Labeling online illegal wildlife advertisements with Active learning-SVM

Sai Manohar Reddy Peddireddy
School of Informatics and Computing
Indiana University Purdue University Indianapolis
Indianapolis, IN, USA
peddires@iu.edu

***Abstract*—This paper explores an application of active learning with support vector machines. Semi-supervised machine learning models are efficient and reduces the effort to label the data. As these models continue to learn by adjusting to the newer data, they are less prone to biases and adopt the natural occurrence. Few supervised learning models have been explored in the prior work to track illegal wildlife trade on the ecommerce websites. The same problem will be addressed, but in a semi-supervised fashion.**

Keywords—active learning, SVM, wildlife, ecommerce, semi-supervised, labeling.

I. Introduction

Labelling data for the purpose of training supervised machine learning models can take lots of time and resources but can be efficient with active learning [5]. Active learning allows the model to adjust its learning parameters with new data, this helps to label the data instead of manually labelling. SVM (support vector machines) is a well-known and proven machine learning model [6]. It has various applications and can be applied in text classification [5]. With the help of kernel trick, SVM classifier can used to classify nonlinearly separated data too [7].

Tracking illegal wildlife trade on internet has been an issue, as there are millions of advertisements posted and deleted every day on hundreds of different websites. Its very difficult to manually check each and every advertisement for illegal activities. With the help of machine learning techniques, we can narrow down this search to identify potential illegal advertisements. E-commerce advertisements at various websites come with different features, usually an ad title, description, image, location, price, and various other features which vary. Images could be used as a preliminary feature but, it is difficult to distinguish between fake products and original products. For example, a fake leather jacket looks similar to real one in an image similarly fake animal linen fur can look same as real fur in an image. For this reason, text is used as main feature as other features from title, description, price, location can tell more about a product.

This problem is further divided into two problems, first is to classify advertisements into wildlife and non-wildlife item and then in the second stage illegal items are determined from the wildlife items which are classified. This paper tries to model for the first problem, that is to build a binary classifier to classify wildlife products from non-wildlife products in the advertisements.

II. Background

A. Wildlife Trade

Internet has become a common place to buy and sell goods globally which is very advantageous if rightly used, but it is now a common place for organizations and

sellers trying to trade illegal or endangered wildlife species online. If you take the case of ivory tusks of elephants or tiger skin, these kinds of products are banned by Interpol and many countries this drove up the cost of these products in black-markets across the globe and the wildlife trade market has evolved to a multi-billion-dollar industry [1] [3]. Given these species are endangered and many legislative authorities banning the trade of these wildlife species, there is a high demand for these rare valuable wildlife products. Bennett describes enforcement on illegal wildlife trade is old-fashioned as internet is aiding organized crimes to buy and sell body parts of various endangered species [2]. This clearly raises a need to tackle this problem which has been evolving over the internet.

B. E-Commerce Marketplace

As of 2017 Pipe Candy's e-commerce database reports 750,000 online e-commerce businesses and estimates 2-3 million of them around the world with 1.89 trillion-dollar market size [4]. This statistic alone tells the sheer size of the online market place and as Bennett mentioned it's a poorly regulated and hard to enforce space [2]. As much of the trade has expanded with many international players it has become easier for the sellers to keep their location anonymous or secret while they operate their business. In the context of illegal wildlife trade, sellers and buyers tend to innovate ways to organize their trade anonymously and keep their transactions buried under other transactions taking advantage of the size of the global market place and number of transactions done globally.

C. SVM for Semi-supervised Learning approach.

Support vector machines is a widely used supervised learning technique for classification. The basic principle behind support vector machines is to build a hyperplane between linearly separable data, non-linear boundaries can also be drawn with kernel trick [6]. When there are lots of data points but limited training examples (labeled data), the hyperplane could be adjusted to the newer data based on the data points that are closest to the current hyperplane in the unlabeled data instead of randomly choosing the unlabeled data. There are many researches having different approaches to this method like pool-based approach [5], expected model reduction [8], expected error reduction, query by committee [9] etc. These approaches help select the next best unlabeled data points to be labeled and added to the training dataset to retrain the model for next iteration.

III. PRIOR WORK

In our prior work we have trained models to answer similar research aims, we tested supervised learning methods like Bag of words MLP (Multi-Perceptron Network) and 1-D CNN (1-dimensional Convolutional Neural networks). These methods require a labeled training dataset for the model to learn. We labeled some data manually and trained our models, the table below shows the results from prior work.

TABLE I. PRIOR MODELS

S.no.	**Performance**	
	Model Name	***Accuracy***
1	Bow-MLP	83.4%
2	1-D CNN	89.2%

Fig. 1. Example of a figure caption. (*figure caption*)

Having 2.5 million advertisements in the dataset is great but labeling them to check for performance of the model drained our resources. In this paper we want to extend our research with the SVM semi-supervised technique. Given the void in tracking illegal wildlife advertisements online and our knowledge on how to address this problem, we hope this will be a better method to efficiently keep track of advertisements online and alert officials about any potential illegal wildlife trade.

IV. EXPERIMENTS AND METHODOLOGY

A. Data Collection Methodology

Having right data is important, we identified 37 websites which are places for potential wildlife advertisements and collected 2.5 million advertisements by web crawling and web scraping. Each advertisement has features 1. Title, 2. Description, 3. Price, 4. Seller name, 5. Seller location, 6. Image URL, 7. URL to the

ad. This data is formatted to JSON for easy transferability. Then, the data is cleaned and preprocessed to identify the key features to train on. As we are exploring learning models for text classification, we utilized 'Title' as the feature. About 20,000 advertisements are sampled with titles as documents.

B. *Data pre-processing and labelling*

In order to prepare the text data for training SVM model, 653 samples with good distribution of wildlife and non-wildlife samples are labeled for initial training of the model. Further 25% of these samples were used as test samples and 75% of the samples were used in the initial training dataset. This labeled data set is combined with the unlabeled samples (titles / documents) and vectorized with the help of TFIDF (Term frequency – Inverse document frequency tf-idf) [10]. For this first a dictionary of words is built by counting all the words from the 20,000 sample and term frequencies with in the document and document frequencies of the terms are calculated.

Term Frequency:

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \quad \text{- [12]}$$

Inverse document Frequency:

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad \text{- [12]}$$

TFIDF

tfidf (t, d, D) = tf (t, d) • idf (t, D) [12]

where, t = term(word)

d = A document

D = All the documents.

At this step all the samples including the initially labeled data are converted into vectorized form of frequency scores.

C. *SVM Model*

Support vector machines work on the principle to find the best possible hyperplane separating the data. This turns out to be a quadratic programming optimization problem. A dual form could be solved by solving langrangian dual of the primal given by

$$\text{maximize } f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i (x_i \cdot x_j) y_j c_j,$$

$$\text{subject to } \sum_{i=1}^{n} c_i y_i = 0, \text{ and } 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all } i. \quad \text{[13]}$$

Here c_i and c_j can be maximized with the help of quadratic programming methods.

D. *Active Learning Query*

Now, the next best possible set of unlabeled data points to retrain the model by adjusting the hyperplane are chosen by looking for the nearest n data samples to the current hyperplane. These n samples are added to the initial training data and retrained in the same way for following iterations.

E. *Experiments*

At each iteration the n (batch size) unlabeled data points which are nearest to the hyperplane are selected to be labeled and the model is retrained in the iteration. Various experiments were designed and conducted to reach best possible parameters including Kernels, Batch size (n for number of unlabeled data points to be labeled), and regularization parameter C.

Kernels can be chosen from 'linear', 'gaussian', and 'polynomial'. Batch sizes of [10, 50, 100, 200, 500] are experimented. Regularization parameter C – [10, 50, 100] are tested.

V. Results

After running the experiments with the above mentioned parameters 'linear' kernel consistently performed well with higher test accuracies averaging about 81.7%.

TABLE II. KERNEL PERFORMANCE

S.no	Parameters			Accuracy
	kernel	***C***	***n***	***Avg %***
1	linear	10	100	81.7%
2	rbf	10	100	79.6%
3	poly	10	100	79.6%

Fig. 2. Table with the three kernels average accuracies with regularization parameter (C) and batch size (n).

With linear kernel further experiments are done with different batch sizes and regularization parameters. Following are the results.

TABLE III. RESULTS

n	Parameters		
	C=10	***C=50***	***C=100***
n=10	82.7%	**83.12**%	79.4%
n=50	81.9%	81.2%	80.8%
n=100	81.01%	80.7%	79.1%
n=200	79.4%	80.6%	77.8%
n=500	75.5%	75.1%	72.4%

Fig. 3. Table of accuracies with linear kernel and varying parameters with c=10, 50, 100. n=[10, 50, 100, 200, 500].

Lower the batch size higher is the time taken for the entire cycle to finish as hyperplane adjusts slowly to lower number of unlabeled data.

VI. Conclusion and Future work

With these experiments we are able to

- We are able to detect extreme data points accurately.
- Achieve lesser false positives.
- With just information like title of an advertisement, we are able to identify wildlife products 83.12% accurately.

But we can improve the model as we have higher false negative rates this means some wildlife products are not easily separable from the non-wildlife advertisements. If we can add other features like description and price the model might pick up the differences between wildlife and non-wildlife, this can be the future work.

Acknowledgment

THIS WORK ALONG WITH PREVIOUS WORKS ARE SUPPORTED BY DR. SUNANDAN CHAKRABORTY AND HIS GRANTS. I THANK HIM FOR HIS HELP AND GUIDANCE.

References

[1] Yu, X., & Jia, W. (2015). Moving targets: tracking online sales of illegal wildlife products in China. TRAFFIC, Cambridge, United Kingdom.

[2] Bennett, E. (2011). Another inconvenient truth: The failure of enforcement systems to save charismatic species. Oryx, 45(4), 476-479. doi:10.1017/S003060531000178X

[3] Wasser, S. K., Clark, B., & Laurie, C. (2009). The ivory trail. Scientific American, 301(1), 68-77.

[4] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

[5] Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, *2*(Nov), 45-66

[6] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 3, Article 27 (May 2011), 27 pages. DOI: https://doi.org/10.1145/1961189.1961199

[7] Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition

learning. *Automation and remote control*, *25*, 821-837.

[8] Bovolo, F., Bruzzone, L., & Marconcini, M. (2008). A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(7), 2070-2082.

[9] Gilad-Bachrach, R., Navot, A., & Tishby, N. (2006). Query by committee made real. In *Advances in neural information processing systems* (pp. 443-450).

[10] Rajaraman, A., & Ullman, J. (2011). Data Mining. In *Mining of Massive Datasets* (pp. 1-17). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139058452.002

[11] https://en.wikipedia.org/wiki/Tf-idf

[12] https://en.wikipedia.org/wiki/Support_vector_machine#Kernel_trick

[13] http://blog.pipecandy.com/e-commerce-companies-market-size/