Classifying Professional Photographers on Instagram: Data Collection and Processing for Computational Learning

Daniel Sánchez-Rodríguez
University of Murcia
Murcia, Spain
daniel.sanchezr@um.es

Sofia Strukova
University of Murcia
Murcia, Spain
strukovas@um.es

José A. Ruipérez-Valiente
University of Murcia
Murcia, Spain
jruiperez@um.es

ABSTRACT

Nowadays, the surge in open data on the internet allows researchers to investigate and broaden the understanding of numerous significant disciplines. However, there remains a notable deficiency in the advancement of methodologies for identifying artistic skills due to their subjectivity and the shortage of available datasets. Thus, our first contribution is a comprehensive, multimodal dataset that encompasses a wide array of attributes from 29 679 Instagram posts, originating from 1042 corresponding user profiles labelled as professional or not professional photographers. Employing this extensive dataset, we explored different machine learning (ML) models to assess their efficacy in classifying these profiles into their respective categories. The Random Forest (RF) model showed the best performance, being able to understand the common structure for professional photographers Instagram profiles. Further statistical analysis revealed significant distinctions between both types of profiles. The most important features for identifying a professional photographer are the number of users tagged, the technical score in their posts, and the height variance of the pictures made. The results obtained in this work hold the potential to significantly inform future research and offer practical applications across multiple real-world scenarios.

CCS CONCEPTS

• **Applied computing** → **E-learning**; **Distance learning**; • **Information systems** → **Information retrieval**; **Web searching and information discovery**; • **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

Instagram, Photography Capabilities, User Expertise, Computational Social Science, Data-driven Evaluation, Data Mining.

ACM Reference Format:
Daniel Sánchez-Rodríguez, Sofia Strukova, and José A. Ruipérez-Valiente. 2024. Classifying Professional Photographers on Instagram: Data Collection and Processing for Computational Learning. In *Proceedings of ACM SAC Conference (SAC'24).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/xx.xxx/xxx_x

SAC'24, April 8 –April 12, 2024, Avila, Spain

ACM ISBN 979-8-4007-0243-3/24/04…$15.00
https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

Over the past decade, the volume of data online has surged, primarily due to advancements in technology, the popularity of social media and the need of information for research projects in almost any study field. Many researchers use these data to develop machine learning (ML) models specialised in particular domains that span from healthcare to finance and beyond. Some studies evidence that the growth of social media usage has been exponential in the last years due to the social interaction, information seeking, and entertainment they provide [12]. Particularly notable has been the rise of photo and video sharing platforms which have become central to daily life for over 60% of the global population [7]. Thus, they present a rich source for exploration and understanding, including the distinction between professional and amateur content creators.

While photo and video sharing platforms have a significant impact on our world and hold immense potential for research, there remains a gap in the availability of comprehensive datasets that contain multiple data types and represent a broad spectrum of platform users. Such datasets would significantly benefit domains such as social behaviour analysis, privacy or data protection. Within the realm of expertise finding, there is a visible lack of solutions that can effectively detect artistic skills and users who have them. We could not discover any expert finding predictor model available on the Internet that makes use of information related to cognitive or abstract capacities, like artistic, photographic or narrative abilities, except a previous study on Flickr platform [16]. This absence can be explained by the fact that there are no datasets with such information and that those capacities are hard to measure given their complex nature and high subjectivity.

The motivation for our study stems from recent advancements made in evaluating specific capacities. Image Quality Assessment (IQA) techniques based on Convolutional Neural Networks (CNNs) have shown impressive results in establishing punctuations to different intrinsic features of pictures, such as sharpness, noise, or overall quality. For instance, the CNN-based model, DeepQA, has demonstrated a high alignment with human perceptual assessments, surpassing traditional techniques [8]. Therefore, we saw an opportunity for the development of a comprehensive dataset that can be used to detect professional photographers. This dataset is based on profiles of a photography-oriented social network that integrates IQA-computed features of posts alongside various social indicators from both the posts and the associated user profiles.

After a thorough review of prominent photo and video sharing platforms, we elected to focus on Instagram due to its pervasive influence. Our first step was to create a multimodal dataset encompassing diverse attributes from 1042 corresponding Instagram user profiles, alongside data from 29 679 postsWe enriched photography,

user-author and crowdsourced features by computing attributes such as technical and aesthetic image scores, as well as NLP features for comments and captions. Furthermore, we labelled our data to indicate for every profile whether it belongs to a professional photographer. With this dataset, we follow a process to find the optimal ML model for predicting if an Instagram profile belongs to a professional photographer or not. It will allow us to answer the following Research Questions (RQs):

- **RQ1**. How do ML algorithms perform at predicting professional photographer profiles on Instagram?
- **RQ2**. Which features contribute the most to the prediction of professional profiles?
- **RQ3**. What differentiates professional from non-professional photographers?

The remainder of this paper is structured as follows. In Section 2, we focus on the background of our study uncovering the subject of photo and video sharing platforms and related works. In Section 3, we present our research methodology. Our findings are outlined in Section 4, while we extend the results in Section 5. Finally, we draw our conclusions and future research directions in Section 6.

2 BACKGROUND

2.1 Photo & Video Sharing Platforms

Photo and video sharing platforms facilitate user interaction with content, incorporating a significant social component through features such as sharing, commenting, and liking. We analysed the most significant ones relevant to our study. Table 1 summarises the most interesting characteristics of several leading photo-based social media platforms, including 500px[1], Instagram[2], Pinterest[3], Flickr[4], EyeEm[5].

Launched in 2004, Flickr provides users with a rich and diverse community of about 60 million monthly active users where they can showcase their creativity and discover inspiring imagesIt offers a user-friendly interface and a wide range of features, facilitating the upload and organisation of photos.

Innovative and influential photography platforms EyeEM and 500px cater specifically to professional photographers, providing them with the opportunity to monetize their work and create portfolios. Both of the platforms offer paid versions that remove the upload restrictions. In contrast, Pinterest, launched in 2010, is primarily focused on facilitating the search of all types of images and curation of personal collections. It has one of the biggest communities, with more than 460 million monthly active users.

Finally, Instagram has quickly grown into a global phenomenon, boasting over two billion active users. The platform enables individuals to share their everyday experiences. Its user-friendly editing tools and a wide range of filters allow users to enhance their photos with ease. Over time, it has also become a relevant channel for professional users and a valuable data source for researchers [5].

[1]https://500px.com/
[2]https://www.instagram.com/
[3]https://www.pinterest.com/
[4]https://www.flickr.com/
[5]https://www.eyeem.com/

2.2 Related work

There exist many studies that demonstrate the value of social media information and how it can be useful in many domains. *Lekkas et al.* recollected Instagram data from individuals who reported having suicidal thoughts in the past, subsequently identifying similar users [9]. Likewise, *Zohourian et al.* effectively predicted the popularity of future Instagram posts [21]. Both studies, like ours, make use of Instagram profile data for the purpose of classifying them.

Social media data facilitate the identification of professional users. Numerous studies reached a good prototype for finding experts in forums or question-and-answer websites, such as Quora (Patil and Lee [14]), or social media platforms such as LinkedIn ([4]).

Nevertheless, this field has not made significant progress in assessing artistic and creative skills. Such expertise is crucial in many domains for enterprises and individual entrepreneurs striving for innovation and maintaining competitiveness [20]. We could not find any study focused on the identification of professional photographers excluding a previous work focused on Flickr [16]. It explored ML models which are able to infer if a user is a professional photographer or not based on self-reported occupation labels.

Finally, it is important to remark that the majority of the studies are making use of single-mode data sources. For example, *Pagolu et al.* applied sentiment analysis and supervised ML principles to tweets for predicting stock market movements [13]. In fact, few studies employed multimodal data, like *Gil-Ramírez et al.*, which analysed YouTube videos from elections to determine if the growth of social participation in political discourse through the new platforms revitalised or degraded the democratic deliberation Gil-Ramírez et al. [3]. Beyond the previously mentioned project, we found scarcely any studies that utilize multimodal data in the context of photo-sharing platforms.

3 METHODOLOGY

Figure 1 shows the methodology of the study. In this section, we present all the comprehensive steps undertaken to address the RQs. First, we describe the platform selected for our study, followed by the infrastructure developed for data acquisition in Instagram. Then, we explain the data collection step and the feature engineering techniques employed to create a solid dataset. Next, we provide a description of the final data collection. We conclude with a search for the best ML model for identifying professional photographers.

3.1 Photo Sharing Platform election

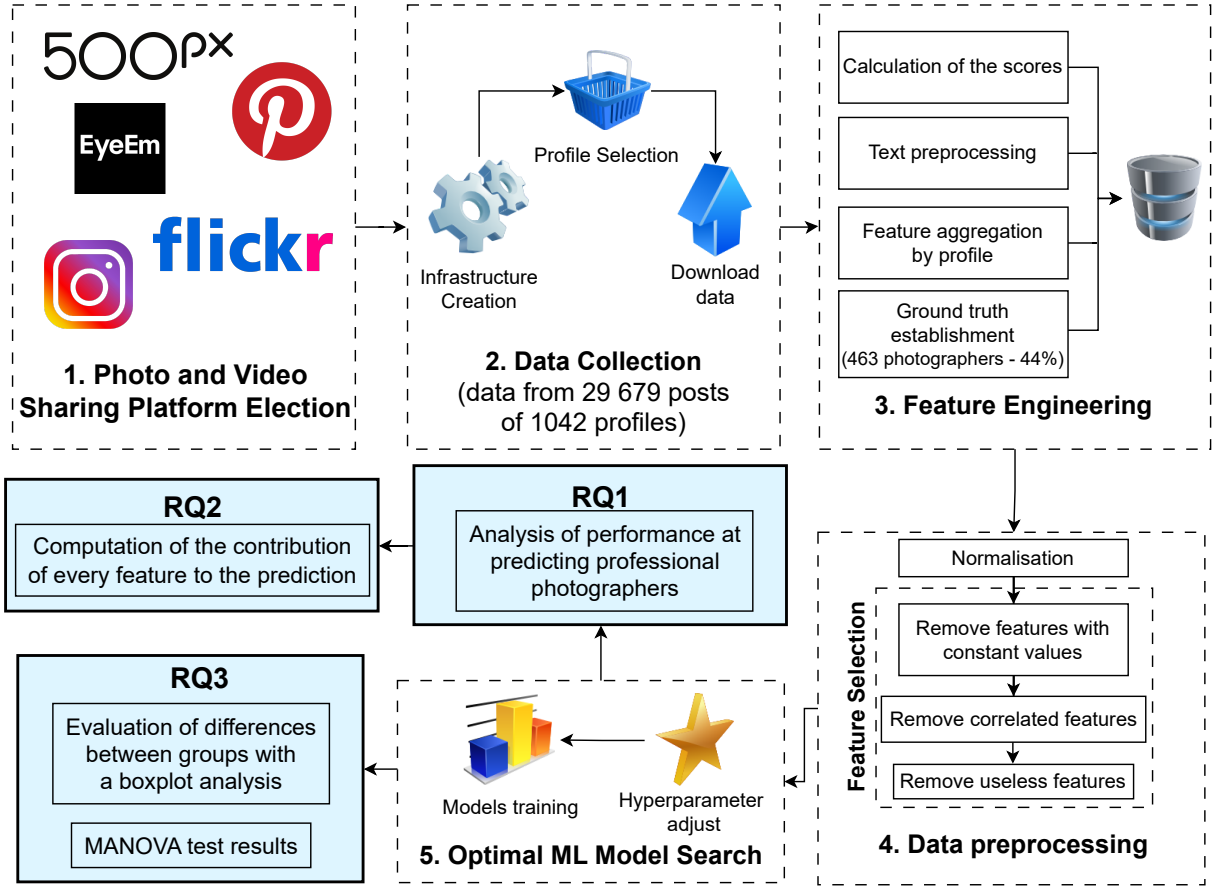
In this study, we are going to use the data of a photo sharing platform for detecting professional photographers' profiles. We took into consideration all the characteristics shown in Table 1 for choosing the social media site. We prioritised the monthly users because we wanted to extract a global and diverse dataset, with information about profiles of all ages, nationalities and occupations. Also, we wanted a photo sharing platform without free-version limits so that it does not influence the behaviour of users.

Even though all platforms were interesting options, we decided to choose Instagram. It has more than two billion users, being the most used photo sharing platform by far. It is available in 234 countries and has 32 different languages to choose from. Also, Instagram does not have any premium version and contains different tools

Table 1: Photo & Video Sharing Platforms Comparison

Portal	Foundation year	Monthly Users	Registration to browse/contribute	Free-version limits	Photo-editing features	Like/rating functionality	API
Flickr	2004	60 millions	✗/ ✓	1000 posts	✗	✓	✓
EyeEm	2011	N/A	✗/ ✓	20 uploads per week	✗	✓	✓
500px	2009	N/A	✗/ ✓	7 uploads per week	✗	✓	✗
Pinterest	2010	463 millions	✗/ ✓	✗	✓	✓	✓
Instagram	2010	2.3 billions	Limited / ✓	✗	✓	✓	Only for verified companies

Figure 1: Overview of the methodology to identify professional photographers in photo sharing platforms



for editing photos. Furthermore, it is not focused only on professional photographers so we could get a balanced dataset between professional photographers and normal users.

3.2 Infrastructure for recollecting data

The first step in order to answer the RQs was to recollect data. Instagram's API is restricted only to verified companies. Therefore, we used a Python module Instaloader [1] that allows downloading data through a set of diverse functions, which internally make online requests to the Instagram server. Making use of Instaloader's functionalities, we created a robust and automated infrastructure for downloading all the data provided by Instagram profiles and their posts. We divided the infrastructure into two different modules. The first one took care of the user recollection, while the second module was used for extracting the most important information about the users and their 30 most recent non-video posts.

The restrictions and request limits of Instagram's server forced us to implement various improvements in order to respect the

restrictions and not slow down the process. We implemented error-handling processes and logs to prevent the infrastructure from collapsing without the possibility of recovery. Furthermore, it was essential to handle request limits by having threads sleep programmatically to avoid timeouts.

3.3 Data Collection

The most critical aspect of our data collection was to achieve a representative sample of Instagram users and ensure a balanced distribution between professional photographers and other profiles.

Having considered utilising Instagram's recommendation page or specific account followers as potential data sources, we finally decided to use hashtags. They assure a representative sample of Instagram users because they are globally used and we could search the most recent posts for collecting only active users.

After an exhaustive analysis of Instagram's photography hashtags, we discovered that those associated with famous camera brands were predominantly used by professional photographers. For each selected account, we used these hashtags to extract their 30 most recent non-video posts and collect data from the respective account owners, assuring a balanced set of profiles.

3.4 Feature engineering for the ML model

3.4.1 Deep learning models. Photography has many intrinsic properties that define the quality of an image, such as composition, shape, technique, blur, lightness and noise which are hard to quantify due to their subjective and abstract components. In this regard, IQA techniques have obtained great results in the last years, especially the supervised learning approaches based on CNNs [6]. Thus, we considered using one of those approaches for evaluating the posts of every user.

In 2018, Google introduced NIMA [17] which open-source implementation utilised the CNN MobileNet, its weights initialised through training on the AVA[11] and TID [15] databases. The AVA database contains over 250 000 images with a great number of aesthetic ratings from professionals. On the other hand, TID includes 3,000 test images also with professional ratings, derived from 25 reference images and subjected to varying distortions. Unlike traditional models that categorise images into simple low/high quality bins, NIMA offers a nuanced rating, predicting on a scale from one to ten both the technical quality and aesthetic appeal of an image.

NIMA's reliability and effectiveness have been confirmed through multiple studies. For instance, recent progress in image super-resolution, which aims to derive high-resolution images from their low-resolution counterparts, has incorporated NIMA to gauge perceptual image quality, reducing the need for manual oversight [18].

In our research, we used NIMA to create two distinct features for each user's downloaded post: the technical score and the aesthetic score, both rated on a scale from one to ten. For sidecar posts, which can contain between 2 and 10 individual photos or videos, we opted to apply NIMA exclusively to the first photo.

3.4.2 Text preprocessing. Instagram allows users to accompany each post with a caption that can describe the post's content, convey personal sentiments, and include pertinent hashtags. Additionally, each post has a comment section. We collected all the comments and captions of each selected profile.

The amount of information within Instagram captions and comments can vary significantly based on the individual user and the underlying purpose of the post. The nature of comments and captions associated with a profile can be useful in identifying professional photographers. Thus, we employed various NLP techniques to generate features capturing the sentiment, complexity, and information density of the text. These features are crucial for differentiating professional photographer profiles. Prior to feature extraction, we converted all emojis into their respective word descriptions in natural language. Then, for the application of NLP techniques, we translated all comments into English. Once we standardised all the texts, we computed the subsequent features:

- Subjectivity – Degree in which the text reflects the author's personal opinion or perspective rather than objective facts. It is calculated with TextBlob [10] which uses a supervised ML approach and a combination of linguistic rules.
- Polarity – Emotional tone conveyed in a text (positive, neutral or negative). It is calculated with NLTK [2] using a vocabulary previously built to assign polarity scores to each word of the text and then calculating the composite scores.
- Difficult words – Number of difficult words in a text. It is calculated with TextStat [19] which compares every word with a list of words considered difficult.
- Reading Time – Average time needed to read a text. TextStat uses the average time to read a character and multiplies it by the number of characters in the text.
- Entropy – Amount of information contained in a text. This feature is calculated in terms of the variability and complexity of the letters used. We created an algorithm that calculates it with the help of Shannon entropy's formula:

$$-\sum_{i=1}^{n} p_i * log_2(p_i) \tag{1}$$

  - p_i =Letter i text frequency.
  - n = Number of words in the English alphabet.

3.4.3 Description of the final data collection. Our final dataset consists of 29 679 posts of 1042 profiles. The features we obtained were divided into three categories: user-author (Table 2), photography (Table 3) and crowdsourced (Table 4).

Table 2: User-author features

Name	Description	Domain
`Followers`	Profiles that follow the account	$\mathbb{N}$
`Followees`	Followed profiles	$\mathbb{N}$
`isBusiness`	It is a business account	Boolean
`isProfessional`	It is a professional account	Boolean
`hasLink`	It has a link in his biography	Boolean
`hideLikes`	It hides the number of likes and views	Boolean
`Category`	Specifies what the account relates to	String

Then, we computed additional features, including the proportion of images, videos, and sidecars attributed to each user, the technical and aesthetic NIMA scores, as well as the average NLP values for the captions and comments of each post.

Table 3: Photography features

Name	Description	Domain
Caption	Description of the post	String
PublicationDate	Date of publication	Unix Time
Height	Height of the image	$\mathbb{N}$
Width	Width of the image	$\mathbb{N}$

Table 4: Crowdsourced features

Name	Description	Domain
Likes	Accounts that liked the post	$\mathbb{N}$
Comments	Accounts that commented the post	$\mathbb{N}$
PhotosInSidecar	Photos/videos in the post	$\mathbb{N}$
TaggedUsers	Users tagged in a photo	$\mathbb{N}$
Location	The post gives a location	Boolean
isAccesible	The post has an accesibility text	Boolean

The next step was to transform all the crowdsourced and photo features (referred to a specific post) into user features that summarise them (referred to the owner of the different posts). As a result, for every user, there is a representation of every crowdsourced and photo feature. For NLP features, we incorporated both the mean and variance. For boolean attributes, we devised new features; these would be designated as 'True' if over 50% of the post values were 'True'. Conversely, if the majority were not, the value would be flagged as 'False'.

In its final form, our dataset is comprised of 42 distinct features.

3.4.4 Ground truth. We used the optional feature *Category* to establish the ground truth values. Instagram offers more than 1500 categories, allowing users to select the one most aligned with their content. There are several categories related to the audiovisual world, which may be adjusted to the profile of a professional photographer. Upon conducting an exhaustive examination of profiles within each category, we identified that the categories predominantly chosen by professional photographers include: Photographer, Camera/Photo, Photography Videography, and Visual Arts.

We tagged as professional photographers all the accounts who used one of the previously referred categories. As a result, 44.4% of the profiles were tagged as professional photographers, ensuring a balanced distribution between both types of profiles, achieved by primarily targeting users through camera brand hashtags.

3.5 ML model to identify professional photographers

The next step of our study was to train a ML model for detecting professional photographer profiles. To achieve this, we first undertook data preprocessing. Then, we selected appropriate supervised learning algorithms for the ML models. Lastly, we determined the hyperparameters that fit better our case study.

3.5.1 Data preprocessing. Data preprocessing involves transforming raw data into structured formats, enhancing the efficiency of ML models. We divided our dataset into a train set (70% of the profiles) and a test set (30% of the profiles). Then, we normalised all the values from each set and applied a feature selection:

- Features with constant values. We deleted all the features with more than 90% of their values repeated.
- Correlated features. We calculated the Spearman correlation coefficient for every pair of features and deleted one of each pair that had a correlation higher than 0.9 or lower than -0.9.
- Useless features. We trained Lasso's algorithm and determined the importance coefficient that it used for each feature, deleting the ones with value 0.

In this way, we reduced the number of features of the dataset from 42 to 30. The above-mentioned steps helped to reduce overfitting and achieve better and more interpretable results.

3.5.2 Supervised Learning Algorithms. For our study, we selected a range of algorithms: Logistic Regression (LR) as a probabilistic algorithm, Decision Trees (DT) as a non-probabilistic algorithm, and two ensemble algorithms based on DT, namely Random Forest (RF) and Gradient Boosting (GB). The combination of their strengths allows for addressing a broad spectrum of performance requirements. These algorithms collectively cover linear relationships, intricate interactions, non-linear patterns, and high-dimensional data, offering a versatile toolbox for achieving optimal results.

To optimise each algorithm's performance, it was crucial to select the best hyperparameters. We constructed four distinct hyperparameter grids, each tailored to the pertinent algorithm, encompassing commonly used values. We applied a 5-fold cross-validation on our training set and trained models for every hyperparameter combination. The best hyperparameter set for each algorithm was determined based on the AUC metric performance because it is one of the most reliable metrics for binary classification. Ultimately, we evaluated the top-performing models using our test set, reporting AUC, accuracy, precision, and recall.

4 RESULTS

4.1 RQ1 - Performance of predicting professional photographer profiles

We categorised all the profiles in the test set using the best ML model for each supervised learning algorithm (the one with the optimal hyperparameters). Table 5 shows the performance metrics for each predictive ML model.

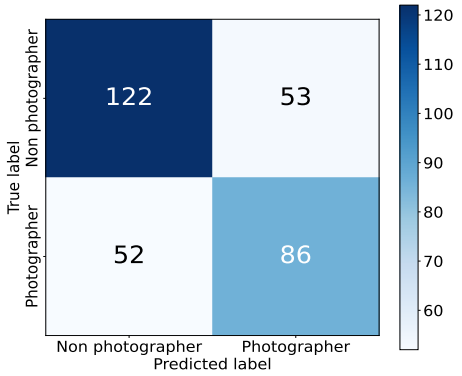
Table 5: Model prediction results

Algorithm	AUC	Accuracy	Precision	Recall
LR	0.675	0.604	0.733	0.159
DT	0.615	0.604	0.547	0.587
RF	0.691	0.665	0.619	0.621
GB	0.646	0.617	0.567	0.551

RF achieves the best AUC value (0.691), followed by LR (0.675), GB (0.646) and finally DT (0.615). Also, we can observe that the model with the best precision value is LR reaching the value of 0.733. However, its recall is 0.159, which means that it mostly predicts the profiles as non-professionals. While it accurately identifies most of

the non-professional profiles, it often misclassifies professional photographer profiles. RF ranks second in precision at 0.619, succeeded by GB at 0.567 and, lastly, DT at 0.547. For both accuracy and recall, RF also shows the best results and therefore we can conclude that RF is the best model for classifying professional photographers. Figure 2 presents the RF confusion matrix predicting 122 of the 175 non-photographers correctly, while the rest were misclassified. It also classified correctly 86 of the 138 professional photographers.

Figure 2: RF Confusion matrix



4.2 RQ2 - Contribution of the features in the final prediction

We used RF to answer RQ2 due to its best predictive power and optimal resources. Figure 3 depicts the importance of every feature in the RF classifier. Each of the values in the figure represents the average of the decrease in Gini impurity achieved by that feature across all trees in RF. All of the features' importances in the graphic are normalised so that they sum up to 1.

From Figure 3, we can observe that `avgTaggedUsers` is the most important feature in the classification task. This feature delineates the average number of profiles that a user tags in their posts. Likewise, `varHeight` is another crucial feature for classification, representing the variance in the height across all posts of a user.

The majority of the features had an importance between 0.05 and 0.03. We observe that both NIMA scores fall within this range, contributing to the model's performance. Most of the NLP computed features are also important, e.g., `cLenght`, `cEntropy` or `captionLength`.

4.3 RQ3 - Differences between professional and non-professional photographers

Figure 4 shows a boxplot distribution for the 15 most important features described in subsection 4.2. It represents a global overview for each metric and compares the distributions of values for professional photographers and non-professional profiles. We can easily detect differences in some metrics. MANOVA test confirms that there are remarkable differences between professional and non-professional photographer profiles ($F > 10^{16}, p < e^{-30}$). A more in-depth look into the influence of every metric using ANOVA tests gives us more detailed information. For example, `followers` metric is statistically different ($F = 6.26, p = 0.012$). The mean is 3113 vs. 1901 for professional and non-professional profiles. There is also a difference in the metric `followees`, with a mean of 941 vs 778 ($F = 4.65, p = 0.03$). Additionally, even though `captionLenght` is similar for both groups, the number of hashtags used in the captions (`captionAvgHashtags`) are statistically different ($F = 10.19, p = 0.001$), using the professional photographer accounts approximately four more hashtags in every caption. Another worth mentioning point is that the `avgLikes` for professional photographers is 198 vs. 103 for non-professional profiles.

5 DISCUSSION

5.1 Obtained results

As we can see in Table 5, ML models provide meaningful results in the classification of professional photographers. If we compare them to other expertise finding studies mentioned in subsection 2.2, our study displays worse performance. Even so, we should remark that those studies are focused on objective skills and technical fields which are easier to measure, while our study focuses on artistic skills like photography knowledge which is ambiguous because the interpretation of a picture could differ depending on the viewer. If we compare our results with the previously mentioned Flickr study [16], we can note that our AUC scores are similar emphasising the fact that it is possible to identify professional photographers based on multimodal data from photo and video sharing platforms. Also, some limitations or decisions explained in the next sections could have negatively affected the prediction performance. Considering this, the study identifies that there is a common structure for professional photographers' Instagram profiles that allows their identification and, like we determined while answering RQ1, RF shows a good potential to predict it.

Focusing on RQ2 regarding which features contribute the most to the prediction, each one of the three feature categories that we established is important. Looking at Figure 3, the importance of `avgTaggedUsers` above the others catches out, being able to conclude that professional photographers use the Instagram tag tool on posts more frequently than the rest of the users. Also, the importance of *varHeight* is caused because normal users usually take all their photos with the same phone while professional photographers may use different cameras to make their pictures. Instead, we can conclude that there is no significant difference between the use of links or accessible captions by both groups.

Finally, talking about RQ3, MANOVA results manifest that there is a statistical difference between both groups based on the features obtained. Figure 4 and ANOVA results provide us with more detailed information, revealing that `avgHeight`, `imagePorcentage`, `avgTechScore` or `captionAvgHashtags` are some of the features in which both groups differ the most. These findings are consistent with theoretical expectations. For example, professionals are likely to use a variety of formats to best suit the subject matter, hence the height variance could be a marker of this flexibility and expertise. Also, the difference in the technical score presumably reflects the quality and compositional elements of the photos, something that professionals are trained to optimise. Besides, the number of hashtags used by professional photographers being statistically higher makes sense. They are likely more attuned to the benefits of using hashtags for visibility and might use them strategically.

Figure 3: Feature importance plot based on RF model

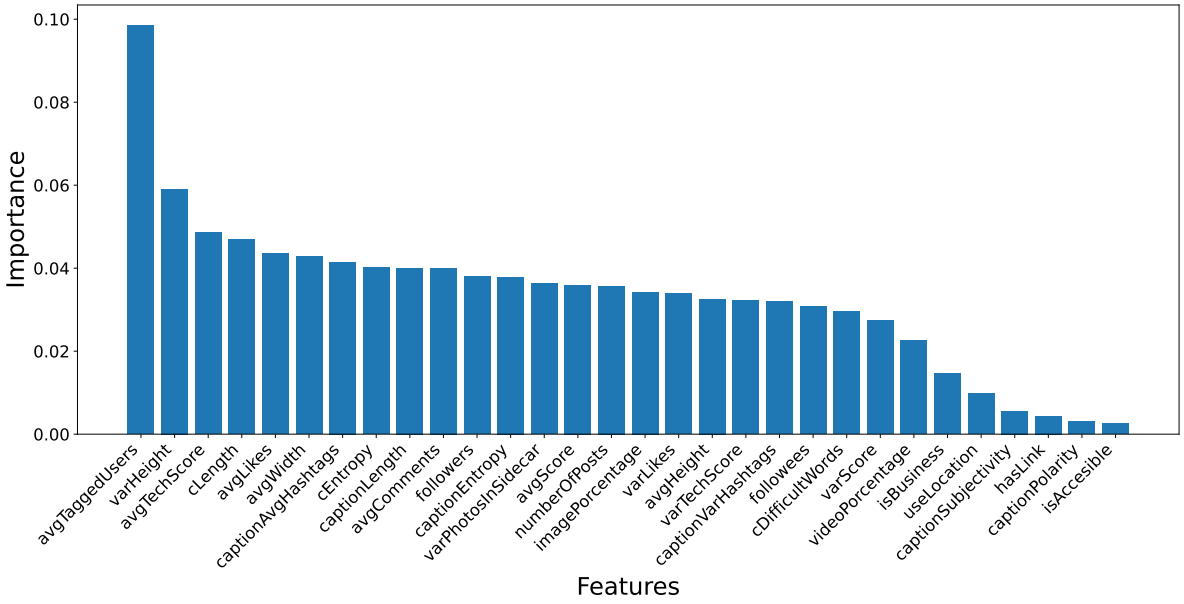
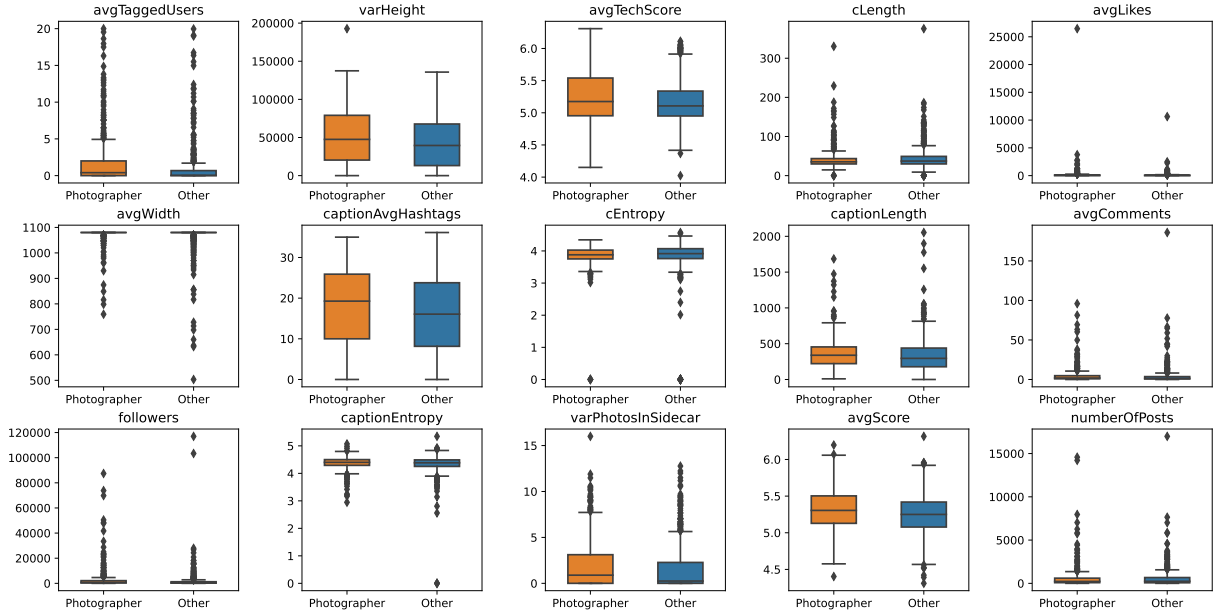


Figure 4: Differences between classes



5.2 Application in real scenarios

Understanding the distinction between amateur and professional photographers offers a myriad of practical applications. For example, we can gain insights into the particular skills, techniques, and styles that differentiate professionals. Such insights can be transformed into targeted training and development programs for amateurs with personalised feedback and resources. Also, digital platforms and photography tutorial websites can use this understanding to deliver content tailored to the skill level of their users.

Besides, websites that sell photographs can implement such models to categorise and rank photographers so that consumers could make informed decisions about purchasing photographs or hiring photographers. On the other hand, companies and agencies in

search of professional photographers can use similar ML models to shortlist potential candidates from platforms like Instagram, saving valuable time and resources. Furthermore, schools and colleges offering photography courses can benefit from this study to understand the current market standards, ensuring that their curriculum remains relevant and up-to-date. Lastly, brands can use this distinction to identify professional photographers for collaborations.

5.3 Limitations

Our study faced several limitations which we would like to discuss. First, we faced restrictions imposed by Instagram, limiting the depth of data collection. Secondly, our dataset might inadvertently favour certain photography styles over others due to inherent biases in Instagram's user base and the selected hashtags. This bias can skew our results, potentially overlooking diverse photographic styles that do not align with popular trends. Further research would benefit from a more varied data source and an inclusive selection approach.

5.4 Ground truth

When differentiating amateur and professional photographers, we grounded our truth on the category dimension, enabling users to pick the category that aligns most closely with their content including photography. We would like to outline several potential biases that could arise from our ground truth determination. Firstly, profiles might self-identify as "Photographer" or "Visual Arts" without necessarily having professional training or earning from photography. Secondly, new or emerging photographers might not yet have identified with the professional categories, even if their work meets professional standards. Thirdly, certain profiles, although categorised under photography category or similar, might be more aligned with videography or other visual arts. Finally, it is crucial to consider the evolution of a photographer's journey. Labelling based on their current category might not capture the spectrum of their skills or their transition phase. These issues could be mitigated by using other indicative elements, such as linkages to a commercial photography website or consistency in posting high-quality content. Moreover, an analysis of content quality and engagement rates, usage of historical data or the analysis of progression in content quality over time can provide a more dynamic understanding.

6 CONCLUSIONS

This work provided a multimodal dataset with information about 1024 profiles and 29 679 posts derived from Instagram. It serves as a foundational platform for the creation of algorithms aimed at discerning various abstract capacities. Utilising a RF classifier, the research successfully identifies a ubiquitous structural pattern among professional photographers' Instagram profiles. Notably, professional photographers exhibit a predilection for double the amount of user tags within their posts compared to average users. They also differ more in the height of the pictures posted, which exhibited better technical features as well. Both MANOVA and ANOVA tests confirm that there are remarkable differences between professional and non-professional photographer profiles.

Future works should consider a multifaceted approach by using both objective metrics and qualitative evaluations. A key step is to manually label a subset of photographer profiles to verify our ML classifications. This will not only provide a solid ground truth but also highlight any discrepancies between manual and computational labelling. Furthermore, exploring the deeper semantic analysis of captions and comments might offer further distinctions between professional and amateur photographers.

REFERENCES

[1] André Koch-Kramer Alexander Graf. 2016. Instaloader: Instagram scraper repository. https://github.com/althonos/InstaLooter

[2] Steven Bird. 2006. Natural Language Toolkit (NLTK). https://github.com/nltk/nltk

[3] Marta Gil-Ramírez, Ruth Gómez-de Travesedo-Rojas, and Ana Almansa-Martínez. 2020. Political debate on YouTube: revitalization or degradation of democratic deliberation? *Profesional de la información* 29, 6 (2020).

[4] Viet Ha-Thuc, Ganesh Venkataraman, Mario Rodriguez, Shakti Sinha, Senthil Sundaram, and Lin Guo. 2015. Personalized expertise search at LinkedIn. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 1238–1247.

[5] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 595–598. https://doi.org/10.1609/icwsm.v8i1.14578

[6] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1733–1740.

[7] Simon Kemp. 2023. *Digital 2023 april global statshot report.* https://datareportal.com/reports/digital-2023-april-global-statshot

[8] Jongyoo Kim and Sanghoon Lee. 2017. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1676–1684.

[9] Damien Lekkas, Robert J. Klein, and Nicholas C. Jacobson. 2021. Predicting acute suicidal ideation on Instagram using ensemble machine learning models. *Internet Interventions* 25 (2021), 100424. https://doi.org/10.1016/j.invent.2021.100424

[10] Steven Loria. 2013. TextBlob: Simplified Text Processing. https://github.com/sloria/textblob

[11] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2408–2415.

[12] Esteban Ortiz-Ospina. 2019. The rise of social media. *Our World in Data* (2019). https://ourworldindata.org/rise-of-social-media

[13] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*. IEEE, 1345–1350.

[14] Sumanth Patil and Kyumin Lee. 2015. Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors. *Social Network Analysis and Mining* 6 (12 2015). https://doi.org/10.1007/s13278-015-0313-x

[15] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30 (2015), 57–77. https://doi.org/10.1016/j.image.2014.10.009

[16] Sofia Strukova, Rubén Gaspar Marco, José A. Ruipérez-Valiente, and Felix Gomez Marmol. 2023. Identifying Professional Photographers Through Image Quality and Aesthetics in Flickr.

[17] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing* 27, 8 (2018), 3998–4011.

[18] Zhihao Wang, Jian Chen, and Steven CH Hoi. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3365–3387.

[19] Alex Ward. 2014. TextStat: NLP Python package. https://github.com/textstat/textstat

[20] Piotr Wesołowski. 2022. Enhancing architectural engineering students' acquisition of artistic technical competences and soft skills. *Cogent Arts & Humanities* 9, 1 (2022), 2043997. https://doi.org/10.1080/23311983.2022.2043997 arXiv:https://doi.org/10.1080/23311983.2022.2043997

[21] Alireza Zohourian, Hedieh Sajedi, and Arefeh Yavary. 2018. Popularity prediction of images and videos on Instagram. In *2018 4th International Conference on Web Research (ICWR)*. IEEE, 111–117.