

A Semi-Supervised Deep Learning Framework for the Automated Detection of Online Human Trafficking

Swetha Revanur, Keanu Spies
(srevanur, keanus)

December 16, 2017

Introduction

Classified advertising has facilitated the vast majority of human trafficking in the United States. The National Center for Missing and Exploited Children found that a whopping 73% of child trafficking reports occur via Backpage.com [1]. A principal reason for this is the lack of robust models to identify keywords for and authors of trafficking-related classified ads. This paper considers online human trafficking detection in a computational context.

In this study, we collect data by crawling Backpage.com and propose a semi-supervised deep learning approach to identify online advertisements that are likely to be human trafficking-related. To the best of our knowledge, this is the first study that employs domain-specific parsing and deep learning techniques to identify human trafficking-related advertisements. We thus make the following technical contributions:

1. We collected real posts from a variety of sections on Backpage.com. Custom preprocessing pipelines were built and applied to clean the data.
2. Following an extensive literature review, we performed primary feature extraction that summarize the characteristics associated with human trafficking.
3. We generated a small labeled data set by relying on expert labeling.
4. We trained various semi-supervised classifiers on the labeled and unlabeled data and validated the results with experts.

The remainder of the paper discusses the above steps in more detail.

Related Work

Recently, there has been a surge of research aimed at elucidating the relationship between the internet and trafficking markets. One of the earliest publications that leveraged computational approaches was a spatio-temporal data analysis of Backpage advertisements during the 2011 Super Bowl [2]. The authors found a stark 136% increase in the number of posts in the Adult section on Super Bowl Sunday compared to average. Another group attempted to use an entity resolution approach to isolate instances of human trafficking online [3]. Specifically, they trained a classifier to predict common authorship of posts, relying heavily on repeated phone numbers. This classifier was then applied to deduce real world entities from digital ones. Nevertheless, prior work does not address the fundamental issue faced by law enforcement officials: determining whether a post is human trafficking-related or voluntary prostitution/benign (neither). This is the challenge we tackle here.

Infrastructure

Data Source

In recent months, major police raids have made it evident that Los Angeles is a hotbed for human trafficking [4]. Thus, the data set is scoped to Backpage posts from the greater Los Angeles area. At the time of aggregation, this encompasses 2738 top posts spanning 7 relevant categories (Dating, MenSeekMen, MenSeekWomen, WomenSeekWomen, WomenSeekMen, Transgender, and Massage) over one month.

Data Scraping

With a high, concentrated load, Backpage’s servers crash. To address this, batch processing was imple-

Every URL in each batch was followed, and the corresponding post’s ID, original text, title, date, location, phone number, and category were parsed from the HTML.









Preprocessing

1. Strip HTML tags: All stray HTML tags and symbols were removed.
2. Casefolding: All alphabetic characters were converted to lowercase script.
3. Expand contractions: A comprehensive dictionary of common English contractions and their corresponding expansions was constructed and applied to the data set.
4. Strip phone numbers: Phone numbers often appear multiple times in the text, so after initial extraction, duplicates were removed.

5. Strip punctuation: All punctuation was removed.
6. Removal of one-character-long words: All one-character-long words were removed.
7. Emoji tokenization: Emojis hold much of the hidden meaning in human trafficking-related Backpage posts [9], so retaining them is key in the preprocessing step. However, emoji parsing is a nascent space, and so custom parsing pipelines were built from scratch to encode emojis and preserve them despite any punctuation and character cleaning.
8. Stop word removal: Stop words are frequently appearing words (such as “the” or “and”) that don’t contribute to the meaning of a text. A comprehensive list of English stop words was constructed and used to filter each post.

Feature Engineering

| Feature Group | Features |
|-----------------------|---|
| Language Pattern | third-person voice, first-person plural pronouns, Shannon entropy to assess text complexity, n -grams with TF-IDF ($n = 3$) |
| Keywords | [young, fresh, new, new in town, new arrival, open minded, petite, exotic, youthful, barely legal, virgin, tiny, incall, in call, new to the game, candy == 1] |
| Countries of Interest | [china, vietnam, korea, thailand, asian == 1] |
| Multiple Victims | [girls, women, men, boys, people, children, babes, dolls, masseuses == 1] |
| Victim Weight | [victim weight ≤ 110] |
| Spa Reference | [spa, massage == 1] |
| Presence of Emojis | [ ,  ,  ,  ,  ,  == 1] |

These 7 groups encompass 19 binary features that are calculated as follows.

The first group of features assesses style and language patterns. For the first and second features, respectively, we identified posts with third-person language (likely to be written by

someone other than the victim) and posts which contain first-person collective pronouns such as “we” and “our” [5]. Both of these features indicate shared management situations.

To ensure anonymity and avoid discovery by human analysts and programs, traffickers often introduce artificial complexity into their posts [6]. To account for this, we extended the notion of Kolmogorov complexity from the field of complexity theory. Kolmogorov complexity is the length of the shortest program capable of reproducing the post on a universal machine [7]. We employed Shannon entropy to estimate the Kolmogorov complexity as shown below, where X is the cleaned post content and x_i is a word in the post.

$$H(X) = \sum_{i=1}^n P(x_i) \log_2 P(x_i)$$

Higher values of $H(X)$ should correspond to human trafficking.

We used word-level n -grams ($n = 3$) to explore common phrases in the posts. Specifically, normalized n -grams was applied along with a term frequency-inverse document frequency (TF-IDF) metric to inform how important each word was to a corpus. Both the entropy value and n -grams results were binarized.

Keywords

After a thorough literature review, certain words were found to be particularly tied to human trafficking: young, fresh, new, new in town, new arrival, open minded, petite, exotic, youthful, barely legal, virgin, tiny, incall, in call, new to the game, and candy. The presence of any one of these words produces a positive feature.

Countries of Interest

Nearly two-thirds of human trafficking victims are from East and Southeast Asia [8]. Therefore, the presence of any one of the words from the set {china, vietnam, korea, thailand, asian} produces a positive feature.

Multiple Victims

The mention of more than one girl in a post was considered evidence of shared management and organized human trafficking.

Victim Weight

Several Backpage posts make references to the weight of the victim. Body weight under 110 pounds is most common in younger girls, and is therefore considered a sign of human trafficking.

References to Spa or Massage Facility

Most human trafficking channels flow through spas and massage facilities. Therefore, the presence of any one of the words from the set {spa, massage} produces a positive feature.

Presence of Emojis

The most prominent emoji indicators of trafficking are the rose, rosette, cherry, cherry blossom, growing heart, airplane, and crown. For example, the growing heart and cherry convey that the victim is underage, while the airplane means the victim is international [9]. The presence of any one of those emojis produces a positive feature.

Following feature extraction, we were able to represent each post as a 19-dimensional binary feature vector.

Unsupervised Filtering

We obtained 1857 records (hereafter referred to as the filtered data set) from our original data set by removing out samples that do not possess any of the binary features (i.e. have little to no sexually-charged content). The word cloud below was created from the filtered data set.



Figure 3: Word cloud from filtered data set

To verify the fidelity of our filtering we applied the t-SNE transformation to create 2D projections of the filtered and junk (other) feature vectors. We clustered these projections using K -means ($K = 2$). We can see a separation between the filtered and

junk data sets, but also separation within the filtered set, indicating that finer classification is possible.

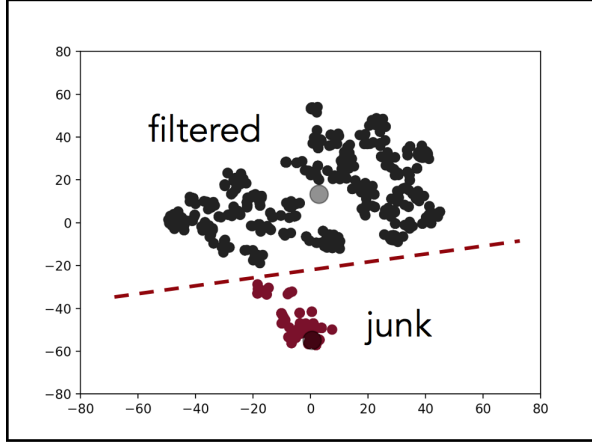


Figure 4: K-means visualized

Phase II: Classification

Expert Labeling

The primary reason that machine learning has not been actively pursued in the human trafficking domain is the dearth of labeled data. 300 of the 1857 records were randomly selected and manually labeled (hereafter referred to as the labeled data set). The other 1557 records make up the unlabeled data set. We have now reduced our task to a semi-supervised binary classification problem (semi-supervised due to the presence of both labeled and unlabeled data). Specifically, the classes are human trafficking (1) and voluntary prostitution/benign (0).

Baseline and Oracle

The baseline was a majority algorithm for post classification. The majority algorithm classified all the examples in the testing set as the majority class of the training set. On the other hand, the oracle was the manual labeling of the entire filtered data set and served as an upper-bound on the performance of our approaches.

Feature Extraction with Topic Modeling

A secondary feature set was extracted using Latent Dirichlet Allocation (LDA). LDA is a probabilistic

model, in which each word in a corpus is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities [10]. In this context, LDA identified the 7 most representative topics from the filtered data set. The optimal number of topics for LDA was tuned by evaluating the metrics proposed by Griffiths, Cao et al., and Arun [11].

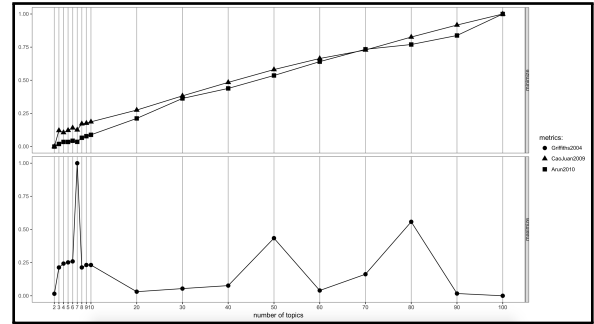


Figure 5: LDA hyperparameter metrics

The LDA document-topic distributions make up the 7-dimensional continuous feature vector for each post that is used in the remainder of this work.

Label Spreading

One of the most popular semi-supervised classification algorithms is label spreading, a type of label propagation [12]. Each record of the filtered data set V represents a node $n \in V$ in graph $G = \{V, E\}$. Here, E is the similarity between nodes represented as a weight matrix W . In label spreading, labels are propagated through G . Label spreading was implemented on two kernels: radial basis function (RBF) and K-nearest neighbor (KNN). The RBF kernel will produce a fully connected graph which is represented as a dense matrix. The KNN kernel will produce a sparse matrix which can reduce run time. The confusion matrices for both kernels are below.

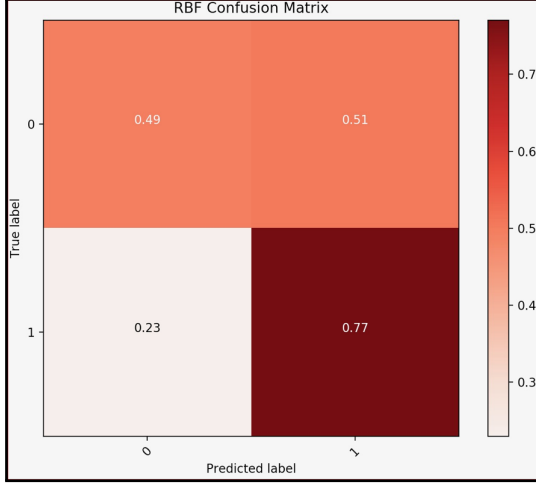


Figure 6: RBF confusion matrix

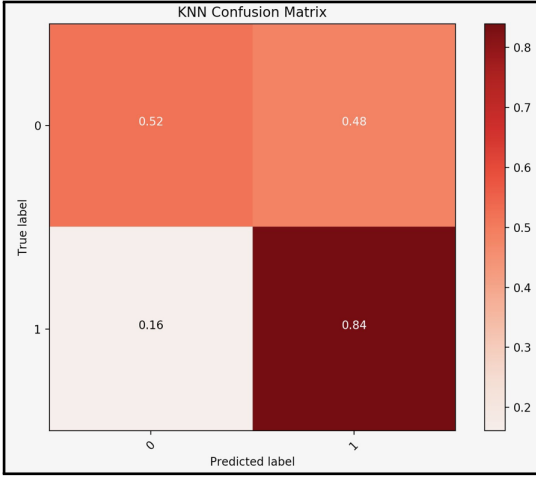


Figure 7: KNN confusion matrix

Deep Ladder Networks

Label spreading was unable to achieve a desirable accuracy (see *Evaluation*), since it historically doesn't perform well in situations with a non-uniform distribution of labels and uniformity is not guaranteed here. So, deep neural networks were explored next given the complexity of the data set with emojis and human trafficking-specific language patterns. In the past, deep learning has been applied to either fully unsupervised or fully supervised problems. Deep ladder networks (DLNs) are a deep neural architecture that combine both [13]. DLNs are trained to simultaneously optimize the sum of supervised and unsupervised cost functions via backpropagation.

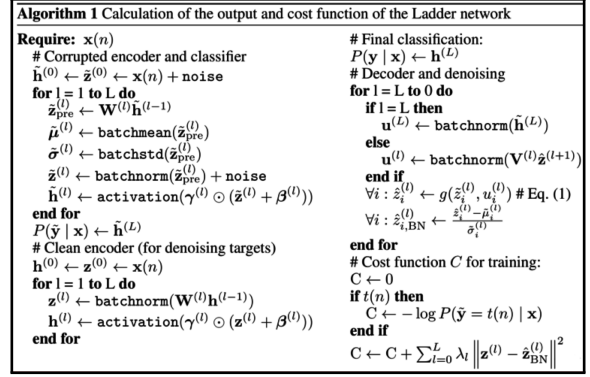


Figure 8: DLN algorithm

DLNs have a 4-phase implementation [14]:

1. Take a feedforward neural network that will handle the supervision. The network consists of clean and corrupted encoder channels. Both behave the same, except that the corrupted encoder introduces Gaussian noise.
2. Take a decoder which handles the unsupervised learning. The decoder uses a denoising function to reconstruct the activations of each layer given the corrupted version. The target at each layer is the clean version of the activation and the difference between the reconstruction and the clean version serves as the denoising cost of that layer.
3. The supervised cost is computed using the output of the corrupted encoder and the output target. The unsupervised cost is the sum of the scaled denoising cost at every layer. The total cost is the sum of the latter two.
4. Use stochastic gradient descent to optimize this total cost.

The DLN classifier was implemented using TensorFlow.

Evaluation

We report the performance of each of the various approaches.

| Approach | Precision | Recall | Accuracy |
|-----------------------------|-----------|--------|----------|
| Baseline Majority Algorithm | --- | --- | 56.667% |
| Oracle | 100% | 100% | 100% |
| Label Spreading (RBF) | 64% | 64% | 63% |
| Label Spreading (KNN) | 70% | 68% | 67% |
| Deep Ladder Networks | --- | --- | 80% |

Our DLN-based learner assigned positive labels to 54% of records in the unlabeled data set. The predicted labels were sent to experts for validation. The DLN classifier achieved 80% accuracy. *Figure 9* is a Backpage post correctly classified as human trafficking. *Figure 10* is a post correctly classified as voluntary prostitution/benign. Identifying information was censored for privacy reasons.

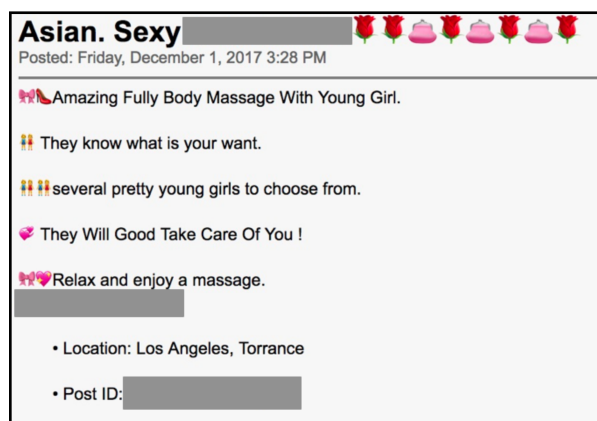


Figure 9: Correctly predicted human trafficking post

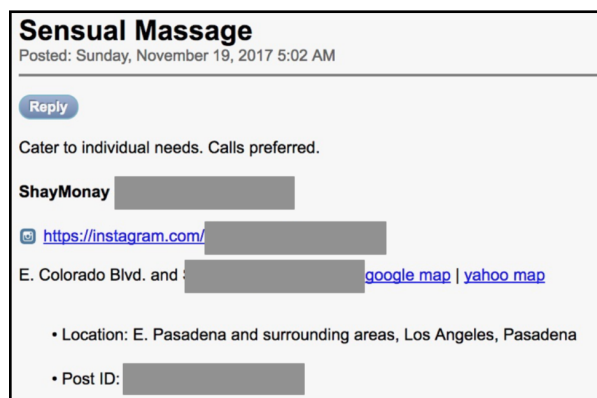


Figure 10: Correctly predicted benign post

Conclusions

Publicly-available online data from escort advertisements can be leveraged to fight and stop human trafficking. In this study, we first developed a comprehensive pipeline to programmatically fetch and clean Backpage data. As far as we are aware, this pipeline includes the most exhaustive and advanced emoji parsing functionality to date. Then we discussed an unsupervised filtering technique and subsequently trained a variety of semi-supervised classifiers on a filtered data set. We found that

the deep ladder networks achieved the best performance, with an accuracy of 80%. The results indicate that our approach has promise when it comes to identifying human trafficking online. This work is available as a Codalab worksheet at <https://tinyurl.com/dl4humantrafficking>.

Further Work

In the future, we seek to perform graph analysis to predict pimp clusters and determine authorship of Backpage posts. This work was specific to the Los Angeles area, but we would like to extend it to other major cities across the country as well and provide a larger training set for the DLN.

References

- [1] A. O'Reilly, "Prostitution still thrives on Backpage despite site shutdown of 'adult' section". *Fox News*. May 2017.
- [2] M. Latonero, "Human Trafficking Online: The Role of Social Networking Sites and Online Classifieds". *USC Annenberg*. Sep 2011.
- [3] C. Nagpal *et al.*, "An Entity Resolution approach to isolate instances of Human Trafficking online". *Carnegie Mellon University*. Jun 2017.
- [4] H. Blume, "More than three dozen arrested in alleged sex-trafficking ring in Compton". *Los Angeles Times*. Sep 2017.
- [5] E. Kennedy, "Predictive Patterns of Sex Trafficking Online". *Carnegie Mellon University*. Apr 2012.
- [6] H. Alvari and P. Shakarian, "A Non-Parametric Learning Approach to Identify Online Human Trafficking". *Arizona State University*. Aug 2016.
- [7] L. Fortnow, "Kolmogorov Complexity". *University of Chicago*. Jan 2000.
- [8] O. Enos, "Nearly Two-Thirds of Human Trafficking Victims Are from Asia". *The Daily Signal*. Nov 2014.
- [9] J.V. Grove, "Emojis and sex trafficking: SDSU researchers crack the code". *The San Diego Union-Tribune*. May 2017.
- [10] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation". *Journal of Machine Learning Research*. Jan 2003.

- [11] M. Nikita, “Select number of topics for LDA model”. Oct 2016.
- [12] “sklearn.semi_supervised.LabelSpreading”. *scikit-learn*.
- [13] A. Rasmus *et al.*, “Semi-Supervised Learning with Ladder Networks”. *The Curious AI Company, Finland*. Nov 2015.
- [14] R. Boney, “Introduction to Semi-Supervised Learning with Ladder Networks”. Jan 2016.