

Leveraging Machine Learning to Improve Unwanted Resource Filtering

Sruti Bhagavatula Christopher Dunn
Chris Kanich Minaxi Gupta Brian Ziebart

UIC Department of
UNIVERSITY OF ILLINOIS
AT CHICAGO Computer Science
COLLEGE OF ENGINEERING



SCHOOL OF INFORMATICS
AND COMPUTING

INDIANA UNIVERSITY
Bloomington

Introduction

Forbes | **New Posts** | **Most Popular** America's Youngest Billionaires | **Lists** The Forbes 400 | **Video** Top-Earning Dead Celebs | Search companies, people and lists | User profile icon

CODE RULES
A Playbook for Marketing at the Crossroads

Read CODE HALOS and recode the customer experience. Get The Book >>>

KEEP CHALLENGING™

Global Business & Technology Services | Cognizant

Top Stories

U.S. | Europe | Asia

Follow Forbes | Facebook | Twitter | LinkedIn | Google+ | YouTube



**Weapons Of Documentation:
Taser Is Pushing Cop Cameras**

Sears Shares Spike As Company Seeks To Raise Cash Through A REIT

Maggie McGrath, Forbes Staff

The Biggest Public Firms In Every American State

Erin Carlyle, Forbes Staff

Humana, Hurt By Sovaldi, Sees ACA And Medicare Boon In 2015

Bruce Japsen, Contributor

Survey Tracks Down One Billion 'Terrible' Web Users



Advertisement 00:28

PAUSE | Full Screen | Volume | MENU

Featured Videos

Power Perks

Waiting for stats.g.doubleclick.net...

Introduction

Forbes

New Posts

Most Popular

America's Youngest Billionaires

Lists

The Forbes 400

Video

Top-Earning Dead Celebs

Search companies, people and lists



Top Stories


U.S. Europe Asia

Follow Forbes



Weapons Of Documentation: Taser Is Pushing Cop Cameras

In the latest 12 months, Taser International netted \$17 million on \$148 million in revenue, but the problem is that stun gun sales aren't as shocking as the company would like. Rick Smith is interested in Taser's new growth business: body cams and digital video storage.

 **Kashmir Hill**, Forbes Staff

Sears Shares Spike As Company Seeks To Raise Cash Through A REIT

 **Maggie McGrath**, Forbes Staff

The Biggest Public Firms In Every American State

 **Erin Carlyle**, Forbes Staff

Humana, Hurt By Sovaldi, Sees ACA And Medicare Boon In 2015

 **Bruce Japsen**, Contributor

Survey Tracks Down One Billion 'Invisible' Web Users

 **Parmy Olson**, Forbes Staff

iDevices Wants To Show Smart Home World How To Play Apple's Game

 **Aaron Tilley**, Forbes Staff

Typical Advertisement

```
▼ <iframe src="http://view.atdmt.com/HAC/iview/477009459/direct/01/1440972580?cli...bYFU60EAqPrJF
marginheight="0" marginwidth="0" topmargin="0" leftmargin="0" allowtransparency="true">
  ▼ #document
    ▼ <html>
      ▶ #shadow-root
        <head></head>
      ▼ <body style="margin:0" marginwidth="0" marginheight="0">
        ▼ <a target="_blank" href="http://clk.atdmt.com/goiframe/344395124/477009459/direct/01"
aomRwQT63p4cY3XUnATPiq4mn6QA7E4WUmXHvZbnW2w5mvU5VrgTcQ9VcjhSA3oTHYSTUFX5bepVaYtVTJbPaMZc
http://t.atdmt.com'">
          
          </a>
          ▶ <script type="text/javascript">...</script>
        </body>
      </html>
    </iframe>
```

Typical DOM structure of an advertisement element in a page.

Ad-Blocking

- URLs matched against filters
- DOM element names matched against element hiding filters
- Iframe content removed
- Resource requests blocked

```
_300_250_  
_300_60_  
300x160  
_300x250-  
300x250.  
_300x250_  
300x250a_  
_300x250b.  
300x250px.  
_300x250v2.  
_300x600.  
_300x600_
```

Blocked Advertisement

```
▼ <iframe id="google_ads_iframe_/16921351/9gag-list-sidebar1-300x250-atf_0__hidden_" name=
width="0" height="0" scrolling="no" marginwidth="0" marginheight="0" frameborder="0" src="
"border: 0px; vertical-align: bottom; visibility: hidden; display: none;">
  ▼ #document
    ▼ <html>
      ▶ <head>...</head>
      <body marginwidth="0" marginheight="0"></body>
    </html>
  </iframe>
  ...
```

After the iframe and images were matched and blocked.

AdBlockPlus Filters

- Typical EasyList general URL filters. (right)
- Multiple filter lists – tens of thousands of filters total.
- Updated every few days with new specific regexes.

```
?view=ad&
?wm=*&prm=rev&
?ZoneID=*&PageID=*&SiteID=
^fp=*&prvtof=
^mod=wms&do=view_*&zone=
_125ad.
_160_ad_
_160x550.
_300x250Banner_
_468x60ad.
_728x90ad_
_acorn_ad_

adwrap.
afd_ads.
affiliate/banners,
affiliate_ad.
afs_ads.
alt/ads/
argus_ad_
assets/ads/
background_ad.
background_ad/

/totemcash1.
/tower_ad_
/towerbannerad/*
/tr2/ads/*
/track.php?click=*&domain=*&uid=$xmlhttprequest
/track.php?uid=*. *&d=
/track_ad_
/trackads/*
/tracked_ad.
/trade_punder.
/tradead_
```

Motivation

- Advertisements are distracting and a potential security and privacy risk.
- Ad blockers use thousands of hand-crafted filters - manually updated through constant advertisement tracking and user feedback.
- Ad blocking assisted by machine learning can improve ad blocking quality and decrease filter crafting effort.

Approach

- Crawl URLs of today and compare with present and historical filters.
- Bootstrap a supervised classifier based on historical regex matches to identify new ads.
- Train multiple classification algorithms to test suitability to the problem.

Related Work

- Classification of advertisement images using C4.9 [Kushmerick '99].
- Classification of advertisements using Weighted Majority Algorithm [Nock et al. '05].
- Rule-based classification of advertisements. [Krammer '08].

Datasets

- Depth 2 web crawl from Alexa top 500
 - 60,000 URLs total
- URLs matched against EasyList filters – binary class labels.
- 2 sets of class labels:
 - “Old” labels – matched against September 23rd, 2013 filter list.
 - “New” labels – matched against February 23rd, 2014 filter list.

Feature Sets

- A. Ad-related keywords (2 features)
- B. Lexical features (2 features)
- C. Related to the original page (2 features)
- D. Size and dimensions in URL (2 features)
- E. In an iframe container (1 feature)
- F. Proportion of external requested resources (3 features)

Select Features

- **Base Domain in URL:**

`http://1.betrad.com/ct/0/pixel.gif?
ttid=2&d=www.livejournal.com&`

- **Ad Size in URL:**

`http://cdn.atdmt.com/b/HACHACYMCAYKC/
Adult_300x250.gif`

Evaluation Methodology

- Evaluate **coverage** using old filters and **improvement** using current filters.
- Bootstrap the classifier using older classifications of EasyList for training.
- Evaluate against classifications based on newer EasyList to evaluate its ability to recognize unrecognized ads.

Evaluation Methodology

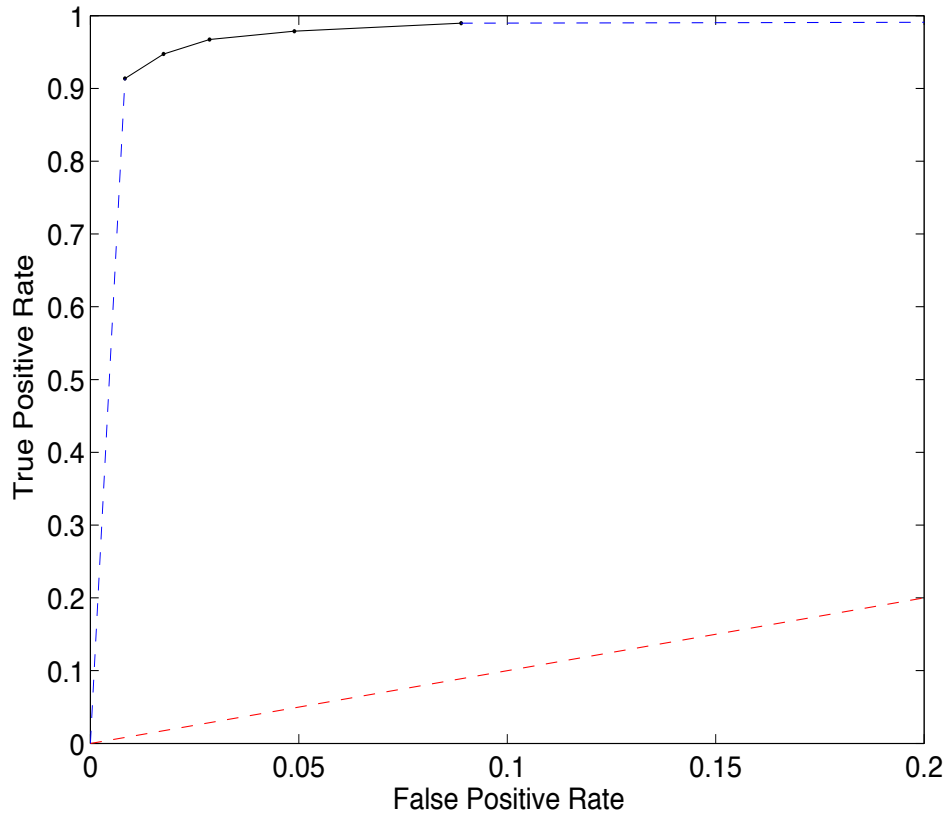
- Specific metrics:
 - **Baseline Accuracy** =
$$\frac{\text{No. of positively classified URLs matched by both lists}}{\text{No. of URLs matched by both lists.}}$$
 - **New-ad Accuracy** =
$$\frac{\text{No. of positively classified URLs matched by the new but not old}}{\text{No. of URLs matched by the new but not old}}$$

Comparison of Classifiers

Classification Method	Avg. Accuracy	Precision	FP-rate
Naïve Bayes	89.50%	89.09%	14.3%
SVM (linear)	92.10%	92.36%	7.4%
SVM (poly)	90.51%	90.56%	7.34%
SVM (rbf)	92.18%	92.43%	7.7%
L2-reg. Logistic Regression	92.44%	92.43%	7.5%
K-Nearest Neighbors	97.55%	98.60%	1.3%

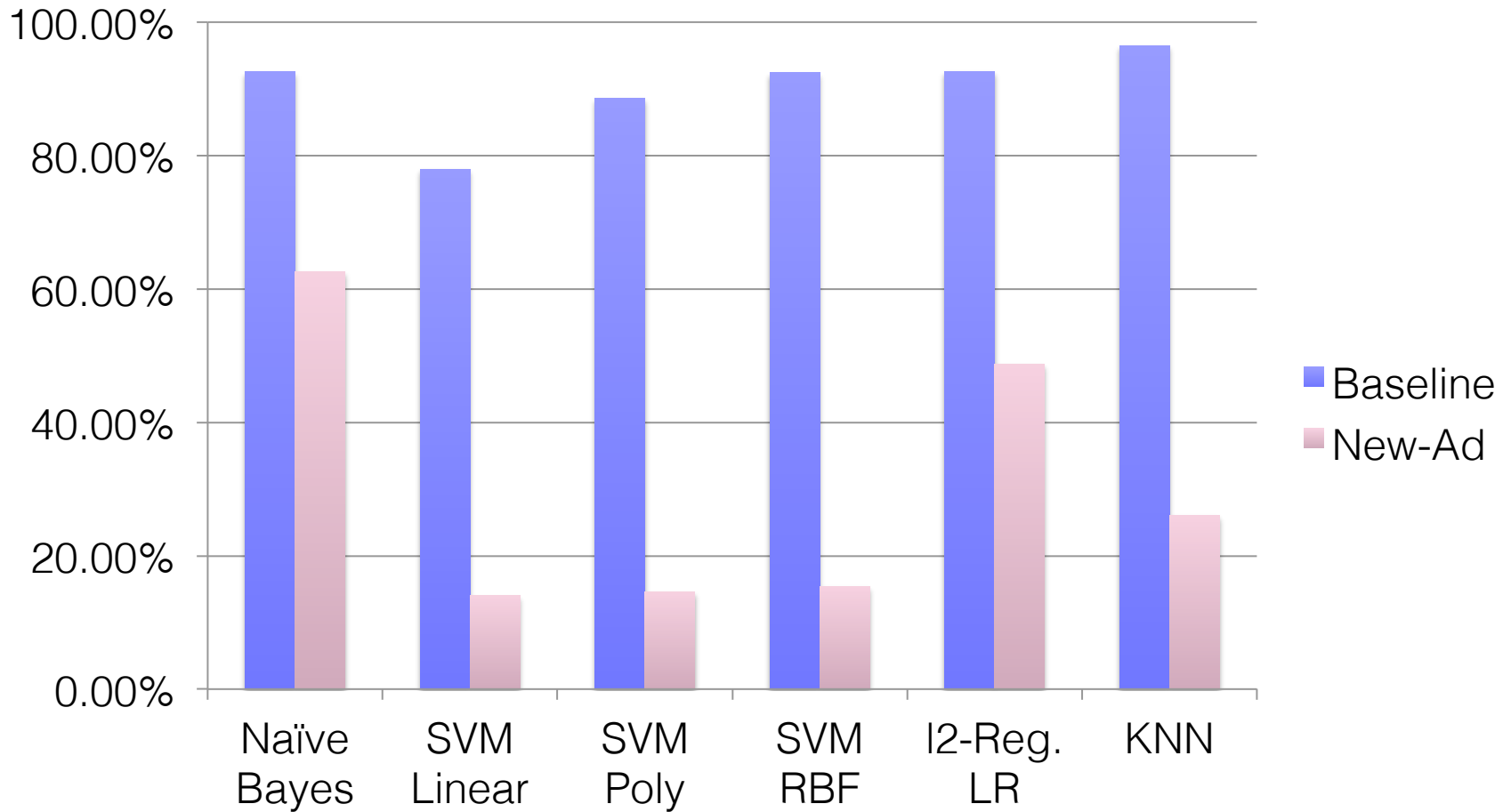
k-Nearest Neighbors had the best overall accuracy and other measures.

ROC Curve



Receiver Operating Characteristic (ROC) curve of the kNN classifier.

Baseline and New-Ad Accuracy



Performance of features with kNN

Feature Set (f)	Avg. Accuracy	Baseline Accuracy	New-ad Accuracy
A	90.21%	81.82%	48.78%
B	97.42%	95.20%	48.78%
C	96.82%	95.16%	34.96%
D	95.94%	93.38%	27.64%
E	96.22%	94.21%	21.95%
F	76.88%	57.50%	9.76%

Table of average accuracy, baseline accuracy and new-ad accuracy without each feature set (f)

Ad-related keywords and **proportion of external resources** feature sets are the most crucial ones.

Minimizing False Positives

- Compared False Positives against very recent filter list from June 7th, 2014.
- Approximately 7% of them were matched by the more recent filters.
- 70% of positively misclassified ads were actually advertisements unrecognized by EasyList.

Future Work

- Incrementally learn accurate and new ads based on user feedback.
- Crowdsourcing feedback on new advertisements and falsely classified resources.

Conclusion

- Machine learning based classifier which was able to automatically learn currently known and unknown ads and up to 50% of new ads.
- Further enable user choice on what ads, tracking beacons, and other undesirable web assets are loaded on their machines, improving the end-user experience and overall web security.

Thank you!

- Questions?

UIC Department of
UNIVERSITY OF ILLINOIS
AT CHICAGO Computer Science
COLLEGE OF ENGINEERING



**SCHOOL OF INFORMATICS
AND COMPUTING**

INDIANA UNIVERSITY
Bloomington