AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

THE 13th INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT)

Ad Service Detection - A Comparative Study Using Machine Learning Techniques

Yadhu Krishna M, Sanjana S
Department of Computer Science and Engineering,
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amritapuri, India
{yadhukrishna, sanjanas}@am.students.amrita.edu

Thushara M.G.
Department of Computer Science and Applications,
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amritapuri, India
thusharamg@am.amrita.edu

Paper ID : 300 | Serial No : 243

Presented By : Yadhu Krishna M

4th October 2022

Presentation Outline

- Introduction
- Research Contributions
- Literature Review
- Methodology
- Results
- Conclusion & Future Work
- References

Introduction

- Number of advertisements & trackers on the internet has gone up.
- They collect personally identifiable information about user, user behavior and sensitive information
- Used to provide personalised advertisements and track the user
- Major threat to user privacy and security
- Ad blockers were developed to prevent this abuse
- Traditional Ad blockers are based on static blacklist

Research Contributions

- Novel approach to develop Ad blockers
- Overcome issues of traditional Ad blockers:
  - Performance issues
  - Inefficiency of static blacklists
  - Frequent updates to blacklists
- Can be combined with traditional Ad blockers
- Comparative study of application of various Machine Learning algorithms to this problem
  - Logistic Regression
  - Random Forest (RF)
  - Support Vector Machines (SVM) classifier
  - Decision Tree classifier
  - K-Nearest Neighbors (KNN) classifier

Literature Review

- Iqbal et al. proposed a graph-based machine learning approach.
- Bhagavatula et al. trained a supervised machine learning model based on keywords extracted from HTTP query string.
- Zain ul abi Din et al. proposed an ad blocker system based on deep learning.
- Mughees et al. performed a study using patterns during DOM changes.
- Gugelmann et al. introduced a machine learning based method for classifying WTA requests - WTAGraph
- Hieu Le et al. [7] proposed AutoFR, that relies on reinforcement learning (RL) to generate filter rules in order to block unwanted URLs.

Methodology

1. Collected dataset from:
    a. Ad blocker dataset - list of websites serving advertisements.
    b. Alexa Top 1M websites - considered to be non-ad serving websites.

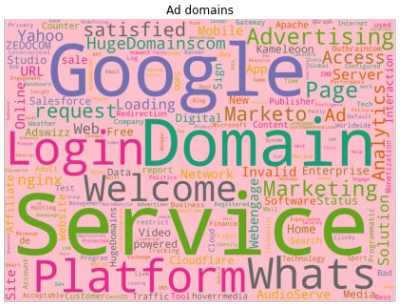
2. Perform web scraping on all the domains to collect data
    a. Static Features (Occurrences of special characters, length of domain, count of digits, etc.)
    b. Dynamic Features (Data extracted from webpage)

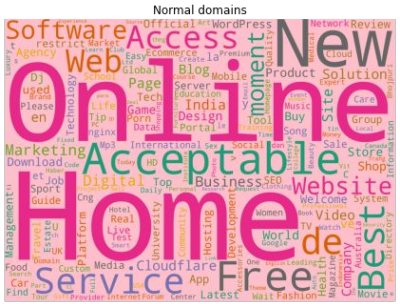
TABLE REPRESENTING FEATURES PRESENT IN THE DATASET

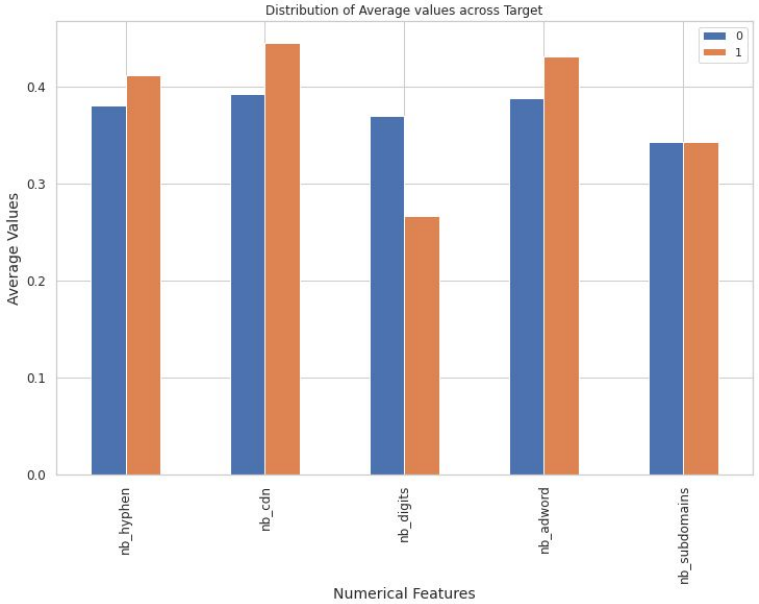
Variable	Datatype	Description	Type
class	Categorical	Describes class of record. Class 0 indicates ad service domains, and class 1 indicates normal domains.	Static
url	String	It contains the URL to which the HTTP / HTTPS request was sent.	Dynamic
status_code	Integer	Contains the HTTP status code returned to the HTTP request. It ranges from 100 - 599.	Dynamic
meta_info	String	Contains metadata retrieved by querying domains.	Dynamic
length	Integer	Length of domain name	Static
nb_hyphen	Integer	Count of hyphens in domain name.	Static
nb_cdn	Integer	Count of the keyword “cdn” in domain name.	Static
nb_digits	Integer	Count of digits in domain name.	Static
nb_adword	Integer	Count of the keywords “ad”, “ads” in domain name.	Static
nb_subdomains	Integer	Count of subdomains	Static

Methodology

3. Perform Exploratory Data Analysis
    - Visualization
    - Compare correlation between different variables
    - Construct word cloud to identify important keywords







Methodology

4. Apply Natural Language Processing on dataset
    a. Extract text data
    b. Apply stemming, lemmatization, remove stop words, lower case conversion, remove special characters

5. Apply and Compare various Machine Learning Algorithms
    a. Logistic Regression
    b. Random Forest
    c. Support Vector Machines (With different SVM Kernels)
    d. Decision Tree classifier
    e. K-Nearest Neighbors (KNN) classifier (With different values of K)

Results

SCORE OBTAINED BY KNN MODEL WITH VARIOUS VALUES OF N

N	Accuracy	F1	Precision	Recall
3	0.63	0.21	0.8	0.13
5	0.82	0.78	0.77	0.81
7	0.82	0.78	0.77	0.81
9	0.82	0.78	0.77	0.8
11	0.83	0.79	0.79	0.8
15	0.82	0.78	0.78	0.79
17	0.82	0.78	0.78	0.78
19	0.82	0.78	0.78	0.78
21	0.82	0.78	0.78	0.78
23	0.82	0.77	0.78	0.78

SCORE OBTAINED BY SVM MODEL WITH VARIOUS KERNELS

Kernel	Accuracy	F1 score	Precision	Recall
Polynomial	0.88	0.85	0.86	0.85
RBF	0.88	0.85	0.86	0.85
Linear	0.87	0.84	0.86	0.83
Sigmoid	0.84	0.8	0.836	0.78

Results

TABLE SUMMARIZING THE SCORES OBTAINED BY VARIOUS SUPERVISED MACHINE LEARNING MODELS

No.	Model	Remarks	Accuracy	F1 Score	Precision	Recall	Training Time	Testing Time
1	**Support Vector Machine Classifier**	Kernel = RBF	0.88	0.85	0.86	0.85	372.78	11.96
2	**Random Forest Classifier**	Estimators = 200	0.88	0.85	0.85	0.85	188.11	4.49
3	**Logistic Regression Classifier**	-	0.85	0.81	0.85	0.78	6.94	0.60
4	**Decision Tree Classifier**	-	0.87	0.84	0.83	0.86	32.36	0.65
5	**K-Nearest Neighbors Classifier**	Neighbors = 11	0.83	0.79	0.79	0.80	5.51	22.69

Best results were given by **Random Forest Classifier & Support Vector Machine (SVM) classifier.**

Conclusion & Future work

- Developed a novel approach for developing ad blockers.
- The approach overcomes three major problems associated with traditional ad blockers
- The approach can be combined with the traditional static blacklist model to get the best of both worlds.
- Identified that Random Forest Classifier & Support Vector Machine (SVM) classifier works best in the scenario.
- The model can be incorporated into a browser extension that can be used in real-world scenarios
- The paper focuses only on metadata that was extracted by using a limited number of features that were derived from a web page.
- Many other features and algorithms can be explored.

References

[1] U. Iqbal, Z. Shafiq, P. Snyder, S. Zhu, Z. Qian, and B. Livshits, “Adgraph: A machine learning approach to automatic and effective adblocking,” arXiv preprint arXiv:1805.09155, vol. 41, 2018.
[2] S. Bhagavatula, C. Dunn, C. Kanich, M. Gupta, and B. Ziebart, “Leveraging machine learning to improve unwanted resource filtering,” in Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, 2014, pp. 95–102.
[3] P. Tigas, S. T. King, B. Livshits et al., “Percival: Making in-browser perceptual ad blocking practical with deep learning,” arXiv preprint arXiv:1905.07444, 2019.
[4] M. H. Mughees, Z. Qian, and Z. Shafiq, “Detecting anti ad-blockers in the wild.” Proc. Priv. Enhancing Technol., vol. 2017, no. 3, p. 130, 2017.
[5] D. Gugelmann, M. Happe, B. Ager, and V. Lenders, “An automated approach for complementing ad blockers’ blacklists.” Proc. Priv. Enhancing Technol., vol. 2015, no. 2, pp. 282–298, 2015.
[6] Z. Yang, W. Pei, M. Chen, and C. Yue, “Wtagraph: Web tracking and advertising detection using graph neural networks,” in IEEE Symposium on Security and Privacy, 2022.
[7] H. Le, S. Elmalaki, A. Markopoulou, and Z. Shafiq, “Autofr: Automated filter rule generation for adblocking,” arXiv preprint arXiv:2202.12872, 2022
[8] Firebog, “The big blocklist collection,” Available at https://firebog.net/.
[9] Anudeep, “Curated hostfile to block trackers and advertisements,” Available at https://github.com/anudeepND/blacklist.
[10] Alexa, “Alex top 1m sites dataset,” Available at http://www.alexa.com/ topsites.
[11] M. Thushara, S. Anjali, and N. Meera, “An analysis on different document keyword extraction methods,” in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019, pp. 933–937.
[12] ——, “A graph based approach for keyword extraction from documents,” in 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). IEEE, 2019, pp. 1–4.
[13] V. Muppavarapu, A. Rajendran, and S. K. Vasudevan, “Phishing detection using rdf and random forests.” Int. Arab J. Inf. Technol., vol. 15, no. 5, pp. 817–824, 2018.

Thank You

Paper ID : 300 | Serial No: 243

Presented By: Yadhu Krishna M

4th October 2022

AMRITA VISHWA VIDYAPEETHAM