???????????????????&#9742;????????? WE do
Outcall : <u>== * == *** == *** == 1154
==.</u>???&#9829;???{{ <a target="_blank"
href="http://www.***.com/">http://www.***.com
</a>}} &#9733; {{ 24/7 No blocked calls, No
emails,pls }} </b></i><b><b>'
```

Fig. 6: This figure shows a query of all the data which finds a multi-city operation in several locations around the New York City metropolitan area. The text from one of these advertisements is shown above. The ads have location-specific URLs and obfuscated phone numbers.

E. Examining Thickness of Tail

We also find that the thickness of the tail of a distribution of text similarities can be indicative of an organization. An example of this can be seen in Table I. After examining this

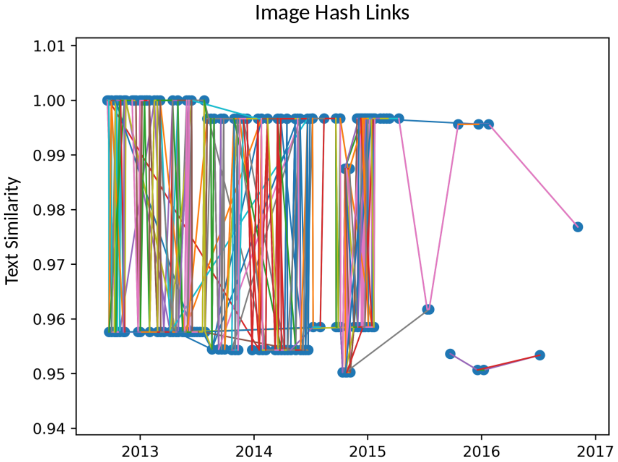


Fig. 7: An organization identified in data from Los Angeles. Its query point corresponds to index 13 in the Table I, which has a Θ value of 47.7

table, we then selected index 13 for investigation, resulting in Figure 7. It is important to note that although this metric indicates the likely existence of an organization, a thin tail does not indicate that no organization exists. In queries with more text noise, several unrelated organizations could appear. Additionally, there might be related ads whose text similarities are not above the threshold t_n.

Index	n_d	n_n	Θ
4	2451	64	2.61
5	1567	23	1.47
6	1980	19	0.96
8	330	153	46.36
9	410	1	0.24
11	988	929	94.03
12	4818	44	0.91
13	153	73	47.71
15	747	352	47.12
16	572	1	0.17

TABLE I: An example of comparing queries using the thickness of tail on a few advertisements in Los Angeles. Each row represents an ad with over 50 textually related ads taken from the first 20 ads. n_n is the number of ads above a similarity threshold of 0.99 and n_d is the number above a threshold of 0.95. Θ is the thickness of tail, $\frac{n_n}{n_d}$.

IV. Conclusion

In this paper, we examine the use of multiple similarity measures in order to find trends and connections indicating organizations within escort advertisements. We demonstrate how this technique can find organizations even as they change names and phone numbers. Additionally, the use of multiple similarities can help to remove erroneous noise from being included in potential organizations. In particular, this framework allows easy incorporation of new similarity techniques, and we leverage this capability in order to draw information from multiple modalities of the data, which can help provide a more complete picture of an organization. This would allow law enforcement to investigate for the existence of sex trafficking organizations, monitor organizations as their activity changes (such as attempting to conceal their identity), and build case evidence against them.

V. Future Work

In the future, our technique can be improved through the use of more features from [5] as well as incorporation of other similarities. The usage of visual information can be improved through similarities based on background and foreground segmentation and matching. Additionally, the use of a weighted average-based paragraph embedding and higher dimensional embeddings may improve the quality of the measure of text similarity.

Acknowledgment

The authors would like to thank the members of the Auton Lab and Robotics Institute Summer Scholars program 2019 for all their insightful help and advice. Additionally, we would like to acknowledge the support of the NSF REU program.

References

[1] I. L. Organization, “Global estimates of modern slavery: Forced labour and forced marriage,” 2017.

[2] M. Latonero, “Human trafficking online: The role of social networking sites and online classifieds,” *Available at SSRN 2045851*, 2011.

[3] K. Hundman, T. Gowda, M. Kejriwal, and B. Boecking, “Always lurking: Understanding and mitigating bias in online human trafficking detection,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018, pp. 137–143.

[4] E. Tong, A. Zadeh, C. Jones, and L.-P. Morency, “Combating human trafficking with deep multimodal models,” *arXiv preprint arXiv:1705.02735*, 2017.

[5] A. Dubrawski, K. Miller, M. Barnes, B. Boecking, and E. Kennedy, “Leveraging publicly available data to discern patterns of human-trafficking activity,” *Journal of Human Trafficking*, vol. 1, no. 1, pp. 65–85, 2015.

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

[7] R. Rabbany, D. Bayani, and A. Dubrawski, “Active search of connections for case building and combating human trafficking,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2120–2129.

[8] P. Szekely, C. A. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, K. Knight *et al.*, “Building and using a knowledge graph to combat human trafficking,” in *International Semantic Web Conference*. Springer, 2015, pp. 205–221.

[9] L. Li, O. Simek, A. Lai, M. Daggett, C. K. Dagli, and C. Jones, “Detection and characterization of human trafficking networks using unsupervised scalable text template matching,” in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 3111–3120.

[10] M. Kejriwal, J. Ding, R. Shao, A. Kumar, and P. Szekely, “Flagit: A system for minimally supervised human trafficking indicator mining,” *arXiv preprint arXiv:1712.03086*, 2017.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[12] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[14] T. Kenter, A. Borisov, and M. De Rijke, “Siamese cbow: Optimizing word embeddings for sentence representations,” *arXiv preprint arXiv:1606.04640*, 2016.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,” 2008.

[19] E. Kennedy, “Predictive patterns of sex trafficking online,” *Dietrich College Honors Theses*, 2012.

[20] “Nationwide sting operation targets illegal asian brothels, six indicted for racketeering,” Jan 2019. [Online]. Available: https://www.justice.gov/usao-or/pr/nationwide-sting-operation-targets-illegal-asian-brothels-six-indicted-racketeering