Combating Human Trafficking with Deep Multimodal Models

Edmund Tong*
Language Technologies Institute
Carnegie Mellon University
edtong@cmu.edu

Amir Zadeh*
Language Technologies Institute
Carnegie Mellon University
abagherz@cs.cmu.edu

Cara Jones
Marinus Analytics, LLC
cara@marinusanalytics.com

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
morency@cs.cmu.edu

Abstract

Human trafficking is a global epidemic affecting millions of people across the planet. Sex trafficking, the dominant form of human trafficking, has seen a significant rise mostly due to the abundance of escort websites, where human traffickers can openly advertise among at-will escort advertisements. In this paper, we take a major step in the automatic detection of advertisements suspected to pertain to human trafficking. We present a novel dataset called Trafficking-10k, with more than 10,000 advertisements annotated for this task. The dataset contains two sources of information per advertisement: text and images. For the accurate detection of trafficking advertisements, we designed and trained a deep multimodal model called the Human Trafficking Deep Network (HTDN).

1 Introduction

Human trafficking "a crime that shames us all" (UNODC, 2008), has seen a steep rise in the United States since 2012. The number of cases reported rose from 3,279 in 2012 to 7,572 in 2016—more than doubling over the course of five years (Hotline). Sex trafficking is a form of human trafficking, and is a global epidemic affecting millions of people each year (McCarthy, 2014). Victims of sex trafficking are subjected to coercion, force, and control, and are not able to ask for help. Put plainly, sex trafficking is modern-day slavery and is one of the top priorities of law enforcement agencies at all levels.

A major advertising ground for human traffickers is the World Wide Web. The Internet has brought traffickers the ability to advertise online and has fostered the growth of numerous adult escort sites. Each day, there are tens of thousands of Internet advertisements posted in the United States and Canada that market commercial sex. Hiding among the noise of at-will adult escort ads are ads posted by sex traffickers. Often long undetected, trafficking rings and escort websites form a profit cycle that fuels the increase of both trafficking rings and escort websites.

For law enforcement, this presents a significant challenge: how should we identify advertisements that are associated with sex trafficking? Police have limited human and technical resources, and manually sifting through thousands of ads in the hopes of finding something suspicious is a poor use of those resources, even if they know what they are looking for. Leveraging state-of-the-art machine learning approaches in Natural Language Processing and computer vision to detect and report advertisements suspected of trafficking is the main focus of our work. In other words, we strive to find the victims and perpetrators of trafficking who hide in plain sight in the massive amounts of data online. By narrowing down the number of advertisements that law enforcement must sift through, we endeavor to provide a real opportunity for law enforcement to intervene in the lives of victims. However, there are non-trivial challenges facing this line of research:

Adversarial Environment. Human trafficking rings are aware that law enforcement monitors their online activity. Over the years, law enforcement officers have populated lists of keywords that frequently occur in trafficking advertisements. However, these simplistic queries fail when traffickers use complex obfuscation. Traffickers, again aware of this, move to new keywords to blend in with the at-will escort advertisements. This trend creates an adversarial environment for any machine learning

* Authors contributed equally.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1547–1556
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1142

system that attempts to find trafficking rings hiding in plain sight.

Defective Language Compositionality. Online escort advertisements are difficult to analyze, because they lack grammatical structures such as constituency. Therefore, any form of inference must rely more on context than on grammar. This presents a significant challenge to the NLP community. Furthermore, the majority of the ads contain emojis and non-English characters.

Generalizable Language Context. Machine learning techniques can easily learn unreliable cues in training sets such as phone numbers, keywords, and other forms of semantically unreliable discriminators to reduce the training loss. Due to limited similarity between the training and test data due to the large number of ads available online, relying on these cues is futile. Learned discriminative features should be generalizable and model semantics of trafficking.

Multimodal Nature. Escort advertisements are composed of both textual and visual information. Our model should treat these features interdependently. For instance, if the text indicates that the escort is in a hotel room, our model should consider the effect that such knowledge may have on the importance of certain visual features.

We believe that studying human trafficking advertisements can be seen as a fundamental challenge to the NLP, computer vision, and machine learning communities dealing with language and vision problems. In this paper, we present the following contributions to this research direction. First, we study the language and vision modalities of the escort advertisements through deep neural modeling. Second, we take a significant step in automatic detection of advertisements suspected of sex trafficking. While previous methods (Dubrawski et al., 2015) have used simplistic classifiers, we build an end-to-end-trained multimodal deep model called the Human Trafficking Deep Network (HTDN). The HTDN uses information from both text and images to extract cues of human trafficking, and shows outstanding performance compared to previously used models. Third, we present the first rigorously annotated dataset for detection of human trafficking, called Trafficking-10k, which includes more than 10,000 trafficking ads labeled with likelihoods of having been posted by traffickers.[1]

[1]Due to the sensitive nature of this dataset, access can only be granted by emailing Cara Jones. Different levels of access are provided *only* to scientific community.

2 Related Works

Automatic detection of human trafficking has been a relatively unexplored area of machine learning research. Very few machine learning approaches have been proposed to detect signs of human trafficking online. Most of these approaches use simplistic methods such as multimedia matching (Zhou et al., 2016), text-based filtering classifiers such as random forests, logistic regression, and SVMs (Dubrawski et al., 2015), and named-entity recognition to isolate the instances of trafficking (Nagpal et al., 2015). Studies have suggested using statistical methods to find keywords and signs of trafficking from data to help law enforcement agencies (Kennedy, 2012) as well as adult content filtering using textual information (Zhou et al., 2016).

Multimodal approaches have gained popularity over the past few years. These multimodal models have been used for medical purposes, such as detection of suicidal risk, PTSD and depression (Scherer et al., 2016; Venek et al., 2016; Yu et al., 2013; Valstar et al., 2016); sentiment analysis (Zadeh et al., 2016b; Poria et al., 2016; Zadeh et al., 2016a); emotion recognition (Poria et al., 2017); image captioning and media description (You et al., 2016; Donahue et al., 2015); question answering (Antol et al., 2015); and multimodal translation (Specia et al., 2016).

To the best of our knowledge, this paper presents the first multimodal and deep model for detection of human trafficking.

3 Trafficking-10k Dataset

In this section, we present the dataset for our studies. We formalize the problem of recognizing sex trafficking as a machine learning task. The input data is text and images; this is mapped to a measure of how suspicious the advertisement is with regards to human trafficking.

3.1 Data Acquisition and Preprocessing

A subset of 10,000 ads were sampled randomly from a large cache of escort ads for annotation in Trafficking-10k dataset. The distribution of advertisements across the United States and Canada is shown in Figure 1, which indicates the diversity of advertisements in Trafficking-10k. This diversity ensures that models trained on Trafficking-10k can be applicable nationwide. The 10,000 collected ads

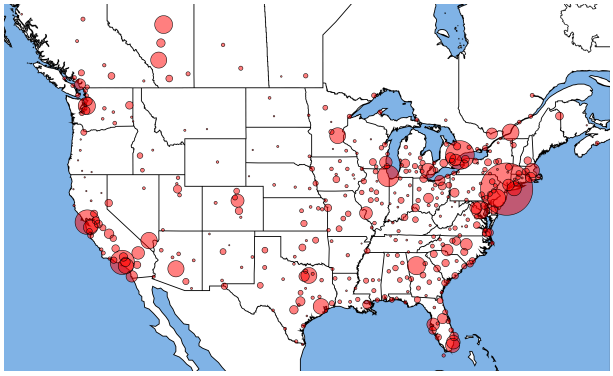

Figure 1: Distribution of advertisements in Trafficking-10k dataset across United States and Canada.

each consist of text and zero or more images. The text in the dataset is in plain text format, derived by stripping the HTML tags from the raw source of the ads. The set of characters in each advertisement is encoded as UTF-8, because there is ample usage of smilies and non-English characters. Advertisements are truncated to the first 184 words, as this covers more than 90% of the ads. Images are resized to 224×224 pixels with RGB channels.

3.2 Trafficking Annotation

Detecting whether or not an advertisement is suspicious requires years of practice and experience in working closely with law enforcement. As a result, annotation is a highly complicated and expensive process, which cannot be scaled using crowdsourcing. In our dataset, annotation is carried out by two expert annotators, each with at least five years of experience, in detection of human trafficking and another annotator with one year of experience. In our dataset, annotations were done by three experts. One expert has over a year of experience, and the other two have over five years of experience in the human trafficking domain. To calculate the inter-annotator agreement, each annotator is given the same set of 1000 ads to annotate and the nominal agreement is found: there was a 83% pairwise agreement (0.62 Krippendorff's alpha). Also, to make sure that annotations are generalizable across the annotators and law enforcement officers, two law enforcement officers annotated, respectively, a subset of 500 and 100 of the advertisements. We found a 62% average pairwise agreement (0.42 Krippendorff's alpha) with our annotators. This gap is reasonable, as law enforcement officers only have experience with local advertisements, while Trafficking-10k annotators have experience with cases across the United States.

Annotators used an annotation interface specifically designed for the Trafficking-10k dataset. In the annotation interface, each advertisement was displayed on a separate webpage. The order of the advertisements is determined uniformly randomly, and annotators were unable to move to the next advertisement without labeling the current one. For each advertisement, the annotator was presented with the question: "In your opinion, would you consider this advertisement suspicious of human trafficking?" The annotator is presented with the following options: "Certainly no," "Likely no," "Weakly no," "Unsure,"[2] "Weakly yes," "Likely yes," and "Certainly yes." Thus, the degree to which advertisements are suspicious is quantized into seven levels.

3.3 Analysis of Language

The language used in these advertisements introduces fundamental challenges to the field of NLP. The nature of the textual content in these advertisements raises the question of how we can make inferences in a linguistic environment with a constantly evolving lexicon. Language used in the Trafficking-10k dataset is highly inconsistent with standard grammar. Often, words are obfuscated by emojis and symbols. The word ordering is inconsistent, and there is rarely any form of constituency. This form of language is completely different from spoken and written English. These attributes make escort advertisements appear somewhat similar to tweets, specifically since these ads are normally short (more than 90% of the ads have at most 184 words). Another point of complexity in these advertisements is the high number of unigrams, due to usage of uncommon words and obfuscation. On top of unigram complexity, advertisers continuously change their writing pattern, making this problem more complex.

3.4 Dataset Statistics

There are 106,954 distinct unigrams, 353,324 distinct bigrams, and 565,403 trigrams in the Trafficking-10k dataset. There are 60,337 images. The total number of distinct characters including whitespace, punctuations, and hex characters is 182. The average length of an ad is 137 words, with a

[2]This option is greyed out for 10 seconds to encourage annotators to make an intuitive decision.

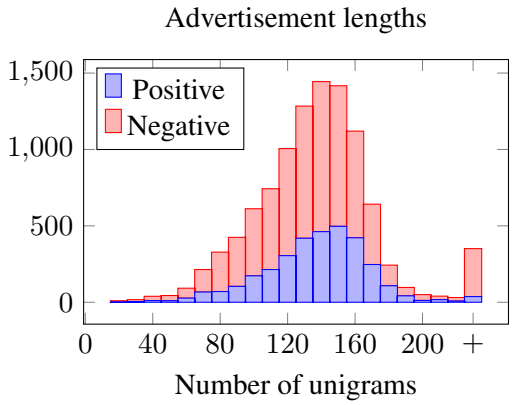

Figure 2: Distribution of the length of advertisements in Trafficking-10k. There is no significant difference between positive and negative cases purely based on length.

standard deviation of 74, median 133. The shortest advertisement has 7 unigrams, and the longest advertisement has 1810 unigrams. There are of 106,954 distinct unigrams, 353,324 distinct bigrams and 565,403 trigrams in the Trafficking-10k dataset. The average number of images in an advertisement is 5.9; the median is 5, the minimum is 0, and the maximum is 90.

The length of *suspected* advertisements is 134 unigrams; the standard deviation is 39, the minimum is 12, and the maximum is 666. The length of *non-suspected* ads is 141; the standard deviation is 85, the minimum is 7, and the maximum is 1810. The total number of suspected ads is 3257; and the total number of non-suspected ads is 6992. Figure 2 shows the histogram of number of ads based on their length. Both the positive and negative distributions are similar. This means that there is no obvious length difference between the two classes. Most of the ads have a length of 80–180 words.

4 Model

In this section, we present our deep multimodal network called the Human Trafficking Deep Network (HTDN). The HTDN is a multimodal network with language and vision components. The input to the HTDN is an ad, text and images. The HTDN is shown in Figure 3. In the remainder of this section, we will outline the different parts of the HTDN, and the input features to each component.

4.1 Trafficking Word Embeddings

Our approach to deal with the adversarial environment of escort ads is to use word vectors, defining words not based on their constituent characters, but rather based on their context. For instance, consider the two unigrams "cash" and "©a$h." While these contain different characters, semantically they are the same, and they occur in the same context. Thus, our expectation is that both the unigrams will be mapped to similar vectors. Word embeddings pretrained on general domains do not cover most of the unigrams in Trafficking-10k. For instance, the GloVe embedding (Pennington et al., 2014) trained on Wikipedia covers only 49.7% of our unigrams. The first step of the HTDN pipeline is to train word vectors (Mikolov et al., 2013) based on the skip-gram model. This is especially suitable for escort ads, because skip-gram models are able to capture context without relying on word order. We train the word embedding using 1,000,000 unlabeled ads from a dataset that does not include the Trafficking-10k data. For each advertisement, the input to the trained embedding is a sequence of words $\hat{\mathbf{w}} = [\hat{w}_1, \ldots, \hat{w}_t]$, and the output is a sequence of 100-dimensional word vectors $\mathbf{w} = [w_1, \ldots, w_t]$, where t is the size of the advertisement and $w_i \in \mathbb{R}^{100}$. Our trained word vectors cover 94.9% of the unigrams in the Trafficking-10k dataset.

4.2 Language Network

Our language network is designed to deal with two challenging aspects of escort advertisements: (1) violation of constituency, and (2) presence of irrelevant information not related to trafficking but present in ads. We address both of these issues by learning a time dependent embedding at word level. This allows the model to not rely on constituency and also remember useful information from the past, should the model get overwhelmed by irrelevant information. Our proposed language network, $\mathcal{F}_l$, takes as input a sequence of word vectors $\mathbf{w} = [w_1, \ldots, w_t]$, and outputs a neural language representation h_l. As a first step, $\mathcal{F}_l$ uses the word embeddings as input to a Long-Short Term Memory (LSTM) network and produces a new supervised context-aware word embedding $\mathbf{u} = [u_1, \ldots, u_t]$ where $u_i \in \mathbb{R}^{300}$ is the output of the LSTM at time i. Then, $\mathbf{u}$ is fed into a fully connected layer with dropout $p = 0.5$ to produce the neural language representation $h_l \in \mathbb{R}^{300}$ according to the following formulas with weights W_l for the LSTM and implicit weights in the fully

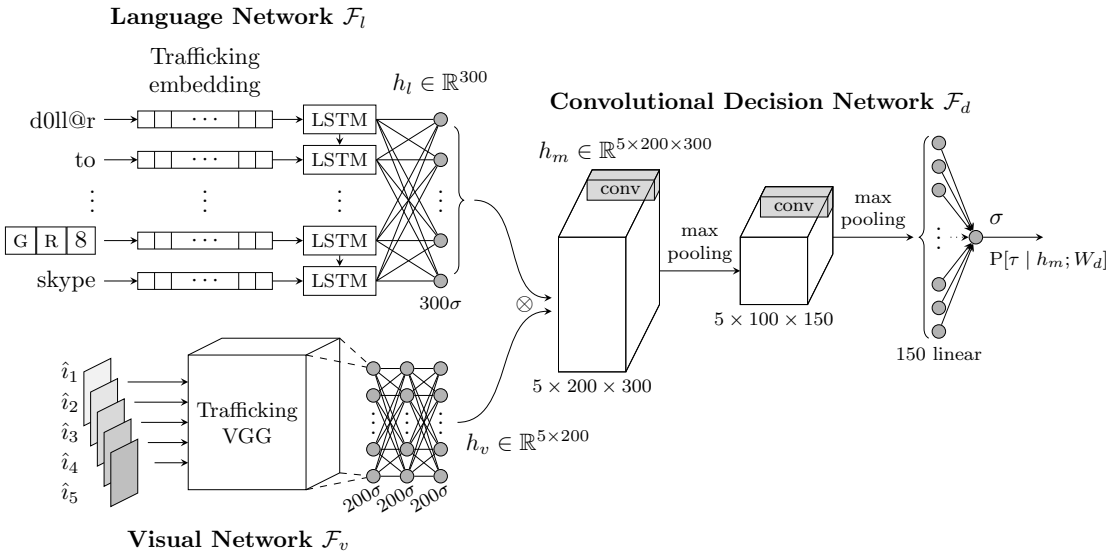

Figure 3: Overview of our proposed Human Trafficking Deep Network (HTDN). The input to HTDN is text and a set of 5 images. The text goes through the Language Network $\mathcal{F}_l$ to get the language representation h_l and the set of 5 images go through the Vision Network $\mathcal{F}_v$ to get the visual representation h_v. h_l and h_v are then fused together to get the multimodal representation h_m. The Convolutional Decision Network $\mathcal{F}_d$ conditioned on the h_m makes inference about whether or not the advertisement is suspected of trafficking

connected layers, which we represent by FC:

$$u_i = LSTM(i, w_i; W_l) \quad (1)$$

$$\mathbf{u} = [u_1, \ldots, u_t] \quad (2)$$

$$h_l = FC(\mathbf{u}). \quad (3)$$

The generated h_l is then used as part of the HTDN pipeline, and is also trained independently to assess the performance of the language-only model. The language network $\mathcal{F}_l$ is the combination of the LSTM and the fully-connected network.

4.3 Vision Network

Parallel to the language network, the vision network $\mathcal{F}_v$ takes as input advertisement images and extracts visual representations h_v. The vision network takes at most five images; the median number of images per advertisement in Trafficking-10k is 5. To learn contextual and abstract information from images, we use a deep convolutional neural network called Trafficking-VGG (T-VGG), a fine-tuned instance of the well-known VGG network (Simonyan and Zisserman, 2014). T-VGG is a deep model with 13 consecutive convolutional layers followed by 2 fully connected layers; it does not include the softmax layer of VGG. The procedure for fine-tuning T-VGG maps each individual image to a label that comes from the advertisement, and then performs end-to-end training. For example, if there are five images in an advertisement with positive label, all five images are mapped to positive label. After fine-tuning, three fully connected layers of 200 neurons with dropout $p = 0.5$ are added to the network. The combination of T-VGG and the fully connected layers is the vision network $\mathcal{F}_l$. We consider five images $\hat{\boldsymbol{i}} = \{\hat{i}_1, \ldots, \hat{i}_5\}$ from each input advertisement. If the advertisement has fewer than five images, zero-filled images are added. For each image, the output of $\mathcal{F}_v$ is a representation of five images $\mathbf{i} = \{i_1, \ldots, i_5\}$. The visual representation $h_v \in \mathbb{R}^{5\times200}$ is a matrix with a size-200 representation of each of the 5 images:

$$h_v = \mathcal{F}_v(\hat{\boldsymbol{i}}; W_v). \quad (4)$$

4.4 Multimodal Fusion

Escort advertisements have complex dynamics between text and images. Often, neither linguistic nor visual cues alone can suffice to classify whether an ad is suspicious. Interactions between linguistic and visual cues can be non-trivial, so this requires an explicit joint representation for each neuron in the linguistic and visual representations. In our multimodal fusion approach we address this by calculating an outer product between language and visual representations h_l and h_v to build the full space of possible outcomes:

$$h_m = h_l \otimes h_v, \quad (5)$$

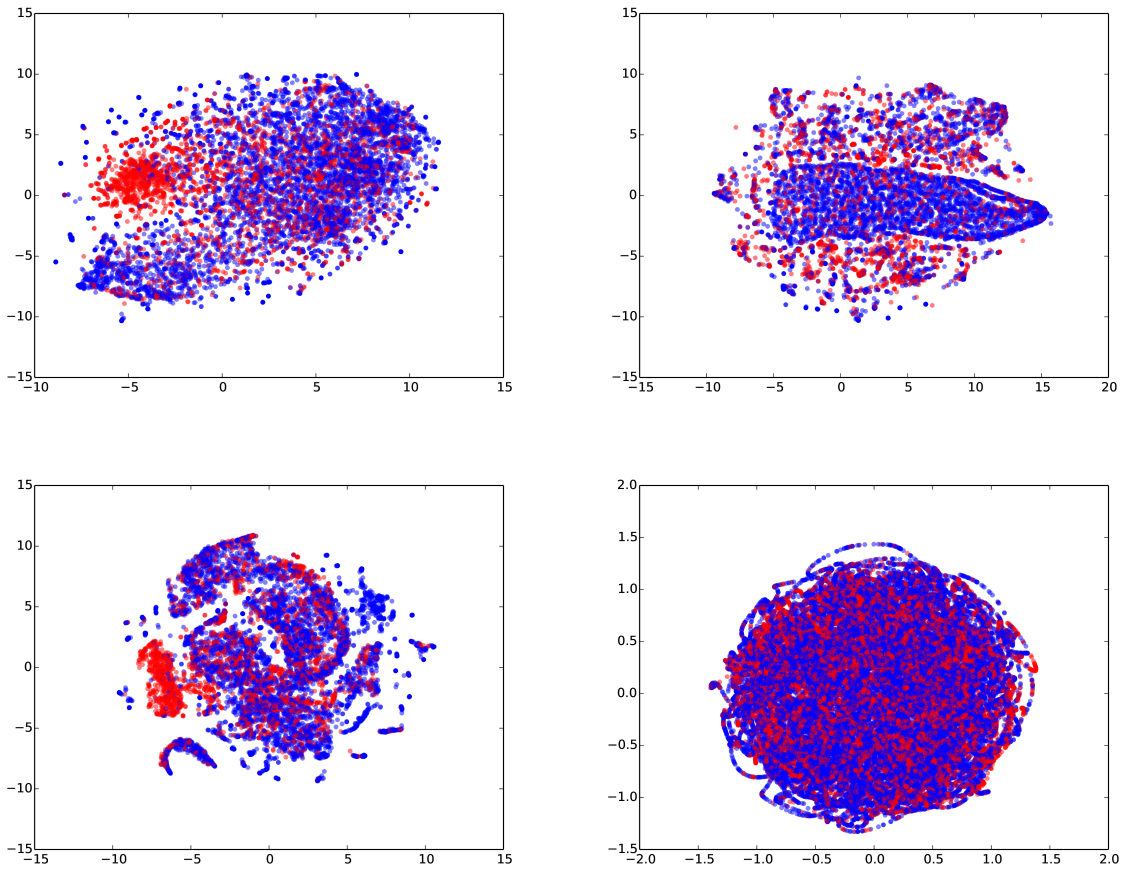

Figure 4: 2D t-SNE representation of different input features for baseline models. Clockwise from top left: one hot vectors with expert data, one hot vectors without expert data, visual features from Vision Network $\mathcal{F}_v$, and average word vectors. These representations show that inference is not trivial in Trafficking-10k dataset.

where $\otimes$ is an outer product of the two representations. This creates a joint multimodal tensor called h_m for language and visual modalities. In this tensor, every neuron in the language representation is multiplied by every neuron in vision representation, thus creating a new representation containing the information of both of them. Thus, the final fusion tensor $h_m \in \mathbb{R}^{5\times200\times300}$ contains information from the joint interaction of the language and visual modalities.

4.5 Convolutional Decision Network

The multimodal representation h_m is used as the input to the convolutional decision network $\mathcal{F}_d$. $\mathcal{F}_d$ has two layers of convolution and max pooling with a dropout rate of $p = 0.5$, followed by a fully connected layer of 150 neurons with a dropout rate of $p = 0.5$. Performing convolutions in this space enables the model to attend to small areas of linguistic and visual cues. It can thus find correspondences between specific combinations of the linguistic and visual representations. The final decision is made by a single sigmoid neuron.

5 Experiments

In our experiments, we compare the HTDN with previously used approaches for detection of trafficking suspicious ads. Furthermore, we compare the HTDN to the performance of its unimodal components. In all our experiments we perform binary classification of whether the advertisement is suspected of being related to trafficking. The main comparison method that we use is the weighted accuracy and F1-score (due to imbalance it dataset). The formulation for weighted accuracy is as follows:

$$\text{Wt. Acc.} = \frac{\text{TP} \times \text{N/P} + \text{TN}}{2\text{N}} \tag{6}$$

Model	Wt. Acc. (%)	F1 (%)	Acc. (%)	Precision (%)	Recall (%)
Random	50.0	-	68.2	-	-
Keywords					
Random Forest	67.0	55.2	78.1	78.2	42.6
Logistic Regression	69.9	57.8	78.4	75.5	46.8
Linear SVM	69.5	57.0	78.6	78.0	44.9
Average Trafficking Vectors					
Random Forest	67.3	54.1	78.0	79.3	41.1
Logistic Regression	72.2	61.7	80.2	79.2	50.6
Linear SVM	70.3	57.7	79.2	80.7	44.9
108 One-Hot					
Random Forest	62.4	60.7	72.6	61.5	60.0
Logistic Regression	62.5	45.1	72.2	60.0	36.1
Linear SVM	61.7	45.1	71.8	58.6	36.7
Bag of Words					
Random Forest	57.6	24.5	70.4	63.2	15.2
Logistic Regression	71.1	24.5	70.4	63.2	15.2
Linear SVM	71.2	24.5	70.4	63.2	15.2
HTDN Unimodal					
$\mathcal{F}_l$	74.5	65.8	78.8	69.8	62.3
$\mathcal{F}_v$ [VGG]	69.1	58.4	74.2	66.7	52.0
$\mathcal{F}_v$ [T-VGG]	70.4	59.5	77.3	78.3	48.0
HTDN	**75.3**	**66.5**	80.0	71.4	62.2
Human	83.7	73.7	84.0	76.7	70.9

Table 1: Results of our experiments. We compare our HTDN model to various baselines using different inputs. HTDN ourperforms other baselines in both weighted accuracy and F-score.

where TP (resp. TN) is true positive (resp. true negative) predictions, and P (resp. N) is the total number of positive (resp. negative) examples.

5.1 Baselines

We compare the performance of the HTDN network with baseline models divided in 4 major categories

Bag-of-Words Baselines. This set of baselines is designed to assess performance of off-the-shelf basic classifiers and basic language features. We train random forest, logistic regression and linear SVMs to show the performance of simple language-only models.

Keyword Baselines. These demonstrate the performance of models that use a set of 108 keywords, all highly related to trafficking, provided by law enforcement officers.[3] A binary one-hot vector representing these keywords is used to train the random forest, logistic regression, and linear SVM models.

108 One-Hot Baselines. Similar to Keywords Baseline, we use feature selection technique to filter the most informative 108 words for detection of trafficking. We compare the performance of this baseline to Keywords baseline to evaluate the usefulness of expert knowledge in keywords selection vs automatic data-driven keyword selection.

Average Trafficking Vectors Baselines. We assess the magnitude of success for the trafficking word embeddings for different classifiers. For the random forest, logistic regression, and linear SVM models, the average word vector is calculated and used as input.

HTDN Unimodal. These baselines show the performance of unimodal components of HTDN. For language we only use $\mathcal{F}_l$ component of the pipeline and for visual we use $\mathcal{F}_v$, using both pre-trained a VGG and finetuned T-VGG.

[3]Not presented in this paper due to sensitive nature of these keywords.

Random and Human. Random is based on assigning the more frequent class in training set to all the test data, and can be considered a lower bound for our model. Human performance metrics are upper bounds for this task's metrics.

We visualize the different inputs to our baseline models to show the complexity of the dataset when using different feature sets. Figure 4 shows the 2D t-SNE (Maaten and Hinton, 2008) representation of the training data in our dataset according to the Bag-of-Words (top right) models, expert keywords (top left), average word vectors (bottom right), and the visual representation h_v bottom left. The distribution of points suggests that none of the feature representations make the classification task trivial.

5.2 Training Parameters

All the models in our experiments are trained on the Trafficking-10k designated training set and tested on the designated test set. Hyperparameter evaluation is performed using a subset of training set as validation set. The HTDN model is trained using the Adam optimizer (Kingma and Ba, 2014). The neural weights were initialized randomly using Xavier initialization technique (Glorot and Bengio, 2010). The random forest model uses 10 estimators, with no maximum depth, and minimum-samples-per-split value of 2. The linear SVM model uses an ℓ_2-penalty and a square hinge loss with $C = 1$.

6 Results and Discussion

The results of our experiments are shown in Table 1. We report the results on three metrics: F1-score, weighted accuracy, and accuracy. Due to the imbalance between the numbers of positive and negative samples, weighted accuracy is more informative than unweighted accuracy, so we focus on the former.

HTDN. The first observation from Table 1 is that the HTDN model outperforms all the proposed baselines. There is a significant gap between the HTDN (and variants) and other non-neural approaches. This better performance is an indicator of complex interactions in detecting dynamics of human trafficking, which is captured by the HTDN.

Both Modalities are Helpful. Both modalities are helpful in predicting signs of trafficking ($\mathcal{F}_l$ and $\mathcal{F}_v$ [T-VGG]). Fine-tuning VGG network parameters shows improvement over pre-trained VGG parameters.

Language is More Important. Since $\mathcal{F}_l$ shows better performance than $\mathcal{F}_v$ [T-VGG], the language modality appears to be the more informative modality for detecting trafficking suspicious ads.

7 Conclusion and Future Work

In this paper, we took a major step in multimodal modeling of suspected online trafficking advertisements. We presented a novel dataset, Trafficking-10k, with more than 10,000 advertisements annotated for this task. The dataset contains two modalities of information per advertisement: text and images. We designed a deep multimodal model called the Human Trafficking Deep Network (HTDN). We compared the performance of the HTDN to various models that use language and vision alone. The HTDN outperformed all of these, indicating that using information from both sources may be more helpful than using just one.

Exploring language through character modeling. In order to eliminate the need for retraining the word vectors as the language of the domain evolves, we plan to use character models to learn a better language model for trafficking. As new obfuscated words are introduced in escort advertisements, our hope is that character models will stay invariant to these obfuscations.

Understanding images. While CNNs have proven to be useful for many different computer vision tasks, we seek to improve the learning capability of the visual network. Future direction involves using graphical modeling to understand interactions in the scene. Another direction involves working to understand text in images, which can provide more information about the subjects of the images.

Given that the current state of the art in this area generally does not use deep models, this may be a major opportunity for improvement. To this end, we encourage the research community to reach out to Cara Jones, an author of this paper, to obtain a copy of Trafficking-10k and other training data.

Acknowledgements

We would like to thank William Chargin for creating figures and revising this paper. We would also like to thank Torsten Wörtwein for his assistance in visualizing our data. Furthermore, we would like to thank our anonymous reviewers for their valuable feedback. Finally, we would like to acknowledge collaborators from Marinus Analytics for the time and effort that they put into annotating advertise-

ments for the dataset, and for allowing us to use their advertisement data.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 2425–2433.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 2625–2634.

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.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*. volume 9, pages 249–256.

National Human Trafficking Hotline. ???? Hotline statistics.

Emily Kennedy. 2012. Predictive patterns of sex trafficking online. *Dietrich College Honors Theses* .

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Lauren A McCarthy. 2014. Human trafficking and the new slavery. *Annual Review of Law and Social Science* 10:221–242.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Chirag Nagpal, Kyle Miller, Benedikt Boecking, and Artur Dubrawski. 2015. An entity resolution approach to isolate instances of human trafficking online. *arXiv preprint arXiv:1509.06659* .

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 1:34.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, pages 439–448.

Stefan Scherer, Gale M Lucas, Jonathan Gratch, Albert Skip Rizzo, and Louis-Philippe Morency. 2016. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing* 7(1):59–73.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

UNODC. 2008. Human trafficking: An overview. Web, New York. http://www.ungift.org/doc/knowledgehub/resource-centre/GIFT˙Human˙Trafficking˙An˙Overview˙2008.pdf.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pages 3–10.

Verena Venek, Stefan Scherer, Louis-Philippe Morency, Albert Rizzo, and John Pestian. 2016. Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing* .

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 4651–4659.

Zhou Yu, Stefen Scherer, David Devault, Jonathan Gratch, Giota Stratou, Louis-Philippe Morency, and Justine Cassell. 2013. Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs. In *Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*. pages 160–169.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* .

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31(6):82–88.

Andrew Jie Zhou, Jiyun Luo, and Lewis John McGibbney. 2016. Multimedia metadata-based forensics in human trafficking web data. *Vanessa Murdock, Charles LA Clarke, Jaap* page 10.