Business & Social Science
IJRBS

Research in Business & Social Science

IJRBS VOL 13 NO 1 (2024) ISSN: 2147-4478

Available online at www.ssbfnet.com
Journal homepage: https://www.ssbfnet.com/ojs/index.php/ijrbs

Improving the prediction of social media engagement in universities by utilizing feature selection in machine learning

Dino Keco [(a)] *Engin Obucic* [(b)] * *Mersid Poturak* [(c)]

[(a)]*Associate Professor, Faculty of Engineering, Natural and Medical Sciences, Department of Information Technology, International Burch University, Francuske revolucije bb., 71000 Sarajevo, Bosnia and Herzegovina*
[(b)]*Faculty of Economics and Social Sciences, Department of Management, International Burch University, Francuske revolucije bb., 71000 Sarajevo, Bosnia and Herzegovina*
[(c)]*Associate Professor, Rector, Faculty of Economics and Social Sciences, Department of Management, International Burch University, Francuske revolucije bb., 71000 Sarajevo, Bosnia and Herzegovina*

ARTICLE INFO

Article history:

Received 22 November 2023
Received in rev. form 20 Jan 2024
Accepted 29 January 2024

Keywords:

Machine learning, social media, Facebook, feature selection, user engagement

JEL Classification:
M3

ABSTRACT

This study aims to examine the importance of feature selection in machine learning, specifically in predicting user engagement with social media post photographs on university Facebook pages. The paper uses a thorough analysis to demonstrate the crucial significance of choosing suitable features and their corresponding algorithms. The research intends to demonstrate how this strategic approach affects the accuracy of prediction findings in social media interaction. The research presents a compelling case study involving 24 leading universities from Australia, the United Kingdom, and the United States. The results underscore the efficacy of the method, stressing that the meticulous selection of characteristics and the use of appropriate algorithms are crucial elements for attaining best results in social media forecasts. Implications: The study's results have important consequences, particularly within the changing environment of machine learning and its use in social media. Feature selection and algorithm choice are vital for optimizing social media initiatives for institutions.

Introduction

Machine learning, a subset of Artificial Intelligence, focuses on creating self-learning skills in computers. It holds a pivotal role at the convergence of computer science and statistics, with strong connections to artificial intelligence and data science (Jordan & Mitchell, 2015). Advances in machine learning have been propelled by the creation of new learning algorithms and theories, together with the growing availability of online data and the cost-effectiveness of processing. Data-intensive machine learning techniques are widely used in several areas such as research, technology, and commerce, resulting in practical applications in marketing, education, financial modeling, and related professions (Jordan & Mitchell, 2015).

By leveraging several pertinent features, machine learning algorithms can create models that demonstrate exceptional performance (Channabasava & Raghavendra, 2022). Feature selection is a critical step in the machine learning process that can greatly influence the model's performance. Choosing and preparing features are essential for making sure they are relevant and appropriate for the particular task (Zheng & Casari, 2018). Machine learning models are known to perform best when working with numerical data due to their natural ability to handle and analyze it efficiently (Géron, 2019). However, quantitative, binary, categorical continuous, ordinal, and other sorts of features might also be important in certain situations. This is especially accurate when the values of a characteristic have intrinsic meaning that cannot be easily represented as numerical numbers (Géron, 2019). Machine learning models often utilize a combination of features to enhance performance by capturing the intricacies of the data. It is crucial to carefully choose and preprocess the characteristics to guarantee their suitability for the selected model (El Bouchefry & de Souza, 2020).

* *Corresponding author. ORCID ID: 0009-0001-9574-3418*

https://doi.org/10.20525/ijrbs.v13i1.3132

This study aims to examine the correlation between visual characteristics in Facebook posts and the associated machine learning techniques used to forecast user interaction. This study aims to find the picture elements that regularly lead to better user interaction on university Facebook image posts by analyzing data from 24 prestigious universities in Australia, the United Kingdom, and the United States. The results offer useful information for enhancing institution social media initiatives.

The study will determine the most significant features for predicting social media user engagement. It will next assess relevant machine learning algorithms to enhance the model's effectiveness. This study will show how important feature selection is in improving the accuracy of machine learning models and will offer practical tips for practitioners to optimize their models.

Literature Review

Machine learning features are crucial components that enable machine learning models to accurately process and understand data patterns. They are used to represent the data in a more meaningful and task-relevant manner. Features offer a method for understanding and analysing properties of the studied data. Features are defined as follows: “Each individual, independent instance that provides the input to machine learning is characterised by its values on a fixed, predefined set of features or attributes” (Witten & Frank, 2002). In this study, features are used as representative of various facets of the Facebook post images. By analysing these characteristics, researchers can gain insight into the visual and contextual factors that may influence user engagement.

Feature selection and engineering are crucial tasks for data scientists when managing real-world challenges (Müller & Guido, 2016). The task involves the identification of the most suitable data format for the specific application. The efficacy of a model can be significantly influenced by the accurate representation of data, which can have a greater impact than the specific parameters chosen by the user. According to (Karagiannopoulos et al., 2004), feature selection refers to the systematic procedure of identifying and eliminating redundant features from a given training dataset. The process of reducing the dimensionality of the data has the potential to enhance the efficiency of regression algorithms (Latiffi et al., 2022). There are cases where the correlation coefficient can be improved, but in other situations, it yields a concise and easily understandable representation of the desired outcome. James et al. (2013) point out that features facilitate the quantitative representation of visual data. Researchers have the ability to employ statistical and machine learning methodologies to process and evaluate data. This involves assigning numerical values to different visual attributes. The utilisation of a quantitative representation allows for a systematic examination and enhances the ability to discern patterns, correlations, and links among the various elements and the level of user engagement with the image.

Matz et al. (2020) suggest that features play a crucial role in the process of hypothesis testing and the assessment of research enquiries. Researchers have the ability to test hypotheses concerning the association between certain qualities and user engagement with Facebook post images by carefully picking features based on existing knowledge and theoretical considerations. This process enables the establishment of empirical evidence and the derivation of conclusions through the examination of observed relationships. The inclusion of features plays a vital role in the advancement of prediction models, as highlighted by El Bouchefry and de Souza (2020). Researchers have the ability to develop algorithms that can predict the degree of user engagement for unfamiliar or unobserved Facebook post images by discerning relevant features. These models can be utilised to understand the elements that contribute to user engagement and formulate informed selections pertaining to social media marketing content creation (Farrel et al., 2022). The inclusion of features facilitates comparisons across various images, universities, and countries. Researchers can evaluate the generalisability of their findings across different contexts by carefully choosing a consistent set of widely applicable features. This facilitates the examination of wider patterns and trends in user engagement on social media sites. James et al. (2013), Murphy (2018) and Bishop (2006) suggest several types of features that are commonly used in machine learning, including:

Table 1. Feature type and description

Feature Type	Description
Numeric features	Numeric features represent numeric values, such as the number of items in a set
Binary features	Binary features have only two possible values, such as 0 and 1.
Categorical features	Categorical features represent categories or classes, such as different types of flowers.
Continuous features	Continuous features continuously change values, like temperature measurements.
Ordinal features	Ordinal features have a defined order, such as scale from 1 to 5.

Source*:* (James et al., 2013); (Murphy, 2018); (Bishop, 2006).

Furthermore, one may also derive more complicated features by using raw data. A machine learning model may extract features that represent the interaction between multiple features or features that capture the overall structure or patterns of the data. Another possibility is that the model may extract features that represent the overall structure of the data.

Feature selection and engineering involves the careful selection and design of features that are pertinent and significant to the current task, while also evaluating the tradeoffs between different kinds of features and their relevance (Zheng & Casari, 2018). Feature selection may also denote the procedure of ascertaining the significance of features for the particular job they are designed to accomplish. Prior to incorporating them into the model, it is essential to carefully choose the appropriate features and perform comprehensive preprocessing (Hastie et al., 2009).

This study used the elements from the Google Cloud Vision API, graphic design and photography literature, and the Gourmet Photography Dataset to assess 24 high ranking universities. The analysis of graphic design and photography literature reveals an interest in how image features specifically may promote user engagement (AL-Ayash et al., 2016).

To understand the impact of image features, Malamed (2011) instructs designers to initiate the processing of visual information by utilising grouping, proximity, shape, line orientation, size, and colour. Malamed recommends that designers use "primitive features" to make their designs distinctive and noticeable (e.g. colour, motion, orientation, size, depth, tilt, closure). The idea is that these basic characteristics enable readers to quickly distinguish between different parts of the message. Malamed suggests that designers should create images that emotionally resonate with the life experiences of the viewer. When individuals can relate to an image, the message becomes more personal and meaningful. The use of features such as "special moment," "touch," and "looks/feels great" can enhance the emotional experience of the visual (Malamed, 2011)**.**

Through the analysis of graphic design and photography literature, one can acquire a greater understanding of the image features that contribute to their success. For example, photography literature suggests that there are numerous ways to improve the quality and aesthetic appeal of images. These characteristics include composition which refers to the arrangement of visual elements within the frame of an image. An image's composition can contribute to its sense of balance, harmony, and aesthetic appeal (Freeman, 2014). The manner in which light falls on a subject can have a substantial effect on the final image. Using natural light, artificial light, or a combination of the two can help create the ideal ambience or mood in an image (Hunter et al., 2015). Colour can be used to create a certain mood or ambience in an image. Choosing the appropriate colour scheme and employing it successfully can be crucial to the creation of aesthetically appealing photographs (Hunter et al., 2015). Moreover, orientation (portrait, landscape), indoor and outdoor shots play an important role in how images appear on different devices and therefore how images are consumed when displayed on a device.

The proper way of including the features in the dataset is suggested by the Gourment Photography Dataset (Sheng et al., 2018). GPD is a large-scale dataset used to assess the aesthetic quality of food photographs, consisting of 12,000 food images with human-annotated labels. In order to maintain the quality of the data, each worker is limited to annotating a maximum of 3,000 images. This also helps to prevent a small number of annotators from subjectively dominating the dataset's aesthetic perception. Only images that have been properly tagged or skipped have been included in the final dataset. In summary, features provide a structured and quantitative representation of the image data, enabling researchers to gain insights, draw conclusions, and make predictions regarding the user engagement of Facebook post images in leading universities.

Conversely, the selection of the right algorithm takes into account the size of the dataset, the intricate nature of the relationships among its elements, the possibility of interactions among those features, as well as the particular goals of the research. To discover which algorithm is the most suitable for a given dataset, it is best to try out a number of different algorithms, and assess how well they work by employing methods such as cross-validation (Caruana & Niculescu-Mizil, 2006). Nevertheless, the experimentation should be conducted within an acceptable theoretical framework, emphasising the need to match features with the corresponding algorithms that best interpret these features. Müller & Guido (2016) provide a comprehensive overview of machine learning features, as well as more complex features that can be derived from raw data:

Research and Methodology

Numeric (Quantitative) Features

Numeric features are used to measure quantities. Statistical analysis and software operations rely on numeric data. As such, numeric values serve as the most inherent method for computer software to interpret and evaluate input data. Numeric features are commonly expressed by various numerical combinations or in a binary format, typically denoted by the values 0 or 1. In addition, it is a common practice to use scaling or normalisation techniques to numeric features, especially when utilising algorithms that are sensitive to variations in feature scales, such as k-NN or SVM (James et al., 2013). Quantitative machine learning problems are best suited for numeric features. Based on the types of features described, the following machine learning algorithms are applicable for each category (Provost & Fawcett, 2013):

Numeric Features

i. Linear Regression: Uses linear relationships to predict numeric values (Ziegel et al., 1997).

ii. Decision Trees: Manages numeric features and recognises nonlinear relationships.

iii. Support Vector Machines/SMO: Manages quantitative data with distinct patterns and outliers.

Binary Features

i. Logistic Regression: used for binary classification tasks.

ii. Decision Trees: handles binary features and recognises complex relationships.

iii. Support Vector Machines/SMO: used for binary classification and high-dimensional data.

Neural Networks, Ensemble approaches such as Random Forest and Gradient Boosting are all capable of handling numeric features. Numeric features are well-suited for quantitative tasks. However, they can also be modified or encoded for classification and regression tasks (Bramer, 2020).

Non-Numeric (Qualitative) Features

Non-numeric features are features that are not expressed in numbers. They are descriptive of qualities used in the analysis. For example, unstructured text data in several languages represents non-numerical features (Felix Biessmann Amazon Research, Berlin, Germany et al., 2018). Brands, names, labels, or colours also represent qualities described by non-numeric features and are used for classification tasks. However, it is noteworthy that many machine learning algorithms, such as Naive Bayes and LSTMs, have the capability to process textual data directly, without the need for prior conversion into numerical representation (Schutze, 2008).

Categorical Features

Categorical features, also known as discrete features, are not numeric (Müller & Guido, 2016). Examples of categorical features may include qualities such as product brand, colour, or department.

i. Random Forest: handles categorical features with multiple classes.

ii. Naive Bayes: andles categorical data analysis and text classification.

iii. k-Nearest Neighbours (k-NN): handles classification tasks involving categorical features.

Continuous Features

Continuous features are measured along a continuous scale. For example, temperature measurements may vary significantly (from -30 to 40 Celsius) and thus fit the category of being a continuous feature.

i. Linear Regression, Decision Trees, and Random Forests handle continuous features well.

Ordinal Features

Ordinal features are expressed through ratings or scales. For example, if a movie is highly preferred by its audience, it will receive the highest rating such as a 5 out of 5-star rating.

i. Ordinal Regression: designed for ordinal data.

ii. Decision Trees: handles ordinal relationships.

iii. Random Forest with Ordinal Encoding: An Ensemble that can manage ordinal features.

Conversion

During the preprocessing stage, that is, before the data is integrated into machine learning models, the process of conversion takes place. This means that non-numeric features are converted into numeric form such as "encoding of categorical variables using the term frequency-inverse document frequency (TF-IDF) technique for text data" (Bramer, 2020). Conversely, numeric features can be transformed or encoded back to non-numeric form to be used in classification tasks as well as regression tasks.

In summary, it seems most profitable to match features with the corresponding machine learning algorithms by paying particular attention to the types of features included in the dataset. Referring back to the relevant literature on feature-algorithm selection, it is possible to quickly match the right algorithm with the corresponding feature.

Research and Methodology

Research Question

This study poses a specific research question in order to understand the relationship between features and machine learning algorithms. The purpose of the research question is to investigate the predictive capabilities of the Facebook image features identified through the literature review. This is specifically conducted for user engagement on the social media platform Facebook. User engagement on social media platforms is an essential indicator of audience interaction and interest. Universities can improve their social media strategies and engage their target audiences more effectively by identifying the image features that substantially contribute to user engagement.

To address the study's research objective, the following research question is posed:

RQ: Which machine learning algorithms are suitable in analysing Facebook post image features to predict user engagement, and how do these algorithms vary in terms of their performance and suitability across universities in Australia, the United Kingdom, and the United States?

This research question seeks to investigate the efficacy of various machine learning algorithms in addressing particular features of Facebook post images and the impact of these features on user engagement. The study also aims to understand the variations in algorithm efficiency among diverse cultural and institutional settings.

A Facebook-centric case study

This case study examines the utilisation of meticulously chosen features in machine learning prediction problems. The aim of this study is to ascertain the key variables that are most relevant in predicting user engagement on university Facebook post images. Additionally, the case study seeks to assess how the selection of these features affects the performance of machine learning models (Obucic et al., 2023). The case study involves the evaluation of different machine learning algorithms, including decision trees and SMO, in order to identify the most effective algorithm based on the chosen features. The purpose of the case study is to demonstrate the significance of Facebook post image feature selection in machine learning. The case study also aims to offer recommendations for enhancing the accuracy of machine learning models used to predict social media user engagement.

The dataset comprises 1,008 images that were manually gathered from 24 prestigious universities in the United States, the United Kingdom, and Australia (Obucic et al., 2023). The selection of these institutions was based on their reputation and the quality of their social media output. The data collection occurred over the period of June to July in 2022. Table 1 lists 24 university Facebook pages and 1,008 images referenced during the data collection process.

Table 2: University Facebook pages referenced in the study

US		UK		Australia	
Princeton	42	Oxford	42	Australian National University	42
Columbia	42	Cambridge	42	Sydney	42
Harvard	42	St Andrews	42	Melbourne	42
Yale	42	London School of Economics	42	New South Wales	42
Pennsylvania	42	Durham	42	Queensland	42
Dartmouth College	42	Warwick	42	Monash	42
Brown	42	Imperial College	42	Western Australia	42
Cornell	42	Bath	42	Adelaide	42

Total number of Facebook university pages: 24
Total number of Facebook images: 1,008

Source: (Obucic et al., 2023)

The selection of specific features of individual Facebook images was based on a comprehensive analysis of several sources, including the Google Cloud Vision API, literature on photography and graphic design, and the Gourmet Photography Dataset study (GPD). The Vision API offered by Google Cloud is a software solution designed to facilitate the seamless integration of image and video analysis functionalities within apps by developers. Vision API employs machine learning techniques to effectively categorise various entities such as objects, individuals, textual content, and other discernible components inside images and videos (Vision AI, 2023). In the present study, the utilisation of the Google Vision API was employed to extract specific features from the images within the dataset and subsequently classify the images. For instance, this technique was employed to classify various entities and spatial contexts within visual representations, encompassing outdoor environments and surrounding scenery. Consequently, it furnished contextual information that established the framework in which the visual material is positioned. The facial detection functionality was additionally employed to examine the emotional state of the individuals portrayed in the Facebook images, thereby offering valuable insights into the conveyed emotional tone. The features obtained from the Google API were subsequently incorporated into the feature set utilised for training by WEKA 3.8.5 machine learning classifiers.

The use of scholarly material pertaining to graphic design and photography indicates an interest in understanding the potential correlation between the technical elements of images and user engagement (Malamed, 2011). Viewer attention can be influenced by several visual characteristics of images, such as composition, lighting, colour, and others (AL-Ayash et al., 2016). Through an analysis of these components within the context of graphic design and photography literature (Bohn, 2006), a deeper understanding of the visual features that enhance user engagement can be attained (Li & Xie, 2020).

The Gourmet Photography Dataset (GPD) was utilised as a comparative benchmark in the present study (Sheng et al., 2018). The dataset utilised in the image analysis research comprises a diverse collection of internet images accompanied by comprehensive metadata. The dataset served as a framework for annotating a compilation of images from university Facebook posts. The purpose of employing annotation in the present study was to replicate the organisation and level of detail found in the GPD, thereby providing a validated framework for data processing. Table 3 below presents a list of 48 features included in the Facebook image analysis.

Table 3. 48 features utilised in the study

48 Features				
Country	Anger1	Roll1	Sorrow2	Headwear2
University	Surprise1	Tilt1	Anger2	Roll2
Image number	Exposed1	Pan1	Surprise2	Tilt2
Joy1	Blurred1	CONFIDENCE1	Exposed2	Pan2
Sorrow1	Headwear1	Joy2	Blurred2	CONFIDENCE2
Dominant Object	Logo (any)	Dominant Colour	R	G
B	Light	Composition	Special moment	Touch
Colour	Black and white	Portrait	Landscape	Indoor
Outdoor	Clear	Blurred3	Looks great	Good
Bad	Skip	Number of page followers		Output Class modelled around the percentage of likes. comments and shares (High, Medium, Low)

Source: (Obucic et al., 2023)

The data collection process focused on acquiring images associated with universities that were publicly shared on the Facebook platform. A total of 1,008 images were tagged by the assignment of numeric values to each individual image. The analysis of each image involved the examination of 48 preestablished features. The software tool WEKA 3.8.5 was utilised for the initial analysis and preprocessing of a total of 1,008 pictures. Preliminary examination of the dataset indicated the presence of outliers, which necessitated their removal. Upon eliminating the outliers, the dataset underwent a reduction, resulting in a total of 936 instances. The 936 instances were then divided into an Output Class with HIGH, MEDIUM, and LOW output values. The percentage of likes, comments, and shares compared to the university page's followers determined the Output Class's valuation.

Data transformation prepares data for analysis, interpretation, and decision-making. WEKA classifiers can handle missing data by using filters like "ReplaceMissingValues" and "NumericToBinary" to replace missing values. Balancing a training dataset is essential for imbalanced datasets (Bardenet et al., 17--19 Jun 2013). One of the ratios to reduce low-class samples while increasing medium- and high-class samples is applied. The ratio reflects a generalisation and may not apply to all cases:

OldLowClass * 0.5 = NewLowClass

NewMediumClass = 400/97 * Old MediumClass

For each class, NewHighClass = OldHighClass * (400/57) where OldLowClass, OldMediumClass, and OldHighClass represent the initial sample count.

This ratio results in an even distribution of 400 samples per class in the final dataset.

Findings and Discussions

The balanced dataset consists of a combination of numeric and binary features derived from Facebook post images. The objective is to predict a categorical label for each instance, belonging to one of three classes: LOW, MEDIUM, and HIGH. In general, it has been observed that a higher level of classification accuracy corresponds to a greater ability of the classifier to accurately assign new instances to their respective classes. The rationale behind using multiple classifiers was based on the observation that the single-classifier technique fails to produce the intended outcomes (ICB-InterConsult Bulgaria Ltd, 2019).

Table 4: The summary of results for three classifiers: SMO, RF and J48

Correctly Classified Instances						
Classifier	Data split 66%	Error	Data split 90%	Error	Cross-validation 10-fold	Error
SMO	94.85	5.15	96.67	3.33	94.85	5.15
RF	97.06	2.94	96.67	3.33	97.08	2.92
J48	88.48	11.52	94.17	5.83	90.92	9.08

Source: (Obucic et al., 2023)

The selection of Sequential Minimal Optimisation (SMO) and Random Forest (RF) is based on their capacity to effectively classify diverse problem sets, encompassing datasets containing both numeric and binary variables. Both SMO and RF algorithms have demonstrated effectiveness in handling numerical and binary information in the context of classification and regression tasks, while J48 showed a weaker ability to predict on the basis of the classified features.

On the initial SMO classifier data split, 66% of data was used for training and 34% for testing, and the SMO classifier achieved 94.85% classification accuracy. In the second split, the SMO classified data with 96.67% accuracy using 90% training data and 10% testing data, which suggests that the classifier works better on the training dataset and is somewhat weaker in the real world environment. The SMO classified data with 96.67% accuracy using 10-fold cross-validation.

RF demonstrated the highest performance in real world situations by achieving 97.08% accuracy rate using a 10-fold cross-validation technique. This means that the algorithm was trained and tested on different data subsets preventing the model from memorizing the data it has been trained on. As such, the model is designed for performance on unseen datasets.

In addition, J48 was chosen as an example of a classifier which demonstrates a somewhat weaker connection between the features and the algorithm. J48 was applied to the dataset employing information theory and a decision tree to solve classification problems (Shaari et al., 2022).

Discussion

The study explores the relationship between machine learning features and algorithm selection for predicting user engagement on university Facebook profiles. It focuses on connecting carefully selected features to the corresponding algorithms. It thus highlights the importance of feature types and the specificity of data in social media contexts (Latiffi et al., 2022). The study is significant for several reasons:

First, it underscores the importance of selecting appropriate features for machine learning analysis. These features are based on AI-driven systems such as the Google Cloud Vision API, offering millions of predefined feature options. They are also based on scholarly literature, allowing for the proper selection of features developed in the literature. The study clearly underscores the fact that the choice of features affects the performance of machine learning models.

Second, the study provides a context for optimising social media strategy employed by universities in three countries. As the importance of machine learning in university social media contexts and beyond continues to grow, the findings of this study serve as a resource for optimising university social media strategies.

Third, the study adds value to the social media strategy literature. In particular, the study explores the nuanced relationship between features and algorithms in an interdisciplinary context. It demonstrates how the proper selection of features and algorithms plays an important role in optimising social media strategy (Poturak et al., 2022). Specifically, it suggests optimisation strategies in the context of predicting user engagement on university Facebook profiles.

Fourth, the study points out that Random Forest (RF) algorithm is the best fit for the Facebook post image dataset considered in this study. This allows for an extended application of the algorithm to social media platforms such as Instagram and TikTok. In summary, the study points out the delicate relationship between machine learning features and algorithms in the context of predicting user engagement on university Facebook profiles. It suggests the key role of feature selection and algorithm matching in the development of machine learning models and their respective performance.

Implications

The research underscores the strategic nature of feature selection in machine learning. It emphasises that the efficacy of predictions is intricately tied to the judicious choice of features. Beyond feature selection, the study highlights the importance of pairing selected features with appropriate algorithms. The synergy between these elements is identified as the key to unlocking accurate social media engagement predictions. The inclusion of a case study involving prominent universities serves to validate the research findings in a real-world context. The results derived from this diverse set of institutions bolster the credibility of the study.

The findings of this study have significant implications for educational institutions working to improve their social media communications strategy. Through the use of individualised machine learning models, educational institutions are able to more effectively interact with their audience, which ultimately results in improved communication and outreach. Not only does this method assist universities in terms of user engagement, but it also provides insights into successful digital marketing strategies that may be used in the Higher Education Institutions (HEIs) setting.

Conclusions

This research highlights the importance of feature selection and algorithmic choice in improving the accuracy of predicting social media engagement for colleges. The results are a useful tool for organizations looking to enhance their social media strategies by using machine learning effectively.

The research offers significant insights, but future studies should further investigate specific features and algorithms, analyzing differences across various forms of social media content. Longitudinal research could enhance our understanding of the changing dynamics in predicting social media involvement.

Acknowledgement

All authors have read and agreed to the published version of the manuscript.
Author Contributions: Conceptualization, D.K, E.O. and M.P.; methodology, D.K, E.O. and M.P.; validation, D.K, E.O. and M.P.; formal analysis, D.K, E.O. and M.P.; investigation, D.K, E.O. and M.P.; resources, D.K, E.O. and M.P.; writing—D.K, E.O. and M.P.; writing—review and editing, D.K, E.O. and M.P.
Funding: This research was funded by the authors.
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.
Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.
Conflicts of Interest: The authors declare no conflict of interest.

References

Al-Ayash, A., Kane, R. T., Smith, D., & Green-Armytage, P. (2016). The influence of color on student emotion, heart rate, and performance in learning environments. *Color Research and Application, 41*(2), 196–205.

Bardenet, R., Brendel, M., Kégl, B., & Sebag, M. (2013). Collaborative hyperparameter tuning. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning (Vol. 28, pp. 199–207)*. PMLR.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bohn, K. (2006). *50 Principles of Composition in Photography: A Practical Guide to Seeing Photographically Through the Eyes of a Master Photographer*. CCB Publishing.

Bramer, M. (2020). *Principles of Data Mining*. Springer London.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning, 161–168*.

Channabasava, U., & Raghavendra, B. K. (2022). Ensemble assisted multi-feature learnt social media link prediction model using machine learning techniques. *Revue D Intelligence Artificielle, 36*(3), 439–444.

El Bouchefry, K., & de Souza, R. S. (2020). Chapter 12 - Learning in Big Data: Introduction to Machine Learning. In P. Škoda & F. Adam (Eds.), *Knowledge Discovery in Big Data from Astronomy and Earth Observation (pp. 225–249)*. Elsevier.

Felix Biessmann Amazon Research, Berlin, Germany, David Salinas Amazon Research, Berlin, Germany, Sebastian Schelter Amazon Research, Berlin, Germany, & Philipp Schmidt Amazon Research, Berlin, Germany, & Dustin Lange Amazon Research, Berlin, Germany. (2018, October 17). “Deep” Learning for Missing Value Imputationin Tables with Non-Numerical Data. *https://doi.org/10.1145/3269206.3272005*

Freeman, M. (2014). *The Photographer’s Eye: Graphic Guide: Composition and Design for Better Digital Photos*. CRC Press.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly Media, Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

Hunter, F., Biver, S., & Fuqua, P. (2015). *Light Science & Magic: An Introduction to Photographic Lighting*. CRC Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2004). Feature selection for regression problems. *Educational Software Development Laboratory, Department of Mathematics, University of Patras, Greece*. Link

Latiffi, M. I. A., Yaakub, M. R., & Ahmad, I. S. (2022). Flower pollination algorithm for feature selection in tweets sentiment analysis. *International Journal of Advanced Computer Science and Applications: IJACSA, 13*(5). https://doi.org/10.14569/ijacsa.2022.0130551

Li, Y., & Xie, Y. (2020). Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *JMR, Journal of Marketing Research, 57*(1), 1–19.

Malamed, C. (2011). *Visual Language for Designers: Principles for Creating Graphics that People Understand*. Rockport Publishers.

Murphy, K. P. (2018). *Machine learning: A probabilistic perspective (adaptive computation and machine learning series)*. The MIT Press: London, UK. https://www.academia.edu/download/62984186/Machine-Learning-A-Probabilistic-Perspective-Adaptive-Computation-And-Machine-Learning-Series-by20200416-47298-618w08.pdf

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Inc.

Obucic, E., Poturak, M., & Keco, D. (2023). Predicting User Engagement of Facebook Post Images in Leading Universities: A Machine Learning Approach. *Revue d’Intelligence Artificielle, 37*(4), 1039–1045.

Poturak, M., Keco, D., & Tutnic, E. (2022). Influence of search engine optimization (SEO) on business performance. *International Journal of Research in Business and Social Science (2147-4478), 11*(4), 59–68.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc.

Schutze, H. (2008). IIR 19: Web Search. https://www.cis.uni-muenchen.de/~hs/teach/14s/ir/pdf/19web.pdf

Sheng, K., Dong, W., Huang, H., Ma, C., & Hu, B.-G. (2018). Gourmet photography dataset for aesthetic assessment of food images. *SIGGRAPH Asia 2018 Technical Briefs*, Article 20.

Witten, I. H., & Frank, E. (2002). *Data mining: practical machine learning tools and techniques with Java implementations*. *SIGMOD Rec., 31*(1), 76–77.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly Media, Inc.

Ziegel, E. R., Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1997). *Applied linear statistical models*. Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences, *39*(3), 342.