Journal of Business Data Science Research

Categorizing Online Advertisements with Machine Learning Models

Mohammad Taghi Taghavifard[a], Pegah Farazmand[b,*], Khatereh Farazmand[c], Saeedreza Jadidi[d]

[a] Department of Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran
[b] Department of Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran
[c] Department of Health, Faculty of Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[d] Department of Mechanical, Industrial and Aerospace Engineering, Concordia University, Montreal, QC, H3G 1M8, Canada

ABSTRACT

Advertising segmentation, also known as audience or market segmentation, constitutes a critical facet of marketing. It involves subdividing a large target population into more manageable groups or segments. By employing advertising grouping, businesses can tailor content and keywords to each group of advertisements, thereby enhancing the effectiveness of their marketing efforts. This project explores two machine learning models for advertising grouping, achieving an impressive efficiency rate of over 90%. The study encompasses a dataset of 2000 online advertisements divided into five advertising groups, with 1600 data points used for training and 400 for testing. The results indicate that the Support Vector Classifier (SVC) and Linear Regression (LR) models perform well in the categorization of advertisements.

KEWWORDS: Machine Learning Models, Advertising Segmentation, Target Audience, Online Ads.

ORIGINAL ARTICLE

Received 28 November 2023
Accepted 25 February 2024
Published 5 March 2024

CITATION

Taghavifard et al. (2024). Categorizing Online Advertisements with Machine Learning Models. Journal of Business Data Science Research, 3(1), 12–20.

2024 © Journal of Business Data Science Research Ltd

1. Introduction

Advertising is a fundamental method employed by firms to introduce potential customers to their products or services, with the ultimate goal of persuading them to make a purchase or utilize the services offered (Shah et al., 2020). Among the various strategies used to manage advertisements, ad grouping stands out as one of the most effective. Ad groups provide intricate data, encompassing ad content and keywords, facilitating the precise targeting of a particular audience. This practice, known as advertising segmentation, involves directing marketing efforts toward specific consumer groups within larger markets (PPCEXPO, 2022). By leveraging advertising grouping, businesses can convey their message more effectively and align their marketing objectives with the preferences of their target audience (PPCEXPO, 2022).

The advantages of advertising segmentation extend across companies of all sizes and play a pivotal role in any successful marketing campaign (PPCEXPO, 2022). A study revealed that businesses with robust market segmentation strategies outperformed their counterparts, increasing their revenue by 10% over a five-year period (PPCEXPO, 2022). To enhance the efficacy of advertisements across diverse target audiences, a multitude of advertising tactics and strategies have been explored (Choi and Lewis, 2017). Notably, the advent of Artificial Intelligence (AI) technologies has revolutionized the advertising landscape by offering enhanced computational capabilities for optimizing digital advertisements. Machine Learning (ML)-based techniques have significantly improved targeting accuracy by predicting the most relevant advertisements for viewers based on contextual or historical user data (Chen et al., 2009). These advancements in AI and data-driven approaches not only enhance the user experience but also alleviate the challenges faced by advertisers.

* Corresponding author: pegahfarazmand1367@gmail.com
• http://jbdsr.com/index.php/home/article/view/13

Numerous studies have explored machine learning algorithms that fine-tune the distribution of tailored advertisements, making them more effective (Capatina et al., 2020). Moreover, in a bid to discern the unique features of advertisements in various industries and effectively present them to the target audience, this study employs machine learning models to categorize five distinct groups of online advertisements (Abrahams et al., 2013).

In summary, advertising segmentation, bolstered by AI and ML, has become an indispensable tool for businesses seeking to connect with their desired customer base and optimize their marketing investments.

2. Previous Research Studies

Online advertising has significantly advanced through the integration of AI technologies, particularly in the realm of expanding target audience reach. Numerous studies have delved into the automatic classification of text documents, advertisements, and various sectors. Abrahams et al.'s work in 2012 is a noteworthy example in this domain. They focused on foreign advertisements and employed a comprehensive set of 103 characteristics to delineate the emotional and semantic content of advertising. Their study used these variables to categorize commercials and distribute them across different media channels through neural networks (Abrahams et al., 2013).

Another significant project was undertaken by Boselli et al. in 2018, emphasizing labor market intelligence, an increasingly vital area for the development and evaluation of EU labor market regulations. Their research revolved around the creation of a ML model for categorizing job openings and outlined their approach to Web Labor Market Intelligence along with three real-world application scenarios. Their approach to labor market intelligence offers several advantages: (I) it reduces time-to-market compared to conventional survey-based analyses; (II) it supports multiple languages, bridging linguistic boundaries between nations using standard classification systems rather than proprietary ones; (III) it enables the mapping of generated knowledge across diverse dimensions, such as territory, sectors, contracts, and more, at various levels of granularity; and (IV) it allows for the assessment and comparison of different labor markets worldwide to facilitate evidence-based decision-making (Boselli et al., 2018).

Shah et al.'s 2020 research endeavors to address the significance of the transition in the advertising process by analyzing before-and-after statistical observations of businesses adopting AI and ML. Their findings suggest that the outcomes themselves serve as a logical source of motivation for businesses yet to embrace this shift. The primary challenge lies in ensuring that AI is used in a manner that doesn't raise concerns about individual privacy. The solution may entail the establishment of regulations, their enforcement, or the creation of a dedicated regulatory body to ensure lawful and ethical use of this technology. The future of advertising appears promising, with AI and ML empowering firms to exercise greater control, ultimately leading to a noticeable and substantial improvement, as perceived by consumers [1].

In the research conducted by Choi et al. in 2020, various ML approaches aimed at enhancing targeted online advertising were explored and categorized. They identified twenty-three ML-based tactics for online targeted advertising, primarily dividing them into content- and user-centric strategies. The report also sheds light on algorithm-based click fraud detection, a somewhat overlooked aspect, illustrating how ML methods can be combined to maintain the profitability of online advertising (Choi and Lim, 2020).

3. Methods and Materials

3.1. Determination of Input and Output Variables

In configuring the variables for this project, we drew inspiration from Ayodeji's work in 2022, which focused on detecting fake news using Python (Ayodeji, 2022). The dataset utilized for the classification of advertisements includes information pertaining to the title of the ads and their bold text (inputs), as well as the label of the ads (output) indicating the category to which each advertisement belongs.

3.2. Data Collection

Data acquisition involved selecting five distinct advertising groups: automotive, home decoration, clothing, movies, and reading materials. We collected a total of 400 advertisements from each group, resulting in a comprehensive dataset comprising 2000 advertisements. Each advertisement was sourced online from platforms such as Google and websites like Amazon. Figure 1 illustrates the distribution of advertisements across these groups.

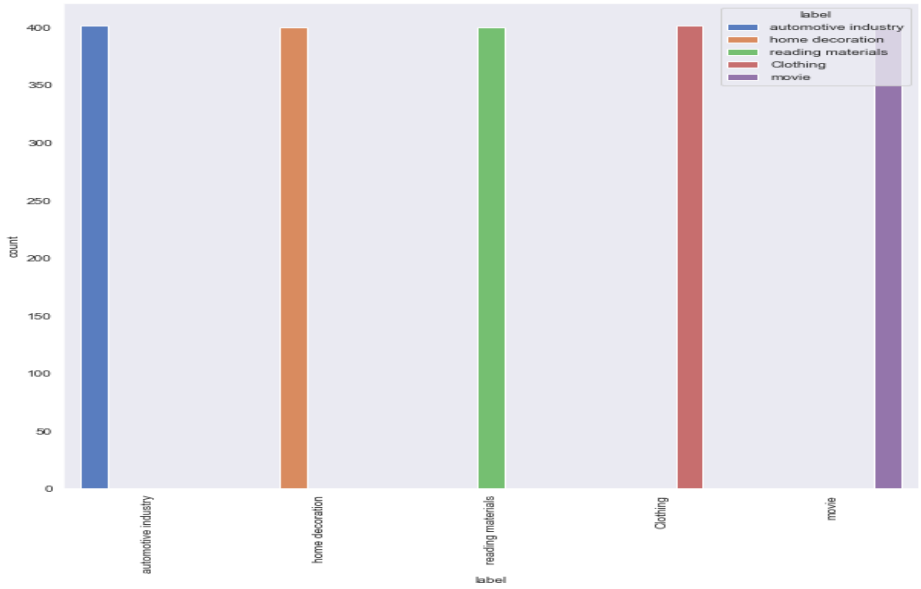


Figure 1: Advertising groups.

To identify the most commonly used terms distinctive to each category of advertisements, we employed Python programming. The frequently used phrases in the automotive sector are depicted in Figure 2.

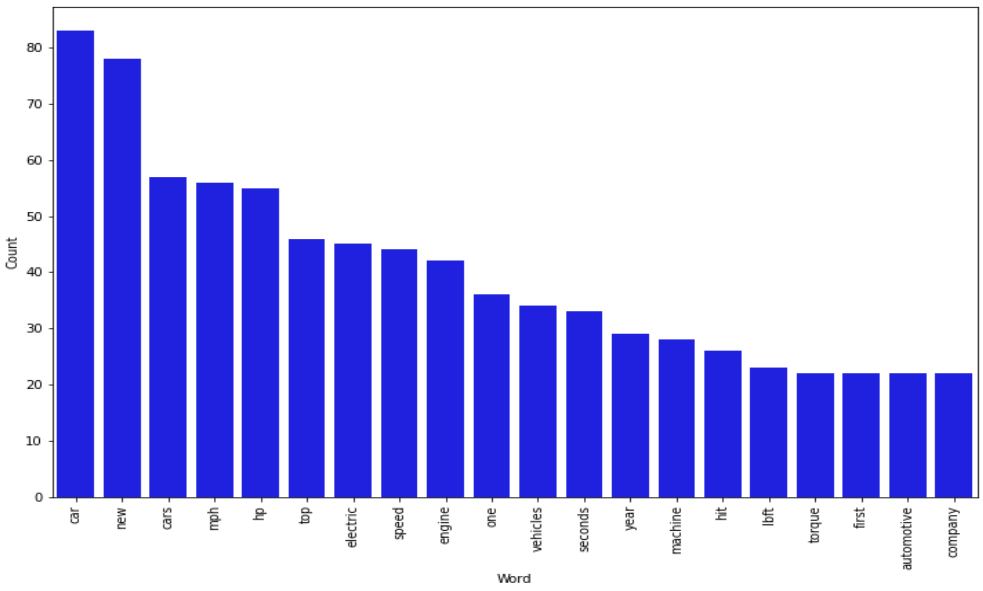


Figure 2: Frequently used phrases in the automotive industry.

Figure 3 highlights the commonly used terms in the home decoration industry.

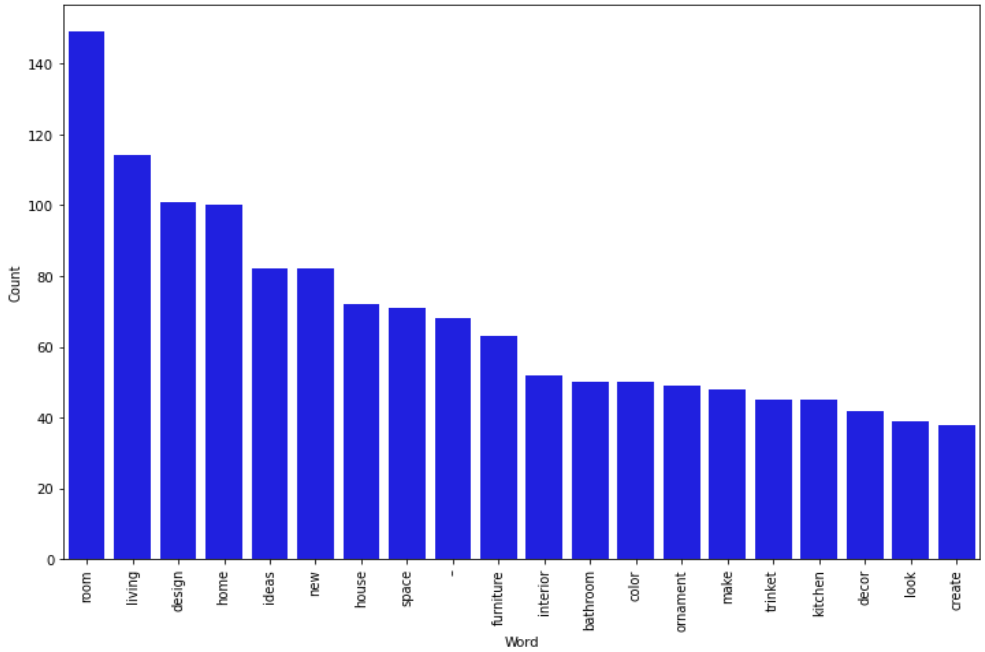


Figure 3: Frequently used phrases in the home decoration sector.

In Figure 4, the terms most frequently used in the reading materials sector can be observed.

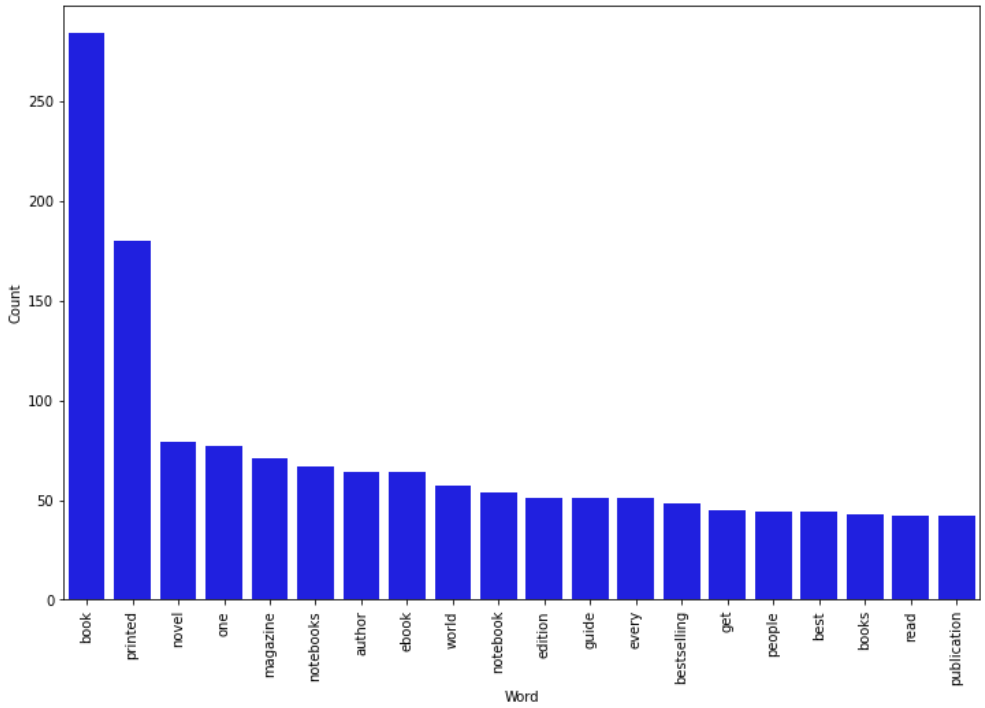


Figure 4: Frequently used phrases in the reading materials sector.

Figure 5 showcases the frequently used terms in the clothing sector.

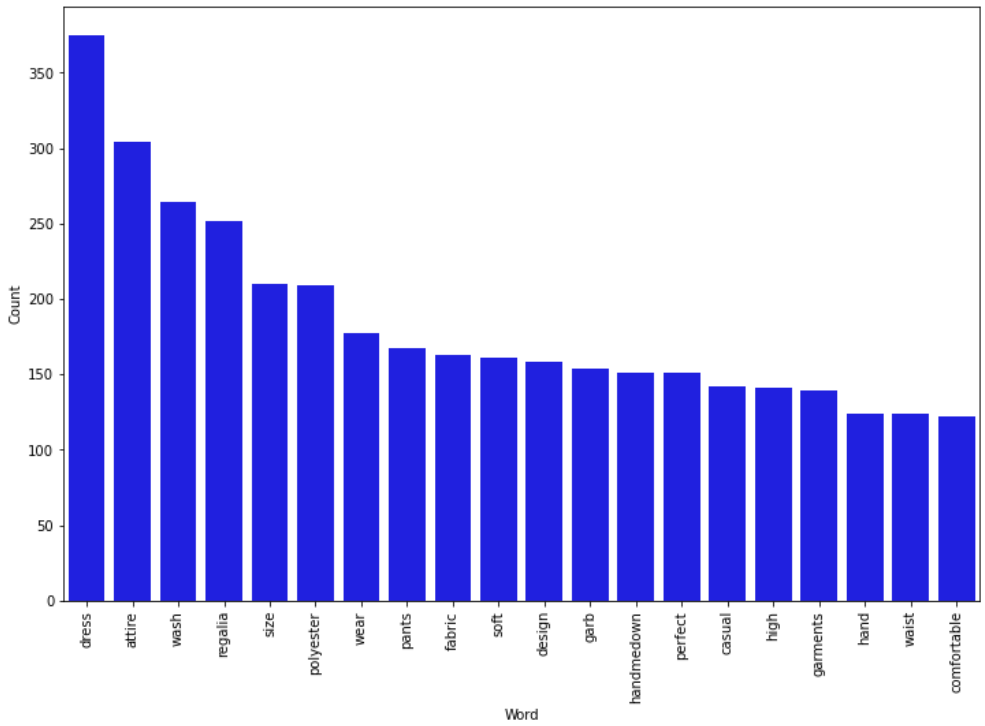


Figure 5: Frequently used phrases in the clothing sector.

The phrases that appear most frequently in the movie industry are presented in Figure 6.

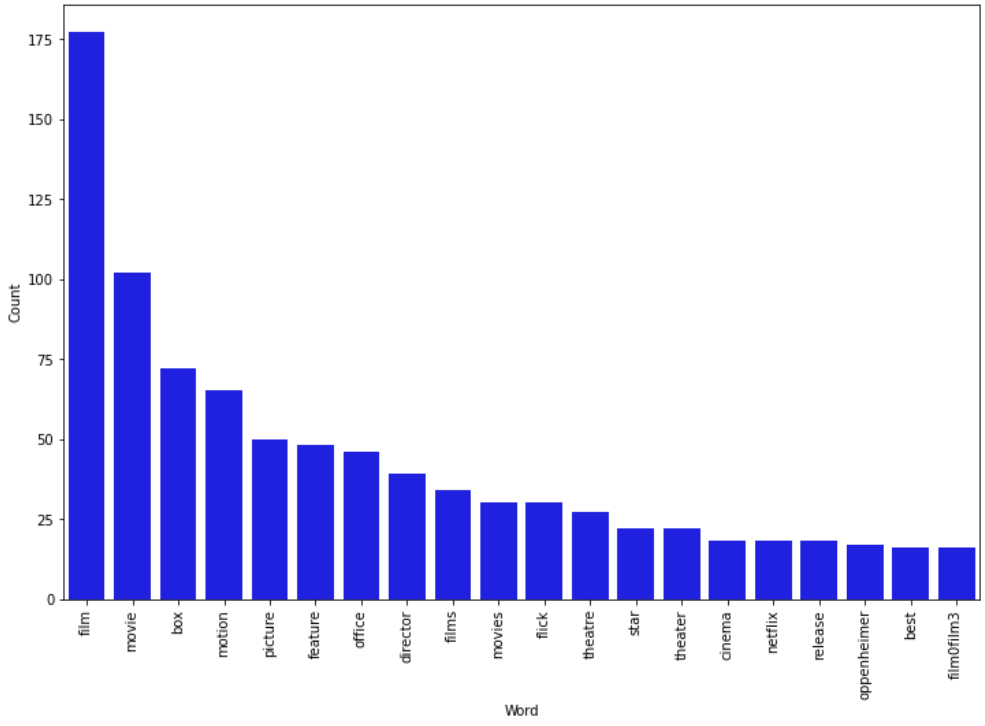


Figure 6: Frequently used phrases in the movie sector.

3.3. Model Selection

Our selection of models for this study was based on their demonstrated performance in ad grouping. The two models chosen were the Support Vector Classifier (SVC) and Linear Regression (LR). LR simplifies the estimation process and provides clear interpretability at the modular level, specifically with regard to weights (Hastie et al., 2009). In its basic form, the linear regression equation is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

where y is the dependent variable (output), x_1, x_2, …, x_n are the independent variables (inputs), and β_0, β_1, β_2, …, β_n are the coefficients.

LR is typically employed in supervised learning when the prediction target is discrete or quantitative. It attempts to establish a linear relationship between the dependent variable (y) and one or more related independent variables (x_1, x_2, …, x_n) by fitting the best-fit regression line (Raj, 2020). To apply these models effectively, we needed to convert words into numerical values, a process accomplished through Python programming, guided by Belhoudi's work in 2021 (Belhoudi, 2021).

Support Vector Machines (SVMs) offer numerous advantages, including their ability to handle high-dimensional data and perform well with smaller datasets. They are also capable of modeling non-linear decision boundaries, expanding their applicability. It is important to clarify the distinction between SVM and SVC, with SVC specifically referring to the linear condition when the hyperplane used for classification is in a linear state. The most suitable ML algorithm for our project was found to be Linear SVC. The primary objective of Linear SVC is to categorize data by determining the "best fit" hyperplane based on the provided data. This is achieved by solving the following optimization problem:

$$minimize \ \frac{1}{2}\|w\|^2 \quad (2)$$

$$subject\ to\ y_i(w.a_i - b) \geq 1\ for\ i = 1,2,\ldots,n$$

where w is the weight vector, b is the bias term, a_i are the feature vectors, y_i are the class labels, and n is the number of data points. This method is particularly well-suited to our project as it allows us to feed features to the classifier after acquiring the hyperplane to predict the class.

3.4. Model Training

The data was collected and organized in an Excel file, which was then saved as a CSV file for use in Python programming, where the training and testing of the data with the selected models were conducted. All necessary libraries for the project were imported, including Pandas for efficient dataset management and data importation, as well as other essential libraries such as NumPy and the Sci-kit Learn library. Of the 2000 data points, 1600 were allocated for training, while the remaining 400 were used for testing. The outcomes of our investigations indicate that the LR model and SVC both achieved impressive accuracy, with accuracy rates of 92.25% and 94.5%, respectively, in classifying the data groups.

4. Results and Discussions

The two selected models underwent initial testing and training with the dataset. To evaluate their performance on the data, a confusion matrix was employed. A confusion matrix is used to assess the effectiveness of a classification model within an $N \times N$ matrix, where N represents the number of target classes. It compares the predictions of the ML model against the actual target values. Based on the data matrix from our research, both models excelled in identifying and forecasting data. However, the LR model exhibited slightly weaker performance than the SVC model in the specific category related to the automotive sector. This variance was relatively minor. The confusion matrices for both the SVC and LR models can be found in Figures 7 and 8, respectively.

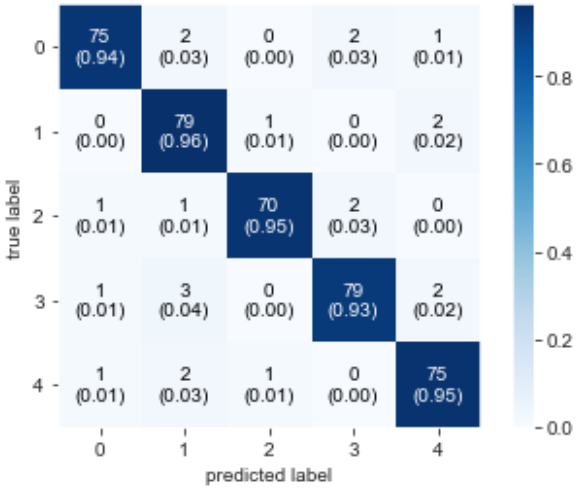


Figure 7: SVC confusion matrix.

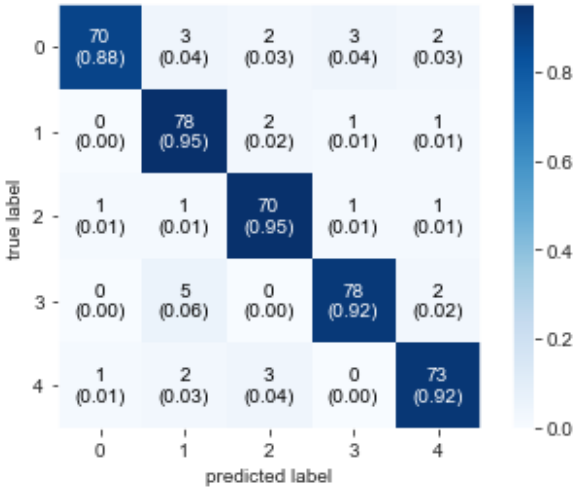


Figure 8: LR confusion matrix.

A strong model demonstrates its strength by achieving low false positive (FP) and false negative (FN) rates, coupled with high true positive (TP) and true negative (TN) rates. TP corresponds to instances where both the projected value and the actual value are positive, indicating accurate positive predictions. On the other hand, TN refers to cases where both the prediction and actual value are negative, signifying accurate negative predictions. Conversely, FP represents situations where the model's prediction is positive, yet the actual outcome is negative, resulting in what is known as a Type 1 mistake. FN denotes scenarios where the model forecasts a negative outcome, but the actual result is positive, constituting a Type 2 mistake. The SVC model notably excelled in achieving a high number of TP predictions, as indicated by the matrix.

It is evident that the LR model performed well in predicting all groups, except for the category associated with the automotive industry, where its performance lagged behind other categories. This discrepancy in performance, albeit relatively modest, can be attributed to the broad vocabulary of the data. It suggests that this weakness could potentially be mitigated by using more relevant keywords. Ultimately, the SVC model, boasting an accuracy of 94.5%, outperformed the LR model, which achieved 92.25% accuracy. This is demonstrated in Figure 9, which compares the performance of these two models. The results indicate that the SVC model offers superior performance, particularly in the context of classifying advertisements within the chosen categories.

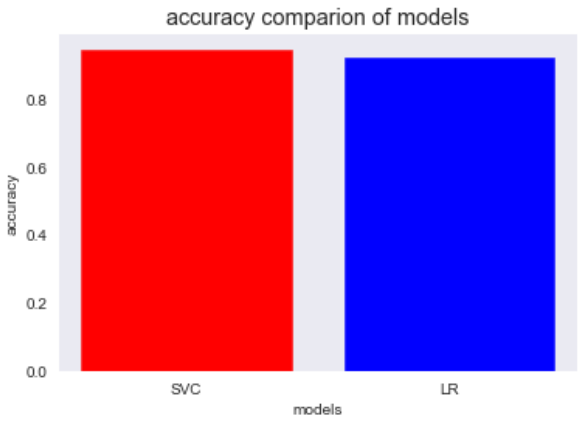


Figure 9: Accuracy of SVC and LR models.

5. Conclusion

The term "to advertise" may initially mean "to show" or "to display," but for businesses, it signifies the act of capturing the attention of the general public, potentially transforming them into customers. Advertising serves a multitude of purposes, including creating demand for products and informing customers about new features and advantages. The specific message a company wishes to imprint in the minds of its target audience determines the choice of their advertising strategy. Advertising is incredibly versatile, capable of addressing a broad spectrum of people, or alternatively, it can be meticulously tailored to a specific, narrowly defined audience segment. This choice hinges on the objectives and the media used for audience engagement (Shah et al., 2020).

Effective management of advertising strategies is pivotal, and one method to achieve this is through the technique of grouping advertisements. Ad groups play a crucial role in instructing search engines to display specific ads in response to particular keyword searches. They also encompass vital details about ad content and the corresponding landing pages. In order to create successful ad groups, businesses must possess an understanding of their target audience or customer base, as well as an insight into the content that piques their interest, thus encouraging prolonged engagement with their websites.

Artificial Intelligence (AI), particularly Machine Learning (ML) - a subset of AI, has emerged as a powerful tool for organizing advertisements. In many cases, ML outperforms human operations in terms of speed and efficiency, thus surpassing natural intelligence in this domain (Jha et al., 2019; Pathan et al., 2020). In the scope of this research, ML models, including Support Vector Classifier (SVC) and Linear Regression (LR), were harnessed for the purpose of grouping internet advertisements. We examined five distinct advertising groups and curated a list of key terms for each category. These models were trained on the data and subsequently utilized to predict the appropriate advertising group for new entries into the network by comparing their textual content to pre-prepared word clusters.

The outcomes of this study indicated that the SVC model demonstrated a marginal advantage over the LR model, although both exhibited exceptional proficiency in classifying and predicting advertising groups. These models hold significant potential for implementation within businesses to enhance their performance in the field of ad grouping.

References

Abrahams, A. S., Coupey, E., Zhong, E. X., Barkhi, R. and Manasantivongs, P. S. (2013), "Audience targeting by B-to-B advertisement classification: A neural network approach", Expert systems with applications, Vol. 40, No. 8, pp. 2777-2791.

Ayodeji, S. O. 2022. Fake News Detection with Python [Online]. Available: https://heartbeat.comet.ml/fake-news-detection-with-python-d7339cf1f018 [Accessed 2022].

Belhoudi, J. 2021. Fake news Classification using Logistic Regression [Online]. Available: https://www.kaggle.com/code/jihenbelhoudi/fake-news-classification-using-logistic-regression [Accessed 2022].

Boselli, R., Cesarini, M., Mercorio, F. and Mezzanzanica, M. (2018), "Classifying online job advertisements through machine learning", Future Generation Computer Systems, Vol. 86, No., pp. 319-328.

Capatina, A., Kachour, M., Lichy, J., Micu, A., Micu, A.-E. and Codignola, F. (2020), "Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations", Technological Forecasting
Social Change, Vol. 151, No., pp. 119794.
Chen, Y., Kapralov, M., Canny, J. and Pavlov, D. (2009), "Factor modeling for advertisement targeting", Advances in neural information processing systems, Vol. 22, No., pp.
Choi, J.-A. and Lewis, R. (2017), "Culture and the star-power strategy: Comparing American and Korean response to celebrity-endorsed advertising", Journal of Global Marketing, Vol. 30, No. 1, pp. 3-11.
Choi, J.-A. and Lim, K. (2020), "Identifying machine learning techniques for classification of target advertising", ICT Express, Vol. 6, No. 3, pp. 175-180.
Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009), The elements of statistical learning: data mining, inference, and prediction, Springer.
Jha, K., Doshi, A., Patel, P. and Shah, M. (2019), "A comprehensive review on automation in agriculture using artificial intelligence", Artificial Intelligence in Agriculture, Vol. 2, No., pp. 1-12.
Pathan, M., Patel, N., Yagnik, H. and Shah, M. (2020), "Artificial cognition for applications in smart agriculture: A comprehensive review", Artificial Intelligence in Agriculture, Vol. 4, No., pp. 81-95.
Ppcexpo. 2022. The Power of Segmentation in Advertising [Online]. Available: https://ppcexpo.com/blog/segmentation-in-advertising [Accessed 2022].
Raj, A. 2020. A Practical Approach to Linear Regression in Machine Learning. A Hands-On Beginner's Guide to Linear Regression [Online]. Available: https://towardsdatascience.com/linear-regression-5100fe32993a [Accessed 2022].
Shah, N., Engineer, S., Bhagat, N., Chauhan, H. and Shah, M. (2020), "Research trends on the usage of machine learning and artificial intelligence in advertising", Augmented Human Research, Vol. 5, No., pp. 1-15.