Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data

Anastasija Mensikova[1], Chris A. Mattmann[1,2]
chris.a.mattmann@jpl.nasa.gov

[1]Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109 USA

[2]Computer Science Department
University of Southern California
Los Angeles, CA 90089 USA

ABSTRACT

We motivate the use of sentiment analysis as a technique for analyzing the presence of human trafficking in escort ads pulled from the open web. Traditional techniques have not focused on sentiment as a textual cue of human trafficking and instead have focused on other visual cues (e.g., presence of tattoos in associated images), or textual cues (specific styles of ad-writing; keywords, etc.). We apply two widely cited sentiment analysis models: the Netflix and Stanford model, and we also train our own binary and categorical (multi-class) sentiment model using escort review data crawled from the open web. The individual model performances and exploratory analysis motivated us to construct two ensemble sentiment models that correctly serve as a feature proxy to identify human trafficking 53% of the time when evaluated against a set of 38,563 ads provided by the DARPA MEMEX project.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Applied computing** → *Computer forensics*;

KEYWORDS

Sentiment Analysis, Human Trafficking, Machine Learning, Feature Identification, Information Retrieval

ACM Reference Format:
Anastasija Mensikova[1], Chris A. Mattmann[1,2]. 2018. Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data. In *Proceedings of ACM Workshop on Graph Techniques for Adversarial Activity Analytics (GTA[3] 2018).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Human trafficking is a global concern involving organized crime, child prostitution, forced labor and servitude, and by 2015 has largely become a household name [5]. The prevalence of human trafficking has also democratized its presence in digital mediums and it is clear that the Internet has become a home for the proliferation of trafficking and for conducting trafficking as a business. Web sites including Backpage.com [7] have been widely used as a digital marketplace for predators and pimps to traffic victims through solicitation of services, especially in the area of sex-trafficking.

Our team was a part of the recently concluded DARPA MEMEX effort which began in 2014 and during its three year timespan produced a comprehensive web corpus and understanding at scale of human trafficking on the web, including ads for escorts, images, videos, and analytics and extraction processes to identify people, places, and things across the variety of digital content present on the web. Our work helped produce indices and web corpora including 80 million web pages and 40 million images, and in this mountain of data, the MEMEX team worked with law enforcement customers helping to mine the data and to provide evidence and information that saved the lives of trafficking victims.

One of the key contributions of MEMEX was the production of ground-truth datasets and in particular, sets of web documents (including images, videos, etc.) wherein which those referenced in the ads were positively or negatively identified as victims of trafficking through our law enforcement partners. Performers within the MEMEX team worked on classification techniques that took this ground truth and derived important features: from visual cues in associated multimedia; to vocabulary choice in the ads, to extracted features about the persons, places, and things in the ads, etc.

We were particularly interested in mining the text associated with the ad, and also in the area of sentiment analysis [8]. Sentiment analysis of web data is an approach to discern the text writer's affinity or negativity as expressed through her use of language and vocabulary. Sentiment can be binary (e.g., positive and negative, or love and hate) or categorical/multi-class (e.g., love, like, neutral, sad, hate, etc.) and can serve as a data reduction proxy for large bodies of text, social media data, etc. We are not aware of broad studies of sentiment analysis as it applies to the area of human trafficking, and our hypothesis was that sentiment analysis could be an important textual cue indicating a web document's potential to describe an actual trafficking scenario. Sentiment analysis could also provide a window into the mindset of both the person writing the ad - the potential predator or pimp; or even the victim.

In this paper, we describe a series of experiments to apply sentiment analysis as an indicator for human trafficking. We applied existing binary e.g., Netflix [9] and categorical e.g., Stanford Treebank [15] sentiment models directly to subsets of web ads from our MEMEX human trafficking corpus for which ground-truth regarding human trafficking was available. We also trained our own binary and categorical human trafficking models using the ad text

GTA[3] 2018, February 2018, Marina Del Rey, CA USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

processed with information retrieval techniques including text-to-tag-ratio (TTR) [17] to isolate the important ad text in the page. We also trained two ensemble sentiment analysis models that used two categorical models, and incorporated additional features including geographic location as identified through named entity extraction [11] and the presence of negation in the text as additional cues. The ensemble models outperformed individual models in accuracy and number of iterations required to converge.

Section 2 highlights background and related work in the area. Section 3 identifies the models and data we used off the shelf and trained for these experiments. Section 4 describes our approach for building HT sentiment models and Section 5 describes our evaluation of the off-the-shelf and custom models against the MEMEX ground truth. Section 6 summarizes our contributions, points to future work and concludes the paper.

2 BACKGROUND AND RELATED WORK

Boiy and Moens [1] perform sentiment analysis on open web data using Apache OpenNLP. They focus on training three-class (positive, negative, and neutral) sentiment related to consumer products on text parsed from blog, review and forum sites. The authors were able to leverage their approach on English, Dutch and French text, with 83% accuracy on the English texts. In their text analysis, the authors leverage unigram features - similar to our own approach. Unlike our approach, however, the authors do not mention how they isolated the appropriate text from the blog, review and forum sites, and performed HTML pre-processing, or if they used Text-to-Tag (TTR) techniques [17] as we employed. Others have also used Apache OpenNLP for sentiment analysis including the Elixa project [16], the work by Wogenstein et al. [18] and Johnson et al. [4]. As noted by Paltoglou, no models are widely available however for sentiment analysis using Apache OpenNLP [14]. Our work contributes freely available Apache OpenNLP sentiment models, available at [12].

Bouazizi and Ohtsuki [2] build a multi-class sentiment analyzer on Twitter data using pattern recognition to go from binary sentiment to categorical sentiment, similar to our own approach. The approach from Bouazizi and Ohtsuki is 56.9% accurate on a seven class identification framework from Tweets, performing better on neutral and sarcastic classes. The authors suggest their work performs well on binary, and ternary class sentiment (positive, neutral and negative), achieving accuracy rates in the eighty three percentile. The work by Bouazizi and Ohtsuki performs part of speech tagging, and uses pattern/word replacement techniques to pre-process the textual data. On the other hand, our technique simply uses Text-to-Tag ratio, and is focused on HTML data, giving us more labeled samples, and the need to remove formatting - whereas the Tweets are already extracted into raw text form. We use proxy features related to human trafficking rather than Part-of-Speech tagging and word replacement. Our approach also differs in that we do not directly assess sentiment accuracy and instead use sentiment as a proxy feature for HT *RELEVANT* and *NOT RELEVANT* classification. Although we do not look at sarcasm, other approaches to sentiment analysis such as Khodak et al. [6], do. They provide an annotated corpus for sarcasm analysis in text via sentiment analysis.

3 APPROACH

In researching how to apply sentiment analysis to our DARPA MEMEX human trafficking related web pages, we discerned two prevailing model categories: (1) binary sentiment - positive/negative; and (2) categorical, or multi-class sentiment - e.g., like, love, neutral, etc. We began our process by examining widely cited and used binary and categorical models for sentiment analysis that are either pre-trained on other existing web and social media data, or that provide the original raw data and allow the user to perform her own training. We narrowed our search down to the Netflix binary sentiment [9] and Stanford Treebank categorical sentiment [15] models.

The MEMEX program also contributed two datasets of Human Trafficking related data. The first dataset that we call *HT ground truth*, is a set of web ads from Backpage.com in which the web pages have an associated *RELEVANT* or *NOT RELEVANT* label associated with them to indicate that the ad contains victims of human trafficking (*RELEVANT*) or not (*NOT RELEVANT*). The second MEMEX dataset, *HT provider reviews* include reviews culled from websites in which consumers of escort-services from Backpage.com and other sites write reviews about the providers of escort-services, akin to the way that a user would write reviews of e.g., a product they purchased on a consumer site like Amazon. These *HT provider reviews* offer insight and textual cues about potential human trafficking victims, and also offer a *rating* or score, to describe the escort provider's "quality" of service as provided by the consumer of it. Both of the MEMEX datasets are law enforcement sensitive, and not public, however we will describe their overall characteristics in this section, along with the public datasets already discussed.

In addition, we categorize our datasets as *training* datasets - built from either public sources including Netflix and Stanford Treebank or the *HT provider review* MEMEX data - and *testing* datasets - the MEMEX *HT ground truth* data, divided into pre-labeled ads, and ads without labels. To be clear, *training* datasets yield sentiment models that we evaluated against *testing* datasets to measure binary or categorical sentiment features as predictors of human trafficking.

3.1 Training: Netflix Binary

The Netflix dataset is a 33.1 MB; 25,000 records. The dataset was collected through the source cited. The source provides the positive and negative records in two separate directories, which allowed us to give each text record a respective label (*positive* or *negative*). We combined the two directories together into one training dataset. Although this dataset is not human trafficking specific, it is widely used in the machine learning community and a good basic model for sentiment based on web-based textual cues and ads.

3.2 Training: Stanford Categorical

The Stanford Treebank data is 12 MB; 239,232 records. This dataset was collected through the source cited. Each record is in JSON format and was given a numerical score from 0 to 1 linking textual features to a score. We used the score and divided the data into distinct categories akin to Facebook's *reaction* recently released sentiment features. Facebook reactions are multi-class sentiment labels to attach to text. We use Stanford's Treebank in a similar categorical sentiment analysis. We took all the reviews with a score

of <= 0.2 and then labeled them as *angry*, those reviews with the score of > 0.2 and <= 0.4 as *sad*, others with > 0.4 and <= 0.6 as *neutral*, text reviews with > 0.6 and <= 0.8 as *like* and finally reviews with the score > 0.8 and <= 1.0 as *love*. Though this dataset was not human trafficking specific it remains widely cited and so this became our baseline categorical sentiment model.

3.3 Training: Human Trafficking (HT) Provider Review Binary / Categorical

The Human Trafficking (HT) Provider Reviews data is 1.9 MB; 1,056 records. This is a private dataset including a subset of provider reviews from various websites. The review web page raw content (HTML) and extracted text is stored in a JSON file along with web page metadata, and a score provided by the escort reviewer. We used the page extracted text and the score of each review, which ranged from 0 to 1. We trained two models. The first, a binary model assessed sentiment as *negative* if the score was <= 0.5 and labeled the rest as *positive*. We also built a categorical model using the same score discretion used for the Stanford Treebank dataset already mentioned with the labels *angry, sad, neutral, like love*.

3.4 Training: HT Ground Truth Binary

The HT Ground Truth binary dataset is 31.3 MB; 22,246 records. The labels in those model are not sentiment related per-se, but correspond to whether each escort ad in the dataset model trained on this dataset was describing an escort involved in human trafficking and was thus labeled as *RELEVANT* or if not, *NOT RELEVANT*. The dataset consisted of 11,123 *RELEVANT* and 11,123 *NOT RELEVANT* ads. All ad HTML was extracted and relevant text isolated using the Text-to-Tag ratio algorithm [17] to get rid of unnecessary and unimportant HTML and remove bias in our models. Though the dataset did not provide a direct correlation to sentiment, it was used as a proxy component in our ensemble models described later in this section.

3.5 Training: HT Ensemble Binary I and II

Taking the same HT Ground Truth dataset, and after looking at model performance for our HT Categorical training data, we manually read through a few of the test ads labeled as *RELEVANT*, and found that our Stanford and HT Provider categorical models classified these ads with sentiment "love". Further, as we will explain in the ensuing section, we found that many of these ads had a specific geolocation mentioned[10] e.g., Las Vegas, and also included the presence of negative words ("negation"). These collective features became reoccurring cues within *RELEVANT* labeled HT ground truth data. Based on this analysis, we trained two new binary models. The first model, *HT Ensemble Binary I* included as features: (1) result of *HT Provider Review Categorical* sentiment, either *love* or not; (2) geolocation as identified by running our prior GeoTopicParser classifier [10] on the ad text which provides a best fit geolocation and a set of N alternate locations, with confidence; and (3) a flag indicating whether negation text was detected or not using a simple regex classifier. In examining the results of building *HT Ensemble Binary I* and later iterations, we built an *HT Ensemble Binary II* model that only considered the first five geolocations output in confidence order from the GeoTopicParser [10].

3.6 Testing: HT Ground Truth - pre-labeled

The HT Ground Truth - pre-labeled - included 22,246 ads. This is the pre-labeled dataset that was used to train the HT Ground Truth Binary and Ensemble Binary I, and II models. Of course in those cases, we withheld this dataset for testing it accuracy and model performance. However, for the Netflix binary, categorical; Stanford binary, categorical; and HT Provider Review binary and categorical model the HT ground truth served as a useful testing dataset, helping us to identify the need to create ensemble models and providing ground-truth to establish correlations between sentiment and human trafficking.

3.7 Testing: HT Ground Truth - label withheld

We used a different subset of MEMEX HT ground truth ads, amounting to 38,563 ads in JSON format in which we withheld the HT *RELEVANT* and *NOT RELEVANT* labels.

4 APPROACH

Our overall approach is divided along our training models and testing models. The first part of our approach were to explore correlations between off-the-shelf sentiment models Netflix, Stanford Binary/Categorical, and HT Provider Review Binary/Categorical and their outputs and the ground truth labels from the HT ground truth dataset. The steps of this portion of the approach are shown in Figure 1.

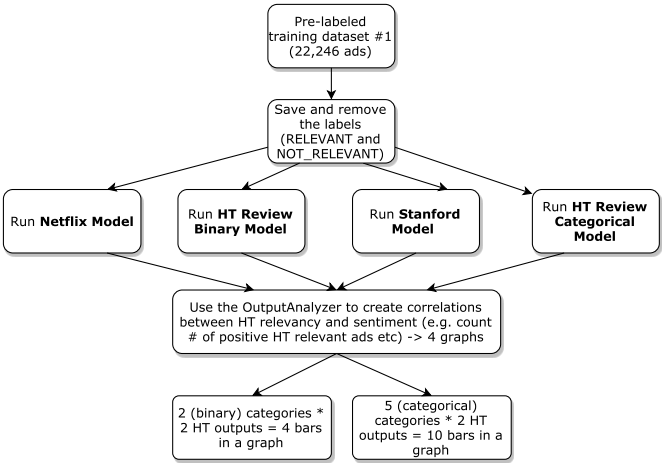

Figure 1: Evaluating trained off-the-shelf sentiment models and HT Provider review models

Using our training model approaches, and HT ground truth training data we used the Apache OpenNLP toolkit [13] to generate classifiers for each of the Netflix, Stanford Binary/Category, HT Provider Review Binary/Categorical models. Apache OpenNLP is an open-source machine learning based toolkit that allows for processing of natural language text. OpenNLP provides the user with various commonly required and used NLP tasks, as well as maximum entropy and perception based machine learning.

The training dataset for each model consists of multiple lines, where each line represents one training element and starts with

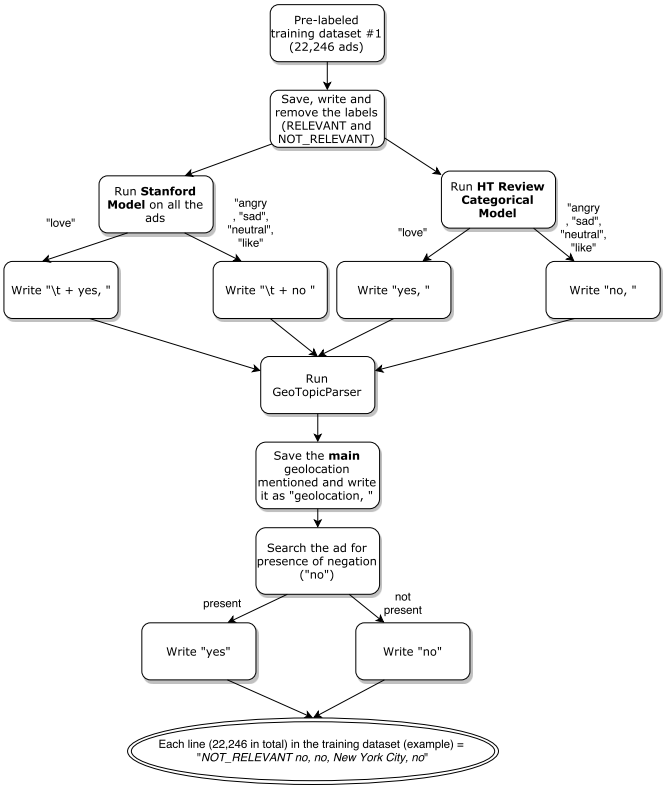

Figure 2: Ensemble Sentiment Model Generation Approach

the category label (due to the supervised nature of the algorithm), followed by a space, the training text and the new line character. Generally, depending on the situation, the more examples there are for each label, the more reliable and accurate the model could be considered. Apache OpenNLP performs 100 training iterations for every model and provides accuracy for each one of the iterations. The accuracy is calculated by dividing the number of correctly predicted events by the total number of events (numCorrect/numEvents).

The very first part of our analysis uses the HT ground truth prelabeled ($RELEVANT$, $NOT RELEVANT$) dataset, runs the training models and calculates the number of positive and negative ads per each of the two initially given labels. The goal of this portion of the approach is to analyze the distribution of sentiment among the HT-relevant and not relevant ads and create correlations. This analysis gives us some basic information on the performance of the models and how they differ depending on many factors, as well as the overall distribution of sentiment among ads - the full distributions can be viewed at [12].

We found that the Netflix model though trained on non human trafficking web data did classify the majority of *RELEVANT* ads as negative and *NOT RELEVANT* ads as positive. The observed trend could be explained by the fact that very often human trafficking ads also include a rather large amount of restrictions (hours of service; only these services provided, etc.) and negations mentioned within them (the text "no" appears often), which might be the reason for excessive 'negativity' from the Netflix model. Analyzing the HT Provider Review binary model, we observe that the majority of all ads in general appear to be positive, which is a trend quite often observed within human trafficking web ads.

The Stanford categorical model shows a general tendency towards *like*, *sad* and *neutral*, however *love* appears to be a good indicator of relevancy with 0 *love NOT RELEVANT* ads and 138 *love RELEVANT* ads. *angry* appears to be a good indicator or irrelevancy with 296 *angry NOT RELEVANT* ads and 112 *angry RELEVANT* ads. The results with the *love* label is, in fact, one of the main reasons for the creation of HT Ensemble I model. Such a trend could be explained by the general 'love' theme of the HT ads.

The HT categorical model strengthens our hypothesis about the importance of *love* label in analyzing HT data. Based on its classification, the *love* label provides a good indicator of human trafficking relevancy (with 0 *love NOT RELEVANT* ads and 93 *love RELEVANT* ads), and angry is a good indicator of human trafficking irrelevancy (with 449 *angry NOT RELEVANT* ads and 377 *angry RELEVANT* ads). The HT Ground Truth Binary model has a general tendency to classify more ads as NOT RELEVANT compared with the HT Ensemble I which produces many more RELEVANT classifications.

After performing exploratory analysis on resulting identified correlations between our training models and test HT Ground Truth data, we determined that the HT Categorical model and Stanford categorical model provided correlations with HT *RELEVANT* ads when the sentiment analysis from both models labeled the ads as *love*. In addition, as previously noted, we also saw in our exploratory analysis a high degree of *RELEVANT* HT ads exhibited repeating geolocations (e.g., Las Vegas). Finally, we noted that the ad text in HT *RELEVANT* ads tended to use extreme negative words and language ("negation"). Given this, we followed with the approach as outlined in Figure 2.

The steps from Figure 2 are as follows (1) take ad text, and run Text-to-Tag-ratio (TTR) algorithm [17] to generate ad-relevant text and remote HTML; (2) to run HT Categorical classier model and the Stanford Categorical model on the ad-relevant text to obtain a sentiment label for each; (3) to convert sentiment label from categorical values to a yes/no value corresponding to yes if the sentiment was *love*, no otherwise; (4) running the cleansed ad text through GeoTopicParser [10] to obtain a set of location labels; (5) scanning cleansed ad text for the presence of negation text and recording a value of *yes* if present; or *no* otherwise, using a simple vocabulary and regular-expression searching for "no"; and (6) providing these features as input to the pre-trained HT Ensemble Binary I model to obtain a *RELEVANT* or *NOT RELEVANT* prediction. In the case of HT Ensemble Binary II model, we only select the top 5 geolocations from step (4),and provide those features to obtain the human trafficking *RELEVANT* or *NOT RELEVANT* prediction.

5 EVALUATION

We evaluated the prediction accuracy and precision (iterations to converge) for all of our produced models. The full accuracy and precision evaluation can be found at [12]. In this section we will focus on using receiver operating characteristic (ROC) curves [3] for model accuracy evaluation and number of training iterations/time to converge as a method of evaluating precision. ROC curves are

one of the most used ways to assess the performance of machine learning models. Evaluation is performed for the HT Ground Truth Binary, and HT Ensemble Binary I and II models. We did not compute ROC or iterations to converge for any of the prior models as they were used to inform our eventual ensemble models and were not intended for direct causal human trafficking prediction.

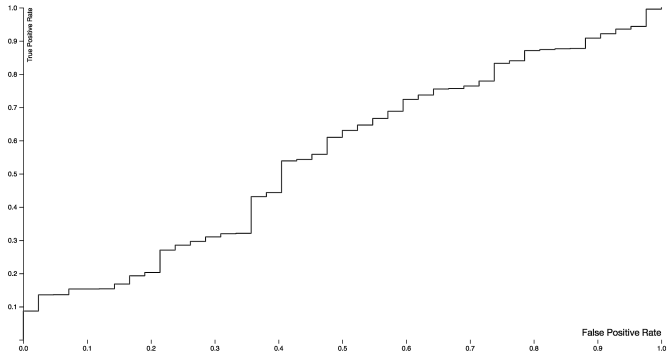

Figure 3: HT Ground Truth Binary ROC curve

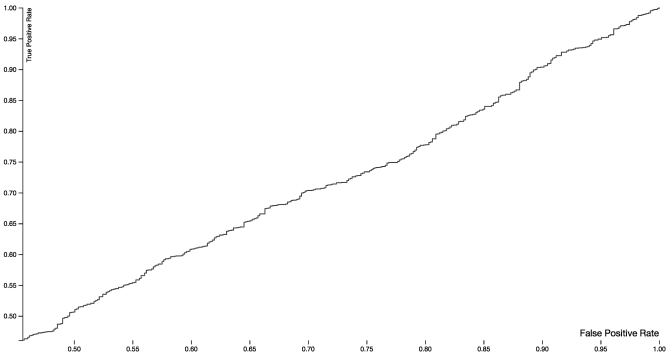

Figure 4: HT Ensemble I ROC curve

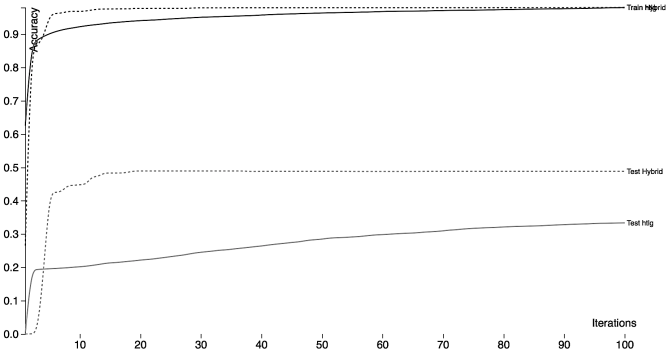

Figure 5: Model Performance Accuracy Curves for HT Ground Truth and HT Ensemble I

For the purposes of prediction accuracy, we define the true positive rate (TPR) as the number of times that the model, given a label withheld ad from the HT Ground Truth testing set of 38,563 ads, is

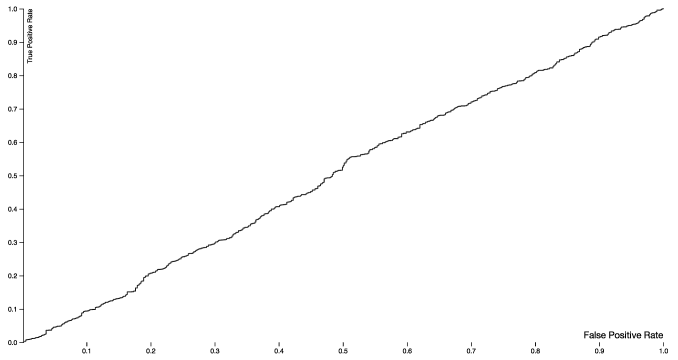

Figure 6: HT Ensemble II ROC curve

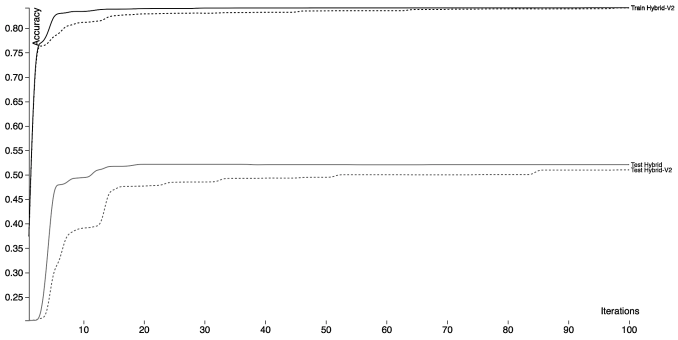

Figure 7: Model Performance Accuracy Curves for HT Ensemble I and II

able to predict that ad correctly as Human Trafficking *RELEVANT* or *NOT RELEVANT* and the false positive rate (FPR) as the inverse of the TPR. Considering only the HT Ground Truth Binary model with this metric follows naturally, but for the ensemble models there are pre-processing steps of course. For the HT Ensemble Binary I and II models, the steps shown in Figure 2 were followed to proxy from $\{ht_{cg}, stan_{cg}, geolocation, negation_{yes|no}\}$ space to {*RELEVANT*,*NOT RELEVANT*} space.

Creating the ROC curves involved breaking up the *HT ground truth* dataset into 80:20, where 80% of the ads were used for training purposes and 20% of them for testing. The labels from the 20% were removed and saved, the model was trained on the 80% and later on run on the newly created test set. The true positive rate (TPR) and false positive rate (FPR) were thereafter calculated for all 4,449 test ads (the 20%) and used to plot the curves.

For the HT Ground Truth Binary model the ROC curve displayed in Figure 3 fluctuates in its true positive rate (TPR) and false positive rate (FPR), indicating unreliability of its prediction. Despite its overall increasing ratio, its false positive predictions are quite high. In analyzing the accuracy of the HT Ensemble I model as shown in Figure 4, we can observe a reduced number of fluctuations and a slightly larger increase in TPR after FPR has reached approximately 0.90. Comparing the HT Ensemble I to the HT Ground Truth Binary ROC curve, HT Ensemble I clearly outperforms HT Ground Truth Binary, indicating that the ensemble features serve as a more accurate proxy than simply the Text-Tag-Ratio (TTR) processed

ad-text. In terms of convergence, as shown in Figure we see that HT Ground Truth Binary model clearly requires at least 100 iterations as its accuracy never stops to increase. In the case of HT Ensemble I, compared with HT Ensemble II shown in Figures 5, 6 and 7, however, training could easily stop after iteration 20 as the curve tapers off after that, indicating its quicker convergence.

6 FUTURE WORK AND CONCLUSIONS

Based on our early work in this area sentiment analysis is a viable classification mechanism and proxy to identify human trafficking in web data. Using open source sentiment models and models trained on human trafficking provider review data we were able to use exploratory analysis to find trends suggesting an approach for what textual, sentiment, geographic and natural-language cues are appropriate features to indicate if an ad is trafficking or not and to build accurate and perfomant ensemble models to automatically identify it. We can conclude that the HT Ensemble I Model, with a training set accuracy of 0.84 at iteration 100 and test set accuracy of 0.52 at iteration 100, clearly outperforms its competitors.

Our work is early and the accuracy of our models is rather moderate and can certainly be improved. Some possible ideas for improvement could include completely removing any bias from the data and looking for more elements within ads that could potentially act as valuable cues of human trafficking apart from geolocation and negation. Several teams are continuing work in reducing bias in the MEMEX datasets, including balancing out geographic features; balancing out ads based on services provided; based on related identified cues from the images related to the ads (e.g., tattoos) and looking at persona analysis to identify properties about the HT ad writer.

ACKNOWLEDGEMENTS

This effort was supported in part by JPL, managed by the California Institute of Technology on behalf of NASA, and additionally in part by the DARPA Memex/XDATA/D3M programs and NSF award numbers ICER-1639753, PLR-1348450 and PLR-144562 funded a portion of the work.

REFERENCES

[1] Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual Web texts. *Information retrieval* 12, 5 (2009), 526–558.

[2] Mondher Bouazizi and Tomoaki Ohtsuki. 2016. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter. In *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 1–6.

[3] Mithat Gönen et al. 2006. Receiver operating characteristic (ROC) curves. *SAS Users Group International (SUGI)* 31 (2006), 210–231.

[4] Christopher Johnson, Parul Shukla, and Shilpa Shukla. 2012. On classifying the political sentiment of tweets. *cs. utexas. edu* (2012).

[5] Kamala Kempadoo, Jyoti Sanghera, and Bandana Pattanaik. 2015. *Trafficking and prostitution reconsidered: New perspectives on migration, sex work, and human rights*. Routledge.

[6] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A Large Self-Annotated Corpus for Sarcasm. *arXiv preprint arXiv:1704.05579* (2017).

[7] Nicholas D Kristof. 2012. Where pimps peddle their goods. *The New York Times* 11 (2012).

[8] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 415–463.

[9] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 142–150.

[10] C. Mattmann. [n. d.]. GeoTopicParser. http://wiki.apache.org/tika/GeoTopicParser. ([n. d.]).

[11] Chris A Mattmann and Madhav Sharan. 2016. An Automatic Approach for Discovering and Geocoding Locations in Domain-Specific Web Data. In *Proceedings of the 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI?16)*. 87–93.

[12] A. Mensikova and C. Mattmann. [n. d.]. USC Data Science - Human Trafficking Lead Generation Analysis. http://irds.usc.edu/SentimentAnalysisParser/htlg.html. ([n. d.]).

[13] Apache OpenNLP. 2011. Apache software foundation. *URL http://opennlp. apache. org* (2011).

[14] Georgios Paltoglou. 2014. Sentiment analysis in social media. In *Online Collective Action*. Springer, 3–17.

[15] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[16] Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. 2017. Elixa: A modular and flexible absa platform. *arXiv preprint arXiv:1702.01944* (2017).

[17] Tim Weninger and William H Hsu. 2008. Text extraction from the web via text-to-tag ratio. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*. IEEE, 23–28.

[18] Florian Wogenstein, Johannes Drescher, Dirk Reinel, Sven Rill, and Jörg Scheidt. 2013. Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 5.