Context-specific language modeling for human trafficking detection from online advertisements

Saeideh Shahrokh Esfahani
Accenture Technology Labs
San Francisco, CA
saeideh.shahrokh@accenture.com

Michael J. Cafarella
Department of Computer Science
University of Michigan
michjc@umich.edu

Maziyar Baran Pouyan
Accenture Technology Labs
San Francisco, CA
maziyar.baran.pouyan@accenture

Gregory J. DeAngelo
Department of Economics
Claremont Graduate University
gregory.deangelo@cgu.edu

Elena Eneva
Accenture Technology Labs
San Francisco, CA
elena.eneva@accenture.com

Andrew E. Fano
Accenture Technology Labs
San Francisco, CA
andrew.e.fano@accenture.com

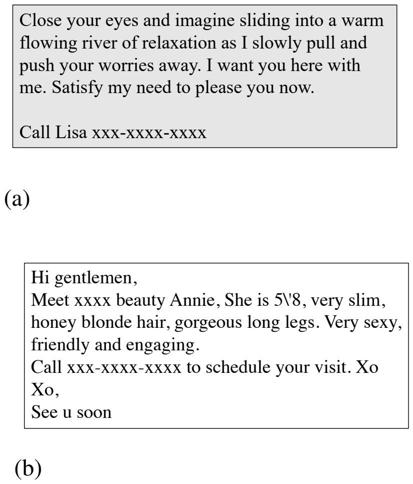
Abstract

Human trafficking is a worldwide crisis. Traffickers exploit their victims by anonymously offering sexual services through online advertisements. These ads often contain clues that law enforcement can use to separate out potential trafficking cases from volunteer sex advertisements. The problem is that the sheer volume of ads is too overwhelming for manual processing. Ideally, a centralized semi-automated tool can be used to assist law enforcement agencies with this task. Here, we present an approach using natural language processing to identify trafficking ads on these websites. We propose a classifier by integrating multiple text feature sets, including the publicly available pre-trained textual language model Bi-directional Encoder Representation from transformers (BERT). In this paper, we demonstrate that a classifier using this composite feature set has significantly better performance compared to any single feature set alone.

1 Introduction

In 2013, the Global Slavery Index reported that 30 million individuals were living in involuntary servitude. Another estimation found that 600,000 women are trafficked in the sex industry per year with the United States being the second most popular destination for these individuals (Kara, 2009); (Schauer and Wheaton, 2006). In the last decade, it has become more difficult for law enforcement (LE) to trace traffickers as they have begun to take increasing advantage of online advertisement platforms for sexual services to solicit clients and become less visible. LE is capable of tracking the posted ads and mining such data to detect trafficking victims. However, the large volume of online unstructured data, the high degree of similarity of ads (Figure 1), and the lack of an automated approach in detecting suspicious activities through advertisements present obstacles for LE to independently develop methods for surveying these criminal activities. Sex trafficking advertisements are unique texts. They have incorrect grammatical structures and misspellings, and are enriched with unconventional words, abbreviations, and emojis. Oftentimes the author uses emojis and emoticons to convey messages to a potential customer. In particular these types of advertisements may also contain equivocal words, e.g., roses as a substitute for dollars. Additionally, dominant keywords from these online ads continuously evolve as traffickers and consenting sex workers alike seek to evade prosecution. While previous researchers have tried to develop automated systems to detect trafficking advertisements, this has proved an enormous challenge for natural language processing and machine learning. In (Whitney et al., 2018), Whitney and colleagues propose to track the use of emojis and their significance in online sex ads as a potential indicator of trafficking. This team processed emojis to determine the meaning of them used

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1180–1184
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics



Close your eyes and imagine sliding into a warm flowing river of relaxation as I slowly pull and push your worries away. I want you here with me. Satisfy my need to please you now.

Call Lisa xxx-xxxx-xxxx

(a)

Hi gentlemen,
Meet xxxx beauty Annie, She is 5\'8, very slim, honey blonde hair, gorgeous long legs. Very sexy, friendly and engaging.
Call xxx-xxxx-xxxx to schedule your visit. Xo Xo,
See u soon

(b)

Figure 1: Two examples of online sex ads describing (a) a trafficking victim and (b) a non-trafficked provider, selected from our labeled ads.

in a sample of online ads, as indicated by interviews with law enforcement officials and individuals combating human trafficking. Taking a different approach, Tong, Zadeh, and colleagues (Tong et al., 2017) collaborated with LE officials and annotated 10,000 ads. With these annotated texts, they proposed the use of deep multimodal models to reach the accuracy of LE officials in identifying suspicious ads. Szekely and colleagues (Szekely et al., 2015) created a large generic knowledge graph from a large database of online sexual ads that allows for visualization and querying data.

In this paper, we present part of an ongoing project. Unlike previous studies, we tested our method on a relatively large number of ads labeled based on the corresponding phone number rather than human interpretation of the text itself. In the following sections, we propose a method relying on extracting feature sets from ads to quantify their context. We later use these feature sets in several predictive models to flag suspicious ads. We also investigate the performance of a newly released pre-trained language model called the Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018) to assess its power in analyzing this type of unstructured data.

2 Advertisement Annotation

We created a dataset of advertisement texts by crawling thousands of ads extracted from various adult websites in 2017. We then performed our analysis to a subset, only including the data from January, February and March of 2017. In order to annotate the ads in our dataset, we further extracted phone numbers from these ads leading to a set of more than 3 million distinct phone numbers. We then used a database consisting of phone numbers associated with trafficking victims, constructed in conjunction with human trafficking domain experts without direct reference to the advertising texts. Afterwards, we created a labeled data set by finding phone numbers that appear in both sets. The overlapping set contains 6,387 phone numbers, which we used to label as trafficking ads (i.e., the positive label in our precision/recall analysis). We limited our analysis to two websites, Backpage and Eroticmugshots, with 4385 ads. We selected non-trafficking's ad examples by randomly sub-sampling from the remaining ads (i.e. not labeled as trafficking) and treated them as negative examples to make a balanced 10K dataset. We assumed a very low prevalence of trafficking ads (less than 5%) in our initial set ($\approx$ 3 million phones). We discuss this decision later in the paper.

After choosing approximately 10K ads, we investigated the basic characteristics of the two labels. The median lengths of ads, including white spaces, are 538 and 401 for positive and negative labels, respectively. After excluding stop-words and lemmatizing the words, we found 24,000 distinct uni-grams in non-trafficking ads, and 9,662 distinct unigrams in the trafficking ads. It should be noted that lemmatizing was only done for calculating the statistics in this section.

3 Text Featurization

In the feature extraction step, the fundamental challenge is to quantify the textual context while retrieving information from unconventional words, abbreviations and equivocals. Here, we revisit different developed feature sets that eventually lead us to our desired contextual model.

3.1 Topic Modeling Via LDA

Our hypothesis is that language patterns, including topics and word usages, can aid in discerning the ads of trafficking victims from those of non-victims. That being said, independent or voluntary sex providers vary in their use of words, context, and topics. To test this hypothesis, we use a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). Our vision was that clustering the words, with the use of LDA to enhance the featurization, would allow us to identify the performance of words in specialized textual contexts. LDA model assigns a score based on the importance of representation of the words within each topic. Therefore, the value of assigned scores to topics indicates which ones dominate throughout the text and create the feature set as $\mathbf{s}_i = [s_{i1}, \ldots, s_{ik}]$, where $\mathbf{s}_i$ is the i-th feature vector for document i containing k scores.

3.2 Average Word Vector

We choose to use word embedding as a key part of our model. Although Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) word embeddings have shown promising results in semantic vector representations of words, when we used these models on our texts we found that they missed many of the novel word uses and abbreviations. Instead, we chose to use FastText (Bojanowski et al., 2017) for our semantic word representation, as it is based on character-level word embedding and the word representation is the sum of vectors. With that said, we hereby define the second feature set for each text as $\boldsymbol{\nu}_i = \frac{\sum_j \boldsymbol{\nu}_{i,j}}{n}$, where n is the number of words in the text i, $\boldsymbol{\nu}_{i,j}$ is the vector representation of j-th word of language model with dimension of p_ν (here set to 100 based on experiment).

3.3 Pre-trained BERT

Thus far, we have defined features which need to be trained using the ads we already had. As our next features set, we propose to use a pre-trained model. Since we believe pre-trained word embedding on general domain is not able to capture all the rare, equivocal, and abbreviated words and phrases in our sexual advertisement text (Tong et al., 2017), we are motivated in finding the most comprehensive deep learning model and chose to assess the newly released Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018). A word representation using BERT is made by using its bidirectional, i.e., left and right, context. BERT is released with two model sizes: (1) $\text{BERT}_{\text{BASE}}$ with 12 layers, 768 hidden layers and 12 self-attention heads, and (2) $\text{BERT}_{\text{LARGE}}$ with 24 layers, 1024 hidden layers, and 16 self-attention heads. One should note that in this study we do not use fine-tuned BERT model to examine the true power of BERT. Here, we choose to use the pre-trained $\text{BERT}_{\text{BASE}}$ model which encodes our document to a vector representation of size 768 for each document i and denote that by $\mathbf{b}_i = [b_{i1}, \ldots, b_{i768}]$.

3.4 Integrating LDA, AWV and BERT

Finally, we propose a new feature set consisting of the three types of features explained above. The rationale behind this composite feature set is to allow for the use of textual context as well as the simpler features. Therefore, we have the final feature vector defined as as $\mathbf{x}_i = [\mathbf{s}_i, \boldsymbol{\nu}_i, \mathbf{b}_i]$, with the dimension of $p = k + p_\nu + 768$.

4 Experiments

In our study, we employ the feature models described above and compare the results of the binary classification corresponding to them. We use logistic regression and compute the precision and recall curve (PRC) to evaluate the performance of different models. Moreover, in this application, it is important to have a model with good recall while keeping high precision, i.e., a high positive predictive value (PPV) to avoid unnecessary actions. To do so, we investigate the sensitivity of models in different high PPVs.

Pre-processing. We choose to not remove stop words or not use any stemming or lemmatization techniques as we are faced with different writing structures which could be informative for our model. We test the impact of emojis and punctuation by training and testing our model by creating two text sets. In the first text set, we keep the emojis and punctuation and remove them in the second set. In the second set, we convert the emojis to words. Numbers in the texts are removed, because: 1) we have made the labels based on phone numbers and 2), the ads are likely to have the same age or same price throughout the texts. We then divide the data into an $80/20\%$ training/testing set. In the following sections, we describe how each set of features is processed while using logistic regression as our fixed classification model.

LDA Features. We begin with features coming from LDA topic modeling scores where we assign it to 12 topics. Gensim LDA is implemented by making a bag of words dictionary of our training set. We find this optimal topic number where we examined the explained LDA feature set via cross-validation on January 2017 alone.

AWV Features. Our FastText model is trained on a set including a minimum count of 2 words and a window size of 3 to give us a vector of dimension 100. After training the FastText model, the average word vector of the training set is computed. Using this saved language model from the training set, we compute the feature test vectors.

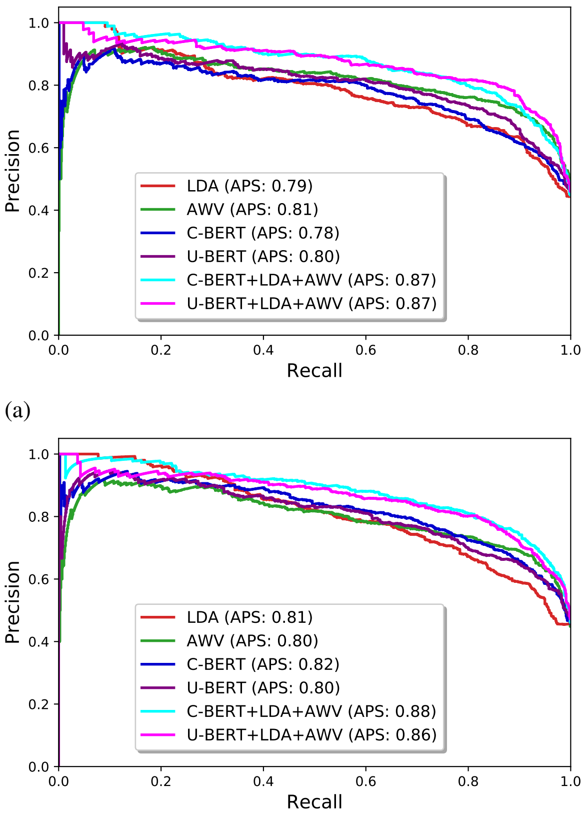
BERT Features. For encoding our texts using BERT, we make a list of all documents and use the BERT service client. We use the weights of the words that $\mathrm{BERT_{BASE}}$ learned in its pre-training to encode each document to a vector of size 768 for both the training and testing sets. We examine encoding texts with both Cased BERT (C-BASED) and Uncased BERT (U-BERT). In the U-BERT, the text has been lower cased, whereas in C-BERT, the true case and accent markers are preserved.

Full Features. In this final step in featurization towards our composite model, we combine all three types of features to build a unified feature set, i.e. combining LDA, AWV and BERT.

5 Results and Discussions

Figure 2 depicts the results of the classifications of the different feature sets. It can be seen that both classification approaches based on LDA and the average word vector features achieve similarly average precision scores (APS). Based on our analysis, keeping the entire text or removing emojis and punctuation do not significantly impact the results. From Figure 2, it can be seen that, despite small improvements, different featurizations provide similar APS values. However, focusing more on the early parts of the PRC, i.e., high precision, we can see that there is a significant improvement of recall. For example, as summarized in Figure 3, at 85% precision, our proposed full model (with U-BERT) achieves 69% and 67% sensitivity on pure text and text without emojis and punctuation, respectively. However, in the composite model with C-BERT, there is an opposite effect where recalls become 65% and 69% for the two scenarios, respectively.

Comparing to the results of the classifiers with different feature sets (under U-BERT), the model utilizing the full feature set provides 26% recall improvement over the three individual ones, i.e. 69% vs 28% – 42%, when precision is set to 85%. A similar observation holds for 90% precision. As a concluding remark, we should emphasize our significant improvement in recall rate over each individual model.



(a)

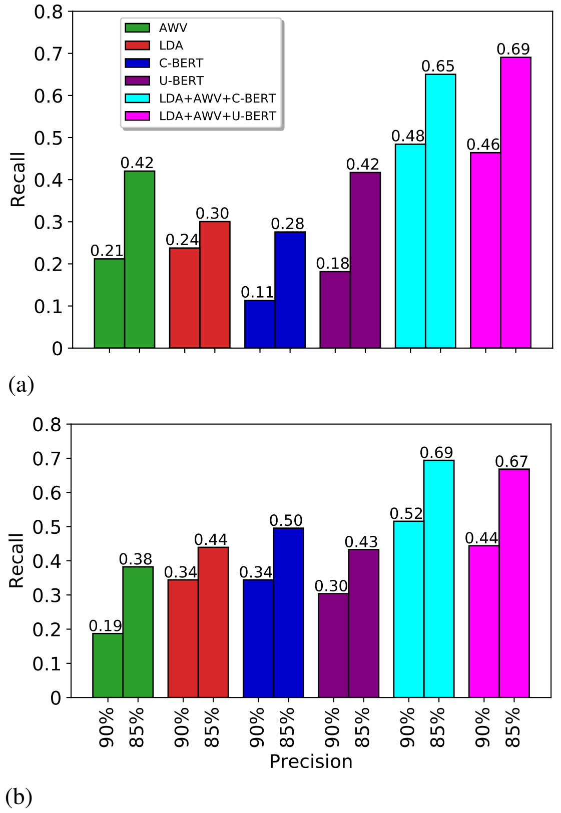
(b)

Figure 2: Precision and Recall curves (PRCs) and their corresponding APS values: (a) pure text, (b) text without emojis and punctuation.

6 Conclusions and Future Work

In this paper, we introduced different models based on different text featurizations where the main goal was to engineer features that allowed for understanding the context of sexual ads and remove the restriction of using keywords. We have proposed a composite model and compared its performance with other simpler models. For more evaluation, we examined the recall rate of models in 85% and 90% of precision. The full feature set, i.e. LDA+AWV+BERT, outperformed others as it indicated that having comprehensive features may be conveying more information about the advertisements.

Thus, we can significantly increase the PPV of our model while maintaining a high recall rate. It also should be noted that our non-trafficking examples may still contain some trafficking ads. We thus note with caution that the false positives in our model may not be truly false. Given that, in



(a)

(b)

Figure 3: Recall rates corresponding to 90% and 85% precision: (a) pure text, (b) text without emojis and punctuation.

our future work, we will be investigating those false positive cases with our collaborators to assess what the correct label for these ads should be. Moreover, since the proposed full feature set involves hundreds of features we plan to increase our sample size to have a better estimation of the performance of our final predictor. We also envision that by including other underlying components from these advertisements, we can assist law enforcement officers with an automated framework to sift millions of sexual advertisements and spend time on especially suspicious activities. Finally, in this study, we tested our model on a balanced data set. However, in the real world, the number of trafficking ads is always far lower than the number of non-trafficking ones. After collecting more labeled data, and tuning our model using anomaly detection techniques like Isolation Forests (Liu et al., 2008), we hope to expand this study to the stage where we are able to use unbalanced data sets.

Acknowledgments

This study was supported by Accenture Labs. We would like to thank Jana Thompson for critical feedback on the manuscript.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Siddharth Kara. 2009. *Sex trafficking: Inside the business of modern slavery*. Columbia University Press.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Edward J Schauer and Elizabeth M Wheaton. 2006. Sex trafficking into the united states: A literature review. *Criminal Justice Review*, 31(2):146–169.

Pedro Szekely, Craig A Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, et al. 2015. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference*, pages 205–221. Springer.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with deep multimodal models. *arXiv preprint arXiv:1705.02735*.

Jessica Whitney, Murray Jennex, Aaron Elkins, and Eric Frost. 2018. Don't want to get caught? don't say it: The use of emojis in online human sex trafficking ads.