FILTERING ADS BASED ON USER BROWSING DATA USING DATA ANALYTICS AND MACHINE LEARNING

Aakash Sharma

Department of Computer Science and Technology SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Kriti

Department of Computer Science and Technology SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

U.M. Prakash

Department of Computer Science and Technology SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Abstract - Digital **Advertising is a data-driven strategy for reaching the target audience. It has gained popularity because of the revenue it generates for the advertising agencies and the companies. It serves as a precious resource as it creates data that could be processed to give insights on the usage and can be altered to receive more audience. On the contrary, for some users, due to the irrelevancy of the advertisements being shown, this experience can leave a distaste for the sites they are visiting. The current Contextual Advertisement model predicts and displays the advertisements based on the current page's surrounding content. Ergo, a dire need to incorporate behavioural targeting with this model to create a more enhanced user experience. Consequently, the advertisements will be a distraction to the user, which increases the probability of reaching the target pool for the companies. With this project, we aim to build a solution prototype depicting Ads' ideal workflow in the Digital environment where the user experience is maintained and the advertising companies are benefited.**

Keywords – Ads Classification, Clustering, Ads Targeting, Machine Learning, Oversampling using SMOTE, data, browser

I. INTRODUCTION

The past decade has seen a huge growth in online advertising. Advertisers are very interested in precisely targeted ads, they want to spend the smallest amount and get the maximum targeted users i.e. the users which are most likely to buy their product. This is resolved by targeted advertising.

Digital advertising is data-driven strategy for target audience. But the user experience can become unpleasable due to irrelevant ads. Search engines store users' browsing data on the cloud. This data is then used to display ads to the users.

A particular website sets a number of ads to be displayed to the user which might or might not be relevant to the portal. We aim to alter this approach, to enhance user experience by filtering ads based on the user's history and relevancy of the current site. This will not only create a good user experience, but will also provide the advertising companies with the target audience.

The aim of this report is to implement other methods to achieve the targeted users to the advertising companies without disturbing the user experience on a particular portal or website.

II. PRIOR WORK

1) Innovative Idea

The paper uses a Machine Learning algorithm to display ads to the user based on their user browsing history as well as website relevancy. It also tries to predict the free space in DOM where the ads can be placed such that there is no hindrance to the original website content and user experience is not disrupted.

2) Purpose

To enhance the current user experience by filtering ads based on browsing history and website relevancy.

3) Scope

The paper displays the ideal behaviour of browsers and search engines. The paper tries to make the user's browsing experience much better that what it is today. One can imagine the number of ads flowing on the internet, which needs to be properly served to the right

users. The proposed idea is cheap and beneficial to all the potential users.

4) Present System

Today, everyone uses browsers for almost all their requirements. And ads are displayed to them, based on their browsing history. This history is stored in the user's device itself which is then parsed by the browser to display the ad, as the user visits some particular website. Now, this ad can be targeted anywhere and on any website, irrespective of its relevance, which creates a bad user experience.

5) Proposed System

The paper proposes a system that is a combination of contextual and behavioural advertisement strategies. The model that is being used today by most popular search engines is contextual advertising which aims to generate ads based on page or keyword relevancy. There is a dire need to incorporate behavioural targeting in order to exploit the available user data. Mining this gold called data we aim to enhance user experience and help companies boost their revenue by meeting their target pool.

III. RELATED WORK

In earlier days of the internet, online advertising was mostly prohibited. Later, it started with online advertisements via emails and then expanded to other internet platforms. Advertising is usually done in the form of text, visual images, banners, animations etc. Theses advertisements frequently target users with particular traits to increase the ads effect.

The common notion of ads is to collect the user's data through the browsing device and use to target the ads to that user. And this essentially forces the user to see these ads on every website or platform they visit, irrespective of its relevancy.

Search engines rank websites on the basis of number of clicks on it, and these clicks add to the market value of the website or portal.

Following are the common targeted methods used to target these advertisements –

1. Demographic Targeting
2. Property Targeting
3. Behavioural Targeting

Advertisers and publishers uses a wide range of payment calculated methods, some of them are –

1. Cost per mille (CPM)
2. Cost per click (CPC)
3. Cost per engagement (CPE)
4. Cost per view (CPV)
5. Cost per install (CPI)

The following data represents the time spend by users on various internet platforms:

TABLE I: Time spend by users on different internet platforms

Social Media	33%
Online TV & Streaming	16%
Music Streaming	16%
Online Press	13%
Others	22%

The following data represents comparison between various online media used for ads:

TABLE II: Comparison between various online media used for ads

Social Networks	37%
Individual retailer websites	34%
Price comparison websites	32%
Multi brand websites	21%
Visual social networks	20%
Travel review websites	16%
Emails from brand/retailers	14%
Deal of the day websites	12%
Mobile Apps	11%
Blogs	11%
Digital press and magazines	6%

We can clearly see the impact of social media, websites and other internet platforms, and one can imagine the type of data that is accessed through these portals. In the name of storing just the browser history, personal and individual details, including name, emails, etc. are also stored. And that is shared through different partners which is a whole market in itself.

Businesses make an average of $2 in revenue for every $1 they spend on Ads. The average click through rate across all industries is 3.17% for the search network and 0.46% on the display network. Every 4 in 10 internet users say that they follow their favourite brands on social media. Over 37% of online shoppers use social media for their product advertisement.

IV. METHODOLOGY

The contextual advertisement uses the content of the current page to predict and display advertisements. The placement of the advertisement itself seems to bear no significance, according to the recent studies. The advertisements thus showed bear relevancy to some extent. But this advertisement method has the potential to improve massively.

In this paper, behavioural targeting is integrated to achieve preferable results. To simplify our procedure, we have divided it into four major parts.

A. Web Portal

The interactive web portal consists of an upload option and an insight option. The company can upload the image advertisement and select/define tags for the same. The company's representative will also receive regular insights for the advertisement. To implement the insight generation, the paper proposes the use of Data Analytics to depict monthly views, clicks and buys through the advertisement.

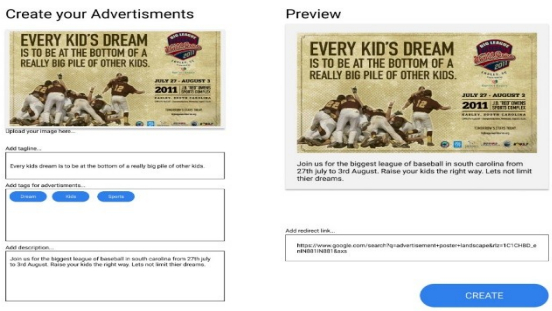

Fig. IV Components of the Web Portal

B. Topic Classification

Topic Classification is a part of Natural Language Processing that uses Deep Neural Networks (or any suitable algorithm that yields the highest accuracy) to predict or classify the text. It is a supervised learning method. Topic Classification in this paper aims to predict the category of the advertisement. Initially, Topic Modelling, an unsupervised learning method, was opted, with predefined tags, but the accuracy of the model was less than unexpected. To improvise, Topic Classification was chosen as a better alternative. To achieve topic classification, multiple NLP algorithms have been implemented and tested on the 'Advertisement Transcripts from Various Industries' Dataset. Our aim is to define the category of an advertisement. For Example, if the advertisement contains phrases like 'cloth,' 'sale,' 'discount' - the same will be classified as an 'E-commerce ad.' This tag helps the model pick this advertisement and display it on any E-commerce site to increase relevancy.

1. *Dataset*

The dataset used for this paper is 'Advertisement Transcripts from Various Industries' from Kaggle. It consists of approximately 2000 advertisement descriptions along with their category and corresponding companies.

The classes are completely mutually exclusive. Consequently, there is no overlap between batches.

2. *Pre-processing*

The dataset consisted of two essential and two inessential columns. The inessential columns ('Advertiser' and 'Product or spot) are dropped during the pre-processing step.

The dataset is visibly highly imbalanced (Fig 3.a). The category in the majority being 'Automotive' and the minority category being 'HealthCare'. To remove the inconsistencies in the data, and to achieve oversampling without over fitting, SMOTE is used at a later stage.

The textual data present in the 'Ad copy' column is cleaned by removing the stop words and bad characters/symbols (some applications of NLP).

TFIDF is used on the target class to provide the relevancy of the words with respect to the entire document thus, can be interpreted by the Machine.

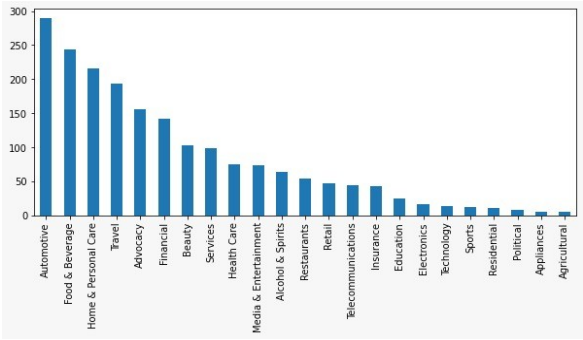

Fig. IV.2.a Imbalanced Categories in the dataset

3. *Models*

Preliminary to testing the model, chi-squared test using unigrams and bigrams are used to find the most relevant/correlated words per category.

The data needs Multiclass Classification. Multinomial Naive Bayes Classifier is used as it uses a multinomial distribution for each of the features.

The paper tests different parameters to yield the best accuracy for the predicted models while maintaining the minimum amount of loss suffered (Fig 2.b). The data is tested against these four models-

A. Logistic Regression
B. (Multinomial) Naive Bayes
C. Linear Support Vector Machine
D. Random Forest

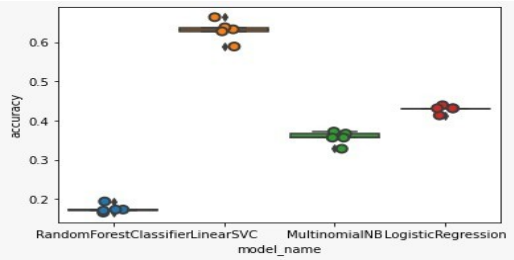

Fig. IV.2.b Accuracy of different models against the test data

The result of the accuracies is given in figure 2.c.

model_name	
LinearSVC	0.630491
LogisticRegression	0.429457
MultinomialNB	0.356072
RandomForestClassifier	0.175194

Name: accuracy, dtype: float64

Fig. IV.2.c Accuracy results of different models

The best result is yielded by Linear Support Vector Machine. A model is built and is cross-validated by a confusion matrix and indicated using a heat map. (Fig 2.d)

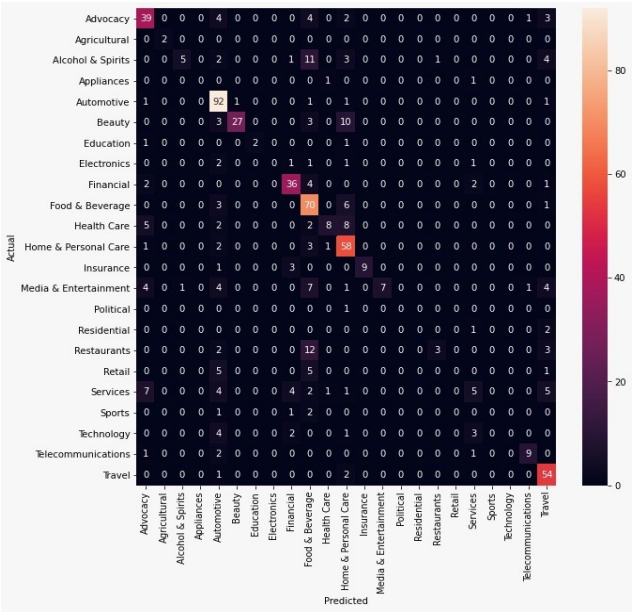

Fig. IV.2.d Linear SVM matrix heat map

The model evaluation score per category displayed in Fig 2.e. The metrics the model is evaluated against are precision, recall, and F1 score. These results are yielded without the application of SMOTE. To further improve the results, SMOTE is applied and the model is again developed and evaluated against the same metrics. Accuracy of up to 84 percent is achievable using the above method.

	precision	recall	f1-score	support
Advocacy	0.64	0.74	0.68	53
Agricultural	1.00	1.00	1.00	2
Alcohol & Spirits	0.83	0.19	0.30	27
Appliances	0.00	0.00	0.00	2
Automotive	0.69	0.95	0.80	97
Beauty	0.96	0.63	0.76	43
Education	1.00	0.50	0.67	4
Electronics	0.00	0.00	0.00	6
Financial	0.75	0.80	0.77	45
Food & Beverage	0.55	0.88	0.68	80
Health Care	0.73	0.32	0.44	25
Home & Personal Care	0.60	0.89	0.72	65
Insurance	1.00	0.69	0.82	13
Media & Entertainment	1.00	0.24	0.39	29
Political	0.00	0.00	0.00	1
Residential	0.00	0.00	0.00	3
Restaurants	0.75	0.15	0.25	20
Retail	0.00	0.00	0.00	11
Services	0.36	0.17	0.23	29
Sports	0.00	0.00	0.00	4
Technology	0.00	0.00	0.00	10
Telecommunications	0.82	0.69	0.75	13
Travel	0.68	0.95	0.79	57
accuracy			0.67	639

Fig. IV.2.e Model Evaluation Score per Category

C. Processing Users Data

Users' data is a pivotal point in behavioural targeting. With the availability of users' browsing data, the model will select the pertinent advertisement from our database. For Example, if the user has visited an E-commerce site, say, 'X' before and has browsed the site for clothes, now when the same user visits any other E-commerce site 'Y' with the same intent, the advertisement from the previously browsed E-commerce site 'X' will be displayed on site 'Y.'

Thus, the user experience is maintained, and the probability of companies reaching their target pool increases exponentially.

To tag the current site, a separate CNN model will be used to classify the site into different topics according to the site's content.

D. Displaying Processed Ads

Now that the data is processed, the machine learning algorithm will send the best suited Ad as a result, this Ad will be then picked up by the service plugin from the AWS S3 bucket and will be injected in the DOM of the website user is currently in. In this way the correct ad will be displayed to the end user.

V. RESULTS AND DISCUSSIONS

In the first iteration, we try to build the architecture design as shown in Fig.V.1, the figure represents an end to end flow. This gives an insight to the micro-service interactions following the model-view-controller design pattern. This will hence enable the service scalable to a large scale without the intervention of other services or dependency upgrades. Here, the three major components are browser-plugin, the cloud service API and the Web Portal for Advertising Company. The diagram represents the flow of interaction of one micro-service to another. The browser sends the user's data

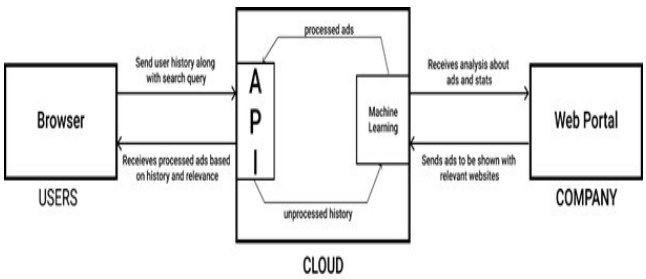

Fig. 1 Architecture Diagram

In the second iteration, we try to automate the UI flow through which ads can be stored directly into the database pool with its tags and description. We followed no-sql database for fast query fetching and traffic lookahead. This ensures quick iteration by the machine learning model to read the tags on advertisements and cluster them accordingly. The UI also provides analytics to the Advertising Company consuming libraries like D3 and others. The images are stored in the Cloud storage to ensure large bandwidth.

In the third iteration, we have connected our UI to the backend that performs CRUD operations to the users and posts data. The backend makes a connection between the MongoDB client and AWS S3 for cloud storage. It listens to various incoming requests from the browser and process them accordingly. It also manage user authentication, to avoid unnecessary creation of ads.

In our fourth iteration, we tried to figure out and classify our data according to accuracy of different models. We try to balance our dataset and then put it to train the model to obtain high accuracy. In this iteration, we also try to figure out different ways to classify the images from the Advertisement post.

In this last iteration, we try to extract user's data from web crawlers and send it to the flask API to process it and give back the recommended ad. This Ad we then parse to the user's website DOM, and hence the ad is displayed to the user. The API consumes Tensorflow.js and has direct access to cloud storage from where it sends the ads to the user. Enabling this not only provide excess of data but the bandwidth is also increased to a greater extent.

VI. CONCLUSION

In this paper, we have demonstrated the advertising structure that uses both contextual and behavioural advertising strategies. However, existing system suffer from large efforts being wasted to find the right audience for the right advertisement. Today, a system is established where it's all about reaching the target goals for posting the ads, which is not profitable for both the entities, but for the advertising platforms. These platforms generate a large revenue just by posting ads, distributing it, and collecting insights. But the advertisement company are still not able to reach their target audience, at the same time it decreases the experience of every other user visiting the platform.

We have established the necessity for a combination methodology to enhance user experience. We have used Machine Learning and NLP algorithms extensively throughout this paper to classify text or topic and to classify the sites as well. The accuracy of 84 per cent has been achieved following the methodology so proposed.

For the future prospects, we also suggest an image recognition system to define tags for an image advertisement. This paper will prove beneficial to those who are looking to incorporate artificial intelligence and big data in the advertisement field.

REFERENCES

[1] Kang S., Jeong, C. and Chung, K., 2020. *Tree-Based Real-Time Advertisement Recommendation System*. Online Broadcasting. IEEE Access, 8, pp.192693-192702.

[2] Sánchez-Núñez, P., Cobo, M.J., De Las Heras-Pedrosa, C., Peláez, J.I. and Herrera-Viedma, E., 2020. *Opinion Mining, Sentiment Analysis and Emotion Understanding in Advertising: A Bibliometric Analysis*. IEEE Access, 8, pp.134563-134576.

[3] Xiong, X., Xie, C., Zhao, R., Li, Y., Ju, S. and Jin, M., 2019. *A Click through Rate Prediction Algorithm Based on Users' Behaviours*. IEEE Access, 7, pp.174782-174792

[4] C. Yin, L. Shi and J. Wang, 2019. *Improved collaborative filtering recommendation algorithm based on differential privacy protection*. Advanced Multimedia and Ubiquitous, Singapore: Springer, vol. 518, pp. 253-258, 2019.

[5] Wang, X., Yang, L.T., Kuang, L., Liu, X., Zhang, Q. and Deen, M.J., 2019. *A tensor-based big-data-driven routing recommendation approach for heterogeneous networks*. IEEE Network, 33(1), pp.64-69.

[6] Kumar, P. and Thakur, R.S., 2018. *Recommendation system techniques and related issues: a survey*. International Journal of Information Technology, 10(4), pp.495-501.

[7] Zhang, Y., Yin, H., Huang, Z., Du, X., Yang, G. and Lian, D., 2018, February. Discrete deep learning for fast content-aware recommendation. In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 717-726).

[8] Mizan, C.M., Chakraborty, T. and Karmakar, S., 2017. *Text Recognition using Image Processing*. International Journal of Advanced Research in Computer Science, 8(5).

[9] Bujlow, T., Carela-Español, V., Sole-Pareta, J. and Barlet-Ros, P., 2017. *A survey on web tracking: Mechanisms, implications, and defences. Proceedings of the IEEE*, 105(8), pp.1476-1510.

[10] Son, J. and Kim, S.B., 2017. *Content-based filtering for recommendation systems using multiattribute networks*. Expert Systems with Applications, 89, pp.404-412.

[11] Zhang, Z., Xiao, Y., Zhu, W., Jiao, X., Zhu, K., Deng, H. and Shen, Y., 2017. *A context-aware recommendation system based on latent factor model*. IEEE 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp. 1-6).

[12] Sun, P., Wen, Y., Ta, D.N.B. and Xie, H., 2016. *Metaflow: a scalable metadata lookup service for distributed file systems in data centres*. IEEE Transactions on Big Data, 4(2), pp.203-216

[13] Chen, G., Cox, J.H., Uluagac, A.S. and Copeland, J.A., 2016. *In-depth survey of digital advertising technologies*. IEEE Communications Surveys & Tutorials, 18(3), pp.2124-2148

[14] Wang, R., Gou, Q., Choi, T.M. and Liang, L., 2016. *Advertising strategies for mobile platforms with "Apps"*. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(5), pp.767-778

[15] Rahman, W.U., Yun, D. and Chung, K., 2016. *A client side buffer management algorithm to improve QoE*. IEEE Transactions on Consumer Electronics, 62(4), pp.371-379.

[16] Tao Mei; Xian-Sheng Hua (2010). *Contextual Internet Multimedia Advertising*. International Journal of Advanced Research in Computer Science, 98(8), 1416–1433

[17] Anastasakos, T., Hillard, D., Kshetramade, S. and Raghavan, H., 2009, November. *A collaborative filtering approach to ad recommendation using the query-ad click graph.* 18th ACM conference on Information and knowledge management (pp. 1927-1930).

[18] Chan, J.C., Jiang, Z. and Tan, B.C., 2009. *Understanding online interruption-based advertising: Impacts of exposure timing, advertising intent, and brand image*. IEEE Transactions on Engineering Management, 57(3), pp.365-379

[19] Danaher, P.J. and Mullarkey, G.W., 2003. *Factors affecting online advertising recall: A study of students*. Journal of advertising research, 43(3), pp.252-267.

[20] Zhang, G. P., 2000. *Neural networks for classification: a survey*. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 30(4), 451–462. doi:10.1109/5326.897072

[21] Yang Wang, & Mori, G., 2009. *Human Action Recognition by Semilatent Topic Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(10), 1762–1774. doi:10.1109/tpami.2009.43

[22] C. Yin, L. Shi and J. Wang, 2019. *Improved collaborative filtering recommendation algorithm based on differential privacy protection*. Advanced Multimedia and Ubiquitous, Singapore: Springer, vol. 518, pp. 253-258, 2019.