ISSN 1822-8038 (online)
INTELEKTINĖ EKONOMIKA
INTELLECTUAL ECONOMICS
2022 , No 16(2), p. 61–75

USING MACHINE LEARNING MODELS TO INVESTIGATE CONSUMER ATTITUDES TOWARD ONLINE BEHAVIORAL ADVERTISING

Christos Ziakis[1]

Department of Economic Sciences
International Hellenic University, 62124, Serres, Greece, ziakis@gmail.com
ORCID: 0000-0001-7744-8215

Dimitrios Kydros

Department of Economic Sciences
International Hellenic University, 62124, Serres, Greece, dkydros@es.ihu.gr
ORCID: 0000-0003-0638-6701

DOI: 10.13165/IE-22-16-2-04

Abstract: *The technique of online behavioral advertising (OBA) is a strategy that has been widely used in the last decade by businesses and advertisers to deliver targeted advertising messages to internet users. It is done by utilizing technology to record the habits of online shoppers, including their searches and the content they visit. Users who browse the internet or use social media view advertisements relevant to their interests, recent searches, and location. We study Twitter users' attitudes about targeted ads using five different machine learning models in this research, applying the CRISP-DM framework. Our primary focus is to develop a benchmark Twitter sentiment dataset related to targeted ads and implement highly accurate machine learning algorithms to predict tweet text sentiments when discussing targeted ads. The machine learning algorithms used are Logistic Regression, Random Forest, Multinomial Naïve Bayes, Multi-Layer Perceptron, and Decision Tree. We use accuracy, precision, recall, and the F1 measure to evaluate their performance. Logistic Regression using the content-based method provides the utmost accuracy of 0.88. We propose a model that allows real-time consumer attitude research regarding retargeting ads. The results show that logistic regression is the most accurate method for predicting customer responses to OBA campaigns and that retargeting and OBA often cause negative feelings in consumers.*

Keywords: *online targeted advertising, behavioral advertising, behavioral targeting, retargeting, machine learning, Twitter sentiment dataset*

JEL Codes: *M31; M37; D83.*

[1] Corresponding author

Introduction

Today, companies that want to advertise to a large consumer audience that uses the internet and social media use online targeted advertising techniques, using advertising platforms such as Google Ads and Facebook Ads (Boerman et al., 2017). In the literature, we often come across similar terms to define this targeting technique, such as online behavioral advertising (OBA), behavioral targeting, data-driven digital advertising, and online profiling (Bennett, 2010). Businesses can target a specific audience based on their demographic profile, online behavior, including their recent searches, and other preferences related to their lifestyle and values (Zuiderveen Borgesius, 2015). Targeted ads therefore target a specific audience who are more likely to be interested in the services and products of the advertised business (Schumann et al., 2014).

The use of this technique based on big data analysis proves to be particularly effective for advertising companies as their advertising message and advertising budget is not disseminated to a wide range of audiences as is the case with traditional advertising methods such as tabloids, television, or the print media. Instead, it is limited to a targeted audience, reducing costs per conversion (Chen & Stallaert, 2014).

Our research focuses on automatically analyzing sentiments using machine learning techniques when people are talking about ads they are getting. This research aims to analyze social media users' opinions about targeted ads, assess whether the user's opinion is optimistic or destructive, and develop a benchmark dataset for sentiment analysis for its potential applications in marketing. We applied and compared multiple machine learning algorithms to find the most optimum technique.

2. Theoretical background

Marketers have been thinking about, arguing over, and implementing personalization tactics to develop effective media strategies (Iyer et al., 2005). According to users' locations, culture, and interests, marketers consume their data to show them ads related to their choice and what they are looking for online, helping companies reach more reliable customers (Johnson & Grier, 2011).

As internet and social media users grow progressively, many agencies and companies use social media to promote their products or services and impact the customer's purchase intention (Alalwan, 2018; Voorveld et al., 2018). Many social media platforms like Instagram, Facebook, Twitter, and YouTube primarily make money by implementing online targeted advertising features (Mitra & Baid, 2009).

Social media users, in particular, supply a wealth of information about their interests on their profiles, making them a niche market for advertisers looking to target them with advertisements based on the information and searches they provide. Technically, advertising platforms can utilize browser history, users' search history, and their interests as manifested from their behavior on social networks using cookies and big data analysis techniques (Boerman et al., 2017). The method of OBA is based on monitoring the websites that a user visits on the internet and the actions that they generally perform online, such as purchases of products and services, to capture their preferences and interests and then promote relevant ads (Varnali, 2021).

However, users seem to have lost confidence in the way businesses use digital media, especially when it comes directly to personal data (Kim & Huh, 2017). A typical example and operative cause of this controversy by consumers towards companies is the scandal that arose from Facebook and Cambridge Analytica (Heawood, 2018). At this point, it is worth mentioning that the state expects companies to give more control to their users regarding the stored cookies and the processing of personal data, as has been legislated in the GDPR for Europe, the Personal Information Protection Law (PIPL) in China, the Data Protection Act (DPA) in the United Kingdom, the Digital Charter Implementation Act (DCIA) in Canada, and the California Consumer Privacy Act (CCPA) in the United States (Barrett, 2019). These legal frameworks are interdependent but move toward protecting personal data (Voss, 2021).

Various studies indicate that there should be a balance between the use of personal data and the possible invasion of privacy, as it has been shown that a privacy protection policy can be beneficial for both clients and marketers (Kox et al., 2017). Other research shows that businesses often should not apply retargeting techniques (Shin & Yu, 2020), and some researchers have shown that the trusted platform matters (Kim et al., 2019). In addition, the importance of proper targeting is highlighted, as if a targeted ad appears to an irrelevant audience, then they will have a negative attitude towards the targeted advertisement and the advertised product (Cyril de Run, 2007). Similarly, the research of Boerman et al. concludes that OBA's success depends on factors controlled by the advertiser, such as the sensitive or non-sensitive information used to personalize the ad, and consumer-related factors, such as the advertiser's trust in the advertiser's perceived usefulness, feelings of intrusion, and concerns about the privacy of personal data (Boerman et al., 2017).

Researchers have followed various research methodologies to study the influence of OBA and consumer responses, including empirical research with questionnaires (Aiolfi et al., 2021; Beak & Morimoto, 2012) or experimental research (Jai et al., 2013). The subject matter of the OBA is and will continue to be of concern to the research community in the coming years with the advent of the IoT, and the introduction of targeted advertising in media other than computers and mobile phones. Hence, the need to further study the phenomenon with new tools and methods is considered significant (Aksu et al., 2018). Kumar and Gupta (2016) suggested using big data to analyze OBA effectiveness, and this research is also moving in this direction. As presented in the following sections, we analyzed 80,000 tweets regarding users' opinions on targeted advertising and studied different machine learning algorithms.

3. Methodology

People use social media platforms to talk about the ads they are interacting with daily – whether they are related to their interests or a source of frustration. Typical examples are tweets where users sometimes express their satisfaction or dissatisfaction with targeted ads.

We used Twitter to extract data from social media platforms, as Twitter is one of the eminent platforms where users tweet daily about their experiences in every part of life. Another reason to select Twitter is that it is an open-source platform that allows us to scrape a large amount of data using the Twitter API (Makice, 2009).

Our research focuses on automatically analyzing sentiments using machine learning techniques when people are talking about ads they are getting. Our primary focus is developing a benchmark Twitter sentiment dataset related to targeted ads and implementing highly accurate machine learning algorithms to predict tweet text sentiments when discussing targeted ads.

Our focus is on automatically generating precise sentiments of tweets related to targeted ads. To develop and deploy an optimized machine learning model, there is a need to follow a structured approach: cross-industry process for data mining (CRISP-DM) is the most basic and widely used framework for machine learning and data science projects (Schröer et al., 2021; Wirth & Hipp, 2000). The CRISP-DM framework embraces six significant phases: business understanding; data understanding; data preparation; modeling; evaluation; and deployment. Although the framework is flexible, the steps involved to complete all these phases must be specific. In the following sections, we follow all stages of the CRISP methodology in detail to execute our research. The flow diagram in Figure 1 describes all of the steps we performed to carry out the task in detail, from data collection to model deployment.

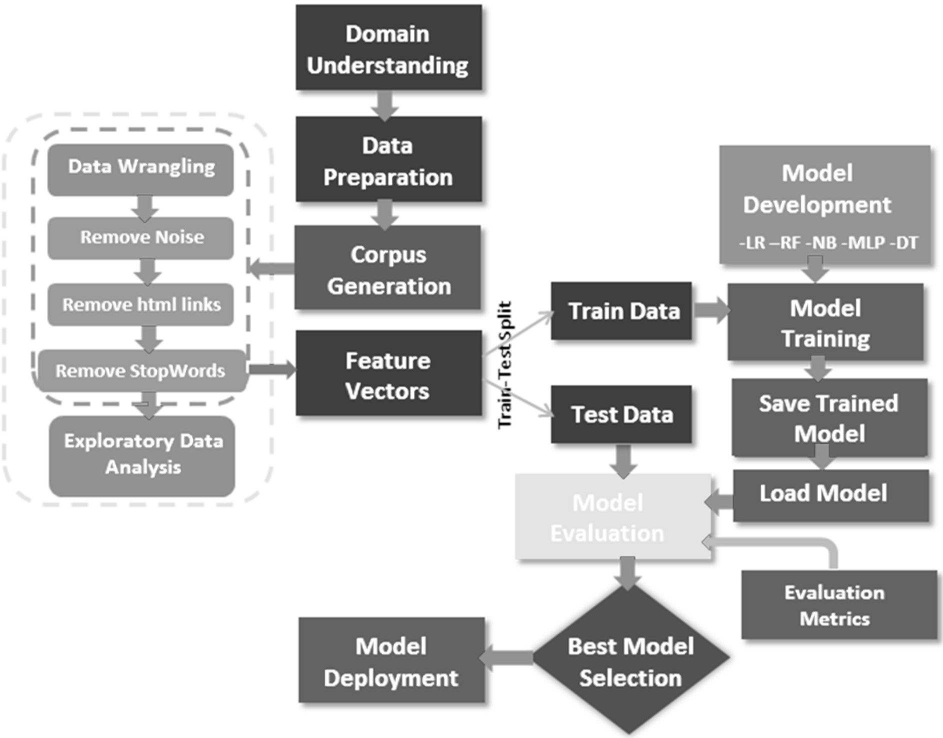

Figure 1. The machine learning model, a detailed flow diagram

3.1. Data Preparation

3.1.1. Corpus Generation Process

3.1.1.1. Data Scraping Methodology and Annotation Process

We scraped Twitter data using the open-source Twitter API. We used the Tweepy library to extract tweets (Wisdom & Gupta, 2016). Some initial conditions were set before scraping: ignore re-tweets; all tweets must be in the English language; timeframe of scraped tweets between 2019 and 2021; tweet must have one of the mentioned tags: #targeted ads, #remarketing, #retargeting, #facebook exchange (FBX), #facebook dynamic ads; only extract text tweets and ignore all images.

The raw data scraped from Twitter had some meta information attached with Twitter text. Thus, the scraped data included 24 features, but after rendering to our requirements, we dropped all other parts and kept only original tweet text. The total number of raw scraped tweets was 80,000.

To produce highly accurate models, we need a large amount of data. So, for the annotation of positive or negative sentiments related to a specific tweet, we used a python-based textblob library that assigns emotions to the dataset according to polarity in the text (Kulkarni & Shivananda, 2021). To analyze the accuracy of the labeled tweets, we manually evaluated 5,000 randomly chosen tweets. After the annotation, we dropped some positive tweets, as positive sentiment had a high weighting in the dataset to keep it balanced. After dropping the positive tweets, the annotated corpus size was 66,000 instances.

3.1.2. Corpus Characteristics and Potential Applications

Data source: we scraped data from Twitter. All tweets that we chose to extract were related to targeted ads and marketing.

Data timeline: all extracted tweets were published between 10 October 2019 and 2 August 2021.

Data tags: to scrape Twitter data we focused on tags including #targeted ads, #remarketing, #retargeting, #facebook exchange (FBX), #facebook dynamic ads.

Applications: the dataset was developed to have potential applications in sentiment analysis tasks not limited to marketing; it can be used in any domain where the problem involves analyzing sentiment from a short text.

3.1.3. Data Wrangling

Some example tweets before data pre-processing:

Had a busy weekend! Just moved to Vancouver... I wonder how my targeted ads will potentially change. haha. #BMC382

Having a strong ***ðŸ’ª*** *, activeðŸ”, and engagingðŸ˜‹* ***Instagram presence is an obvious restaurant marketing tip. But are you finding new customers with great images of your food in targeted ads on Instagram?***

To sanitize the dataset, the following data cleaning steps were performed: remove non-ASCII characters; remove HTML links, all #tags, and @tags; remove all punctuation marks; remove all numeric values; remove single characters; remove more than one whitespace in the text (including double space, \t, \n); lowercase letters; remove leading and trailing whitespaces; remove all null values; eliminate stop words; and exclude tweets with a length lower than three characters.

3.1.4. Exploratory Data Analysis

All the statistics mentioned here are from after data pre-processing. Table 1 depicts the top 10 most frequent words in the dataset. It demonstrates that the most frequent words were "targeted" and "ads" because we intentionally focused on these keywords during data scraping.

Table 1. Top 10 most common words in the dataset excluding stop words

#	Common Words	Frequency
1	targeted	64,745
2	ads	52,130
3	ad	20,209
4	facebook	8,116
5	get	8,090
6	like	7,254
7	getting	7,141
8	know	4,322
9	people	4,174
10	one	4,001

Table 2. Lexical analysis of the dataset

#		Positive	Negative	Total
1	Words (Tokens)	435,850	493,915	929,765
2	Word Types (Unique Tokens)	30,405	30,431	43,961
3	Average Tweet Length	13.77	15.12	14.46

Table 3 depicts the total instances in the dataset and the ratio of positive sentiments and negative sentiments in our dataset.

Table 3. Positive to negative tweets ratio in the pre-processed dataset

#	Sentiment	Frequency
1	Positive	32,654
2	Negative	31,636

3	Total	64,290

Some examples of positive tweets:

1. targeted ads facebook wedding dresses engagement rings;
2. targeted ads love candle;
3. interesting ad placement talk targeted product placement;
4. best targeted campaign ad ever;
5. these targeted ads are really pulling through helping me make my xmas list;
6. targeted ads have given me some amazing book recommendations.

Some examples of negative tweets:

1. phone spying enough give wildly specific targeted ads least spy enough;
2. absolutely insane targeted ad;
3. these targeted ads are getting weird;
4. i hate targeted ads so much;
5. these targeted ads are getting creepy how do they know about my life;
6. how do I stop getting targeted ads for universities i am already a part of.

As can be observed from some of the above example tweets, some of the pre-processed sentences do not make sense according to standard grammatical English language rules. We removed stop words and punctuation and performed other data cleaning steps. However, the clean data is more meaningful to machine learning algorithms. For example, stop words are used to build up sentence structure, but these words have no relevance when identifying sentiments in the text.

3.1.5. Content-based Method

A TFIDF (Term Frequency–Inverse Document Frequency) vectorizer was used to extract the frequency ratio of word unigrams from the dataset. The total number of features we found in the corpus was 43961. To keep the most important and relevant features and reduce complexity during model training, we tweaked the values of TFIDF parameters to max_dif = 0.8 and min_dif = 7. Max_dif excludes tokens with a frequency greater than the given threshold, whereas min_df excludes tokens with a frequency lower than given threshold. After setting the parameter values, we reduced the number of features to 9,407.

3.1.6. Train-Test Split

The dataset was split into two parts to train and evaluate the models' performance. Using 80:20 ratios, we randomly split the data into train and test groups. Our prediction models were then built with training data and tested against a test set.

3.2. Modelling

After cleaning and pre-processing the data in a machine-understandable format, it was directly passed to our models. In this step, machine learning models were developed and optimized to be trained on the dataset.

3.2.1. Model Development

In consideration of automated sentiment analysis systems, the tweets in the dataset have a long sequence of words or tokens. After applying feature vectors techniques, every token has a different frequency distribution, so there is a need to identify models that can learn the sequence and recognize the significant relation between tokens. Various machine learning models may show different behavior with a dataset. In order to explore which model best fits our dataset, four different machine learning algorithms were evaluated. In addition, Artificial Neural Network techniques produce encouraging results in datasets with complex features, so neural network models with multiple hidden layers were also built using our dataset. The following algorithms were used in our research:

- Logistic Regression;
- Random Forest;
- Multinomial Naïve Bayes;
- Artificial Neural Network;
- Decision Tree.

3.2.1.1. Logistic Regression

Logistic Regression is a parametric model with specific parameters that depend upon input features to predict the output. For binary classification problems, logistic regression is preserved as binomial logistic regression. The following equation depicts the behavior of the algorithm (Hosmer Jr et al., 2013):

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

3.2.1.2. Random Forest

Random forest can be applied in both regression as well as classification problems. Random forest is an example of an ensemble learning method where multiple decision trees are designed to make predictions, and results from all trees are pooled together. For classification problems, the random forest output is the class chosen by all trees (Svetnik et al., 2003).

3.2.1.3. Multinomial Naïve Bayes

Multinomial Naïve Bayes is a classification algorithm built on Bayes' Theorem by a hypothesis of individuality amongst predictions. In modest expressions, a Naïve Bayes algorithm undertakes that the occurrence of a specific feature in a label is isolated to the presence of any other supplementary features (Abbas et al., 2019).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

- $P(c|x)$ *is the posterior probability*
- $P(c)$ *is the label's prior probability*
- $P(x|c)$ *is the likelihood*
- $P(x)$ is the prediction's prior probability

3.2.1.4. Artificial Neural Network

Artificial neural networks (ANN) can be developed and designed for supervised classification and regression tasks. ANNs can have multiple layers (da Silva et al., 2017). The simplest form of ANN with no hidden layer is perceptron, whereas an ANN with one or more hidden layers is called a multi-layer perceptron (Zou et al., 2008). ANNs learn a predictive function after training on a given dataset.

$$f(.) : R^m \rightarrow R^o$$
$$where\ m = input\ dimension$$
$$m = output\ dimension$$

Given an input vector vector X= $x_1, x_2, \ldots x_m$ and a target value y, a multi-layer perceptron can learn a nonlinear function to make future predictions. In addition to the linear relation between features, an MLP can also learn nonlinear relations between input and output variables.

3.2.1.5. Decision Tree

Decision trees are non-parametric methods designed for supervised machine learning tasks and work in a tree structure (Safavian & Landgrebe, 1991). The tree structure splits the complete dataset into subsets and continues until all possible paths or rules are met, or the tree reaches its maximum depth and makes a decision towards the final prediction.

3.2.2. Model Training

Training data was used to train the models developed in the previous step.

3.3. Evaluation

3.3.1. Evaluation Methodology

The problem of identifying sentiment from an author's tweet was treated as a supervised classification problem. The sentiment analysis task was considered a binary classification issue, with the goal of distinguishing between two classes: 1) positive, and 2) negative. We used a train-test split to experiment towards the better assessment of content-based methods. Five diverse machine learning techniques were evaluated for the classification task: Logistic Regression, Random Forest, Naïve Bayes, Multi-Layer Perceptron, and Decision Tree. For the content-based method, word gram features extracted using the TFIDF vectorizer were used as inputs for machine learning algorithms.

3.3.2. Evaluation Metrics

We used the following evaluation metrics for supervised binary classification problems: 1) accuracy, 2) precision, 3) recall, 4) F_1 (Hossin & Sulaiman, 2015).

3.3.3. Model Evaluation

Test data was used to check the performance of the model built and trained in the previous steps. Multiple evaluation metrics were used to compare the model's performance.

To perform real-time evaluations, we ensured that test data was not taken from training data; both of these datasets were disjointed.

During the testing phase, only input features were passed to the model.

We ensured that the structure and number of input features for the test data was the same as the training data to avoid any errors or false results.

3.4. Deployment

3.4.1. Application Phase

When all models were trained and evaluated, the best model with the highest accuracy and F_1 score was trained on the complete dataset and deployed for future predictions. F_1 was chosen to maintain a balance between precision and recall.

3.4.2. Predictions on Future Data

To evaluate the performance of the best model on real-time data, we took an unseen input in text format from a user and predicted positive or negative sentiment.

4. Results

The findings of all five machine learning models on test data that we examined using the content-based strategy are shown in Table 4. The Logistic Regression model outperformed all other models. F1 was also used to evaluate both positive and negative sentiment data. The results showed that the Logistic Regression classifier has the highest average accuracy (0.88), highest average precision (0.87), highest average ROC AUC score (0.90), and highest average F1 score (0.88).

Table 4. Comparison of the results of machine learning models

Model Name	Accuracy	Precision	Recall	F_1
Logistic Regression	0.883419	0.87142	0.897255	0.884149

Random Forest	0.873153	0.865936	0.880471	0.873143
Naïve Bayes	0.799113	0.837127	0.73851	0.784732
Multi-Layer Perceptron	0.876575	0.873246	0.878588	0.875909
Decision Tree	0.827112	0.833013	0.814588	0.823697

Figure 2 plots a graphical visualization of the performance of all models on the test dataset and exposes their comparison. The benchmark model with detailed results is available online at https://github.com/ziakis/retargeting_sentiment_analysis (Appendix).

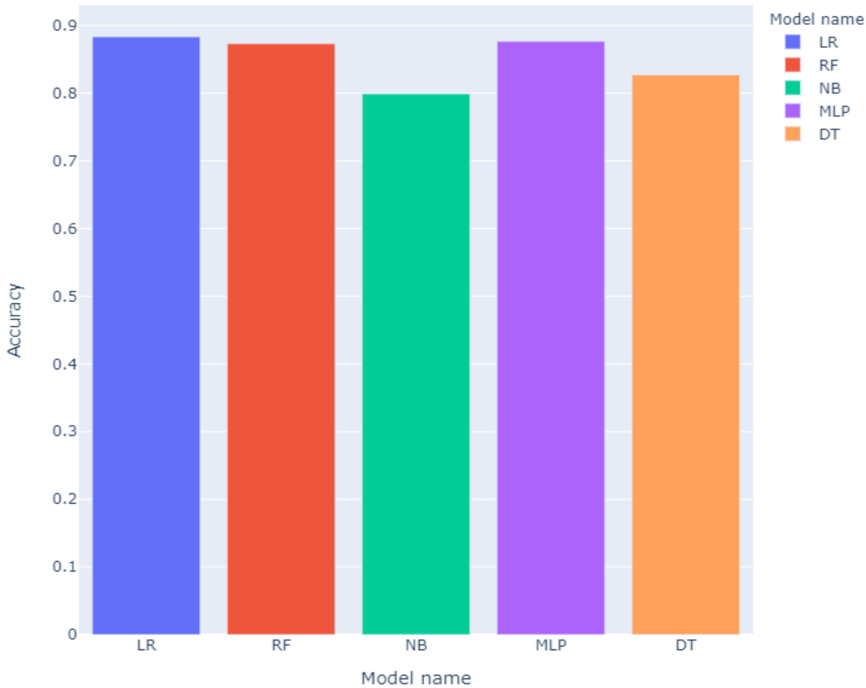

Figure 2. Accuracy comparison of multiple models

5. Conclusions and Discussion

In this project, we developed a benchmark English language Twitter sentiment analysis dataset related to tweets about targeted ads to build and assess sentiment analysis techniques. After cleaning and pre-processing, the developed dataset comprised 32,654 (51%) negative and 31,636 (49%) positive tweets. These results support the assumptions derived from our theoretical foundation, according to which we assumed that retargeting and OBA often cause negative feelings to consumers.

Furthermore, we applied and compared five existing machine learning algorithms, including Logistic Regression, Random Forest, Multinomial Naïve Bayes, Multi-Layer Perceptron, and Decision Tree, using content-based methods to pass input data to machine learning models. We used accuracy, precision, recall, and the F_1 measure to evaluate their performance. Logistic Regression using the content-based method provides the utmost accuracy of 0.88. We propose a model that allows real-time consumer attitude research regarding retargeting ads.

The study's limitations are related to the methodology used and the number of hashtags, tweets, and sources of user-generated content analyzed. In future work, we aim to develop our own proposed machine learning algorithm and address the limitations by enhancing the dataset with a more extended observation period and more social media platforms.

Our findings are relevant for advertisers and digital marketers who are advised to cautiously consider all available information sources, including social media, and utilize sentiment analysis as part of their business strategy.

References

1. Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security*, *19*(3), 62–67. http://dx.doi.org/10.13140/RG.2.2.30021.40169
2. Aiolfi, S., Bellini, S., & Pellegrini, D. (2021). Data-driven digital advertising: Benefits and risks of online behavioral advertising. *International Journal of Retail & Distribution Management*, *49*(7), 1089–1110. https://doi.org/10.1108/IJRDM-10-2020-0410
3. Aksu, H., Babun, L., Conti, M., Tolomei, G., & Uluagac, A. S. (2018). Advertising in the IoT Era: Vision and Challenges. *IEEE Communications Magazine*, *56*(11), 138–144. https://doi.org/10.1109/MCOM.2017.1700871
4. Alalwan, A. A. (2018). Investigating the impact of social media advertising features on customer purchase intention. *International Journal of Information Management*, *42*, 65–77. https://doi.org/10.1016/j.ijinfomgt.2018.06.001
5. Barrett, C. (2019). Are the EU GDPR and the California CCPA becoming the de facto global standards for data privacy and protection? *Scitech Lawyer*, *15*(3), 24–29.
6. Beak, T. H., & Morimoto, M. (2012). Stay away from me, examining the determinants of individual avoidance of personalized advertising. *Journal of Advertising*, *41*, 59–76.
7. Bennett, S. C. (2010). Regulating online behavioral advertising. *The John Marshall Law*

Review, *44*, 899–962.
8. Boerman, S. C., Kruikemeier, S., & Borgesius, F. J. Z. (2017). Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising*, *46*(3), 363–376. https://doi.org/10.1080/00913367.2017.1339368
9. Chen, J., & Stallaert, J. (2014). An economic analysis of online advertising using behavioral targeting. *Mis Quarterly*, *38*(2), 429–449.
10. Cyril de Run, E. (2007). Ethnically targeted advertising: Views of those not targeted. *Asia Pacific Journal of Marketing and Logistics*, *19*(3), 265–285. https://doi.org/10.1108/13555850710772932
11. da Silva, I. N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L. H. B., & dos Reis Alves, S. F. (2017). Artificial Neural Network Architectures and Training Processes. In I. N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, & S. F. dos Reis Alves (Eds.), *Artificial Neural Networks: A Practical Course* (pp. 21–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-43162-8_2
12. Heawood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity*, *23*(4), 429–434. https://doi.org/10.3233/IP-180009
13. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.
14. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1–11. http://dx.doi.org/10.5121/ijdkp.2015.5201
15. Iyer, G., Soberman, D., & Villas-Boas, J. M. (2005). The Targeting of Advertising. *Marketing Science*, *24*(3), 461–476. https://doi.org/10.1287/mksc.1050.0117
16. Jai, T.-M. (Catherine), Burns, L. D., & King, N. J. (2013). The effect of behavioral tracking practices on consumers' shopping evaluations and repurchase intention toward trusted online retailers. *Computers in Human Behavior*, *29*(3), 901–909. https://doi.org/10.1016/j.chb.2012.12.021
17. Johnson, G. D., & Grier, S. A. (2011). Targeting without alienating. *International Journal of Advertising*, *30*(2), 233–258. https://doi.org/10.2501/IJA-30-2-233-258
18. Kim, H., & Huh, J. (2017). Perceived Relevance and Privacy Concern Regarding Online Behavioral Advertising (OBA) and Their Role in Consumer Responses. *Journal of Current Issues & Research in Advertising*, *38*(1), 92–105. https://doi.org/10.1080/10641734.2016.1233157
19. Kim, T., Barasz, K., & John, L. K. (2019). Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness. *Journal of Consumer Research*, *45*(5), 906–932. https://doi.org/10.1093/jcr/ucy039
20. Kox, H. L. M., Straathof, B., & Zwart, G. (2017). *Targeted Advertising, Platform Competition, and Privacy* (SSRN Scholarly Paper ID 3021392). Social Science Research Network. https://doi.org/10.1111/jems.12200
21. Kulkarni, A., & Shivananda, A. (2021). Exploring and Processing Text Data. In A. Kulkarni & A. Shivananda (Eds.), *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning Using Python* (pp. 31–62). Apress. https://doi.

org/10.1007/978-1-4842-7351-7_2

22. Kumar, V., & Gupta, S. (2016). Conceptualizing the evolution and future of advertising. *Journal of Advertising*, *45*(3), 302–317. https://doi.org/10.1080/00913367.2016.1199335
23. Makice, K. (2009). *Twitter API: Up and running: Learn how to build applications with the Twitter API*. O'Reilly Media, Inc.
24. Mitra, P., & Baid, K. (2009). Targeted advertising for online social networks. *2009 First International Conference on Networked Digital Technologies*, 366–372. https://doi.org/10.1109/NDT.2009.5272165
25. Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674.
26. Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199
27. Schumann, J. H., von Wangenheim, F., & Groene, N. (2014). Targeted Online Advertising: Using Reciprocity Appeals to Increase Acceptance among Users of Free Web Services. *Journal of Marketing*, *78*(1), 59–75. https://doi.org/10.1509/jm.11.0316
28. Shin, J., & Yu, J. (2020). *Targeted Advertising and Consumer Inference* (SSRN Scholarly Paper ID 3688258). Social Science Research Network. https://doi.org/10.2139/ssrn.3688258
29. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958. https://doi.org/10.1021/ci034160g
30. Varnali, K. (2021). Online behavioral advertising: An integrative review. *Journal of Marketing Communications*, *27*(1), 93–114. https://doi.org/10.1080/13527266.2019.1630664
31. Voorveld, H. A., Van Noort, G., Muntinga, D. G., & Bronner, F. (2018). Engagement with social media and social media advertising: The differentiating role of platform type. *Journal of Advertising*, *47*(1), 38–54. https://doi.org/10.1080/00913367.2017.1405754
32. Voss, W. G. (2021). *The CCPA and the GDPR Are Not the Same: Why You Should Understand Both* (SSRN Scholarly Paper ID 3769825). Social Science Research Network. https://papers.ssrn.com/abstract=3769825
33. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, *1*, 29–39.
34. Wisdom, V., & Gupta, R. (2016). *An introduction to twitter data analysis in Python*. Artigence Inc. Working Paper. http://dx.doi.org/10.13140/RG.2.2.12803.30243
35. Zou, J., Han, Y., & So, S.-S. (2008). Overview of artificial neural networks. In D. J. Livingstone (Eds.), *Artificial Neural Networks: Methods in Molecular Biology* (Vol. 458, pp. 14–22). Humana Press. https://doi.org/10.1007/978-1-60327-101-1_2
36. Zuiderveen Borgesius, F. (2015). *Improving privacy protection in the area of behavioural targeting*. https://dx.doi.org/10.2139/ssrn.2654213

Appendix

List of files on GitHub:

- https://github.com/ziakis/retargeting_sentiment_analysis/blob/main/Twitter_sentiment_analysis_targeted_ad.html
- https://github.com/ziakis/retargeting_sentiment_analysis/blob/main/Twitter_sentiment_analysis_targeted_ad.ipynb
- https://github.com/ziakis/retargeting_sentiment_analysis/blob/main/twitter_sentiment_dataset.csv
- https://github.com/ziakis/retargeting_sentiment_analysis/blob/main/twitter_sentiment_preprocessed_dataset.csv