CS221 Project Progress Report - Distributed Word Representations for the Automated Detection of Human Trafficking

Swetha Revanur, Keanu Spies
(srevanur, keanus)

November 19, 2017

I. 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. 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.

The feature space for the task of human trafficking detection is largely undefined. The Distributional Hypothesis states that words that appear in the same contexts share similar semantic meaning. Here, we use vector space models (VSMs) to exploit this theory and optimize feature engineering. VSMs embed words in a continuous vector space in such a way that encodes the semantic information of the texts. We propose the use of these word embeddings to enable keyword identification and authorship attribution of Backpage posts.

II. Infrastructure

Data Source

In recent months, major police raids have made it evident that Los Angeles is a hotbed for human trafficking. Thus, the data set is scoped to Backpage posts from the greater Los Angeles area. At the time of aggregation, this encompasses 6466 top posts spanning 22 randomly selected categories and one month.

Data Scraping

With a high, concentrated load, Backpage's servers crash. To address this, batch processing was implemented. A web crawler was built and deployed to recursively collect URLs for the nearly 6500 posts. The URLs were then split into batches of size 100. Between each batch process, a two-minute wait time was forced.

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.

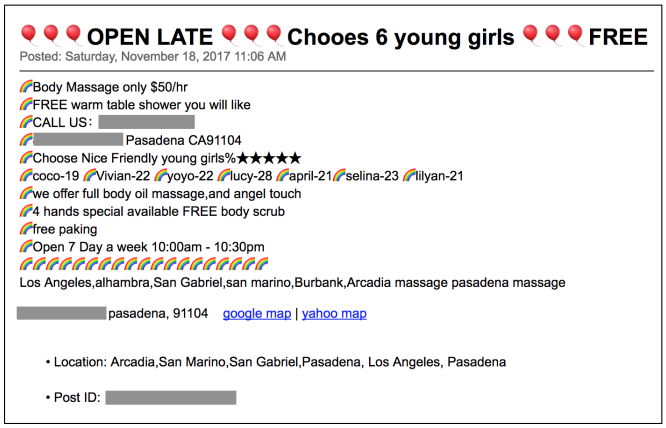

Figure 1: Sample Backpage.com post

In addition, some posts have an "Other ads by this user" section. Post IDs from this section were grabbed. However, exploration of the data showed that identical posts were sometimes made from different accounts and posts were often deleted and re-posted, so the "Other ads by this user" section alone was deemed insufficient for complete automated authorship attribution.

Preprocessing

As shown in Figure 1, posts are frequently written informally and are largely unstructured. To normalize and reliably extract relevant textual data, preprocessing was conducted. Preprocessing was a six-stage procedure:

1. Casefolding: All alphabetic characters were converted to lowercase script.
2. Contraction expansion: A comprehensive dictionary of common English contractions and their corresponding expansions was constructed and applied to the data set.
3. Stop word removal: Stop words are frequently appearing words (such as "the" or "a") 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.
4. Punctuation cleaning: All punctuation was removed.
5. Duplicate character removal: After examination, it was observed that Backpage posts often contain mistyped and non-standard words that have duplicate characters (such as "queeeeeeen"). These duplicate characters were removed.
6. Leetspeak translation: Leetspeak is an alternative alphabet that uses some characters to replace others in ways that play on the similarity of their glyphs via reflection or other resemblance. For instance, "h3ll0" was converted to "hello".
7. Emoji parsing: Emojis hold much of the hidden meaning in the human trafficking-related Backpage posts, so retaining them is key in the preprocessing step. Emoji characters are encoded at a byte level and preserved despite any punctuation and character cleaning.

III. Methods

Baseline

The baseline model solely uses information from the *otherAds* attribute. All posts linked to a single user account are assigned to the same cluster, and those coming from different accounts belong to separate clusters. However, this does not account for ads authored by individuals using multiple accounts or capture linguistic and syntactic patterns specific to human trafficking ads, and is therefore a lower bound on our model's performance.

Oracle

On the other hand, the oracle is a manual tagging of ads that reflect human-distinguishable differences in writing style as well as high trafficking likelihood. The oracle captures multi-user account authorships, enabling documents to participate in more complex relationships compared to the simpler baseline. Ultimately, the aim is to tune a machine learning approach so that it approaches the oracle's upper bound.

Approach

The Distributional Hypothesis is leveraged with a predictive neural probabilistic language model. Specifically, we consider Word2vec, which is a particularly computationally-efficient predictive model for learning word embeddings from raw text. The Skip-Gram version of Word2vec predicts source context-words from the target words.

With skip-gram, we train the model using a logistic regression classification paradigm to discriminate the real target words w_t from k noise words w_n, in the same context. The objective function is:

$$J_{NEG} = log Q_\theta(D = 1 \mid w_t, h) + k \cdot \mathbb{E}_{w_n \sim P_{noise}}[log Q_\theta(D = 0 \mid w_n, h)]$$

where $Q_\theta(D = 1|w, h)$ is the binary logistic regression probability of observing w in the context h in the data set D, calculated in terms of the learned embedding vectors θ.

The objective function is defined over the entire data set, but we can optimize this with stochastic gradient descent using a minibatch technique.

The goal is to make an update to the embedding parameters θ to maximize the objective function. We update the embeddings by taking a small step in the direction of the gradient. Intuitively, this has the same as 'moving' the embedding vectors around for each word until the model can successfully differentiate real words and noise words.

We visualized the skip-gram results by projecting the word embeddings onto a two-dimensional

plane using the t-SNE dimensionality reduction technique as shown below.

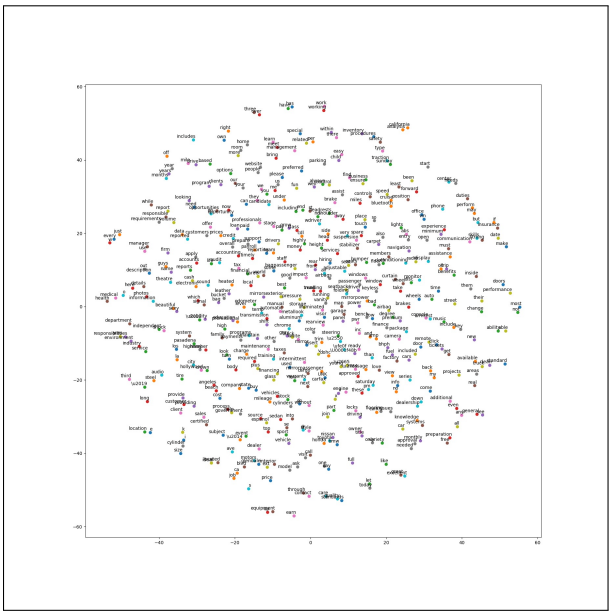

Figure 2: Word vector projections

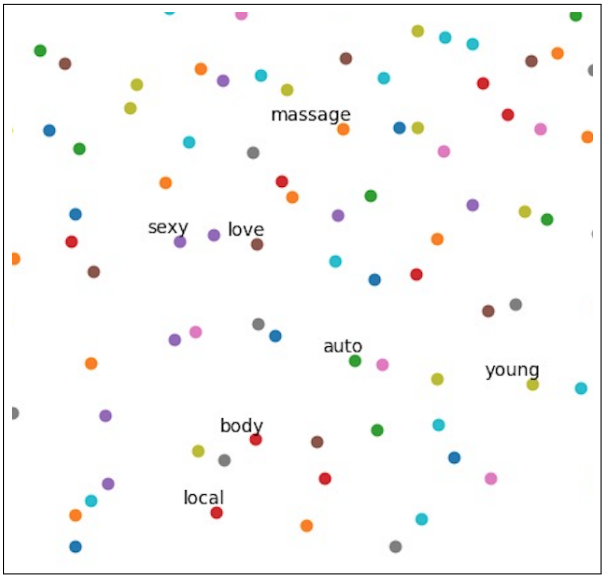

Figure 3: Cluster of semantically similar words

IV. Evaluation

During training, we compute the noise-contrastive estimation loss. After 50000 training iterations, the NCE loss converged to 2.52. To evaluate embeddings, we directly used them to predict syntactic and semantic relationships, in a technique called analogical reasoning.

V. Future Work

We begin with a minimal set of known features related to human trafficking. Word2vec's semantic clusters enable the discovery of previously unknown features. From the significance of the features, we bubble up to score each document.

We then want to score the relationship between all pairs of documents. This edge weight between documents A and B will be high if they are likely to be written by the same author. Computing edge weights creates a graph, from which we can use graph analysis to determine authorship.