

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

Sofia Strukova^{1*} , Daniel Sánchez-Rodríguez² , José A. Ruipérez-Valiente² 

¹ Institute for Implementation Science in Health Care, University of Zurich, Zurich (Switzerland)

² Department of Information and Communication Engineering, University of Murcia, Murcia (Spain)

* Corresponding author: strukovas@um.es

Received 28 January 2024 | Accepted 23 December 2025 | Early Access 9 March 2026



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, particularly in the field of expertise finding, due to their subjectivity and the shortage of available datasets. Thus, we saw an opportunity in the popularity of photo sharing platforms to create a dataset for the identification of professional photographers' profiles. 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.

KEYWORDS

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

DOI: 10.9781/ijimai.2026.2211

I. 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 for information for research projects in almost any study field [1]. 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 recent years [2] due to the social interaction, information seeking, and entertainment they provide [3]. Particularly notable has been the rise of photo and video sharing platforms because they satisfy five primary social and psychological components of the human being: social interaction, archiving, self-expression, escapism, and peeking [4]. Also, these platforms have become central to daily life for over 60% of the global population [5]. 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 platforms 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 find any expert-finding model available online that makes use of information related to cognitive or abstract capacities, like artistic, photographic or narrative abilities, except a previous study on the Flickr platform [6]. 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)

Please cite this article as:

S. Strukova, D. Sánchez-Rodríguez, J. A. Ruipérez-Valiente. Classifying Professional Photographers on Instagram: Data Collection and Processing for Computational Learning, International Journal of Interactive Multimedia and Artificial Intelligence, (2026), <http://doi.org/10.9781/ijimai.2026.2211>

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 [7]. Similarly, Brisque stands out in its ability to predict image quality without using reference images (Non-Reference IQA) [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 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 posts [9]. We 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 II, we focus on the background of our study uncovering the subject of photo and video sharing platforms and related works. In Section III, we present our research methodology. Next, we depict the final data collection and describe ML algorithms to identify professional photographers. Our findings are outlined in Section IV, while we extend the results in Section V. Finally, we draw our conclusions and future research directions in Section VI.

II. BACKGROUND

A. 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. Given the abundance of photo-sharing platforms available today, we analysed the most significant ones relevant to our study. We discarded platforms such as Snapchat and BeReal because they are centred around limited-time photo-sharing among friends and therefore are unsuitable for the needs of professional photographers. Table I summarises the most interesting characteristics of several leading photo-based social media platforms, including 500px¹, Instagram², Pinterest³, Flickr⁴, EyeEm⁵.

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, discover inspiring images, and connect with like-minded individuals from all over the world. It offers a user-friendly interface and a wide range of features, facilitating the upload and organisation of photos. With its vast collection of images covering various genres and subjects, the FlickrAPI and the amount of information available, Flickr has become an invaluable resource for researchers [10].

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 for all types of images and the curation of personal collections. It also has a vast quantity of images across all domains and styles, and is 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, transforming ordinary snapshots into visually appealing masterpieces. Over time, it has also become a relevant channel for professional users and a valuable data source for researchers [11].

B. 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.* collected Instagram data from individuals who reported having suicidal thoughts in the past, subsequently identifying similar users [12]. Likewise, *Zohourian et al.* effectively predicted the popularity of future Instagram posts [13]. 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 Reddit ([14]) or Quora ([15]). Similarly, *Ha-Thuc et al.* determined LinkedIn professionals based on their profiles and associated social metrics [16].

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 [17]. We could not find any study focused on the identification of professional photographers excluding a previous work focused on Flickr [6]. It explored ML models which are able to infer if a user is a professional photographer or not based on self-reported occupation labels.

It is also 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 [18]. Besides, *Idrees et al.* analysed the time series data of the Indian stock market and built a statistical model that could efficiently predict the future stocks [19]. In fact, few studies employed multimodal data, like *vand Dijk et al.*, who demonstrated the utility of using textual, behavioural and time-aware features in StackOverflow [20]. Also, *Gil-Ramírez et al.* analysed YouTube videos from the Andalusian elections to determine if the growth of social participation in political discourse through the new platforms revitalised or degraded the democratic deliberation [21]. Beyond the previously mentioned project, we found scarcely any studies that utilise multimodal data in the context of photo-sharing platforms.

While few studies have focused specifically on identifying artistic skill, there are intersecting research directions that offer useful

¹ <https://500px.com/>

² <https://www.instagram.com/>

³ <https://www.pinterest.com/>

⁴ <https://www.flickr.com/>

⁵ <https://www.eyeem.com/>

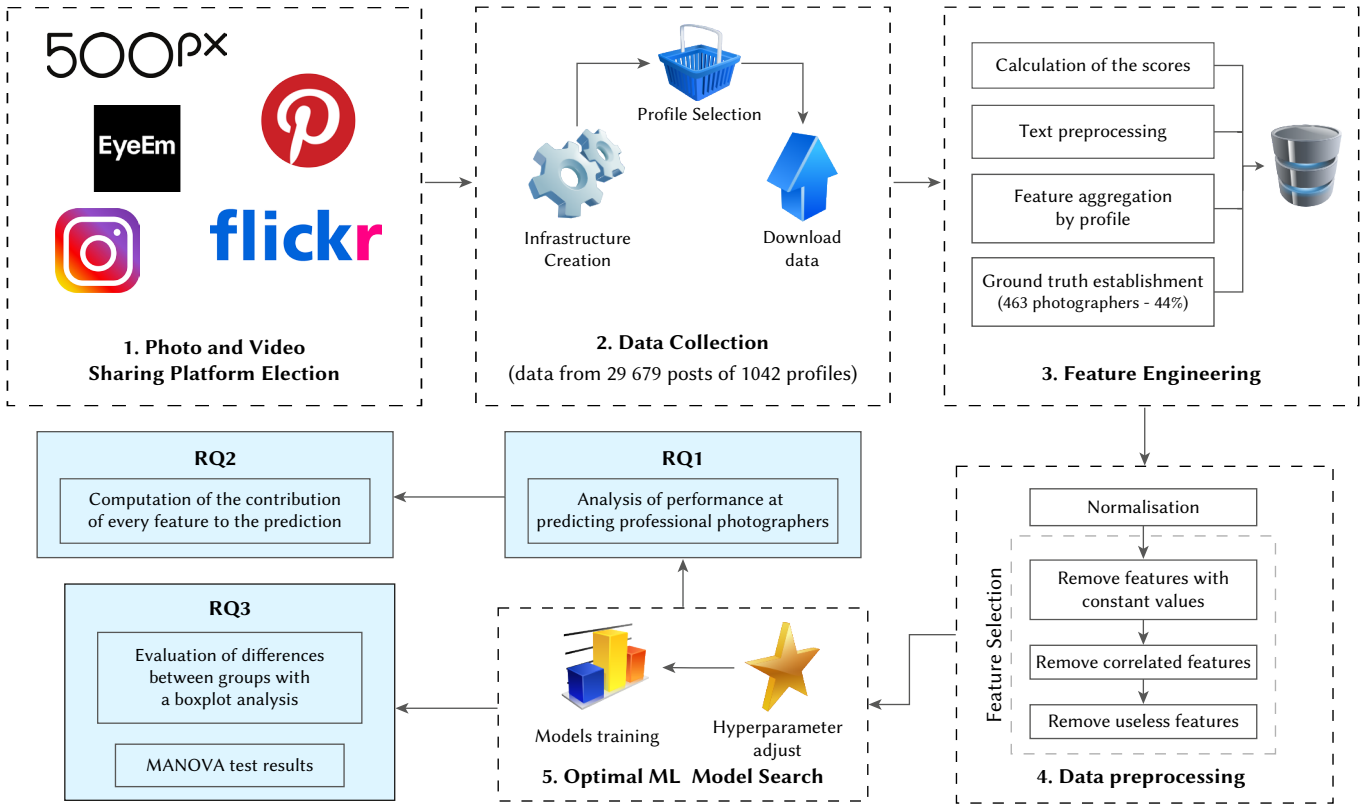


Fig. 1. Overview of the methodology to identify professional photographers in photo sharing platforms

TABLE I. 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

perspectives. One such direction is multimodal user modelling, where researchers combine visual, textual, and interaction-based signals to infer traits such as personality, influence, or interest profiles. These approaches have been applied to social media platforms to model user behaviour and identity based on diverse content modalities [22]. A second relevant area is automated competence and expertise evaluation in informal or creative domains. For example, recent work has investigated the use of large language models (LLMs) to assess student essays in educational settings, revealing both the promise and current limitations of automated skill evaluation when compared to human grading [23]. Such research underscores the broader challenge of modelling nuanced human competencies from text or media and highlights the need for domain-specific feature engineering. Our study contributes to this growing body of work by applying a data-driven, feature-rich approach to detecting professional-level photographic skill in social media contexts.

III. METHODOLOGY

Fig. 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.

A. Platform Selection

In this study, we use the data of a photo sharing platform for detecting professional photographers' profiles. We took into consideration all the characteristics shown in Table I 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 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.

B. Infrastructure for Collecting Data

The first step in order to answer the RQs was to collect data. Instagram's API is restricted only to verified companies. Therefore, we used a Python module *Instaloader* [24] that allows downloading data through a set of diverse functions, which internally make online requests to the Instagram server. This is a powerful, intuitive and flexible tool due to its configuration options that adjust its behaviour.

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 collection, 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 delay execution programmatically to avoid timeouts.

C. 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.

D. Feature Engineering for the ML Model

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 [25], especially the supervised learning approaches based on CNNs [26]. Thus, we considered using one of those approaches for evaluating the posts of every user.

In 2018, Google introduced NIMA [27] which open-source implementation utilised the CNN MobileNet, its weights are initialised through training on the AVA [28] and TID [29] 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. Then, the last layer of the network is substituted by a layer of 10 randomly initialized neurons fully connected followed by softmax activations. 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 [30].

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. These features represent the IQA metrics that will be further explored in response to RQ2 in subsequent sections. For sidecar posts, which can contain between 2 and 10 individual photos or videos, we opted to apply NIMA exclusively to the first photo.

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. However, to achieve this, the textual data must be transformed into quantifiable features. Thus, we employed various NLP techniques [31] 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 corresponding natural language descriptions. Then, for the application of NLP techniques, we translated all comments into English. Once we standardised all the texts, we used various Python libraries and our proper algorithms to compute 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* [32] 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* [33], [34] 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* [35] which compares every word with a list of words considered difficult.
- **Reading Time** – Average time needed to read a text. *TextStat* estimates reading time based on average character reading speed and the total number of characters.
- **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 (1) [36]

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

p_i = Letter i text frequency.

n = Number of words in the English alphabet.

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 II), photography (Table III) and crowdsourced (Table IV).

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 II. 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
hideLike	It hides the number of likes and views	Boolean
Category	Specifies what the account relates to	String

TABLE III. PHOTOGRAPHY FEATURES

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

TABLE IV. USER-AUTHOR 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.

4. Ground Truth

We used the optional feature Category to establish the ground truth values. This is a special attribute that professional Instagram profiles could contain. 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, many were excluded since they pertained to other professional types, such as camera shops, cinema directors and camera brands, among others. Through this rigorous selection process, we identified that the categories predominantly chosen by professional photographers include: Photographer, Camera/Photo, Photography Videography, and Visual Arts.

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

E. 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.

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.

2. Supervised Learning Algorithms

For our study, we selected a comprehensive suite of algorithms to build a robust classification model. The selection includes foundational models like Logistic Regression (LR) and Decision Trees (DT) as strong baselines. We heavily explored ensemble methods, including bagging with Random Forest (RF) and several powerful boosting techniques: standard Gradient Boosting (GB), XGBoost, LightGBM, and CatBoost. Furthermore, a Stacking Classifier was implemented to combine the predictive power of different base models. To capture potentially deep and complex patterns within the tabular data, we also incorporated TabNet, a deep learning-based algorithm. This diverse set of algorithms covers linear relationships, intricate non-linear patterns, and high-dimensional data, providing a versatile toolbox for achieving optimal results.

To optimise each algorithm’s performance, we conducted a hyperparameter tuning process using GridSearchCV. A tailored hyperparameter grid was constructed for each of the selected algorithms, encompassing a range of commonly used values. We applied a 5-fold cross-validation on our training set, training 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. Also, although it is not a problem for our dataset, AUC is more robust against class imbalance. Ultimately, the top-performing model for each algorithm was evaluated on the held-out test set, with performance reported using metrics such as accuracy, precision, recall, F1-score, and AUC.

IV. RESULTS

A. 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 V shows the performance metrics for each predictive ML model.

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. Fig. 2 presents the RF confusion matrix predicting 124 of the 174 non-photographers correctly, while the rest were misclassified. It also classified correctly 73 of the 139 professional 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. Fig. 2 presents the RF confusion matrix predicting 124 of the 174 non-photographers correctly, while the rest were misclassified. It also classified correctly 73 of the 139 professional photographers. The hyperparameters used in the model are as followed: `class_weight: balanced`, `criterion: gini`, `max_depth: 6`, `max_features: log2`, `min_samples_leaf: 0.005`, `min_samples_split: 0.005`, `n_estimators: 200`, `n_jobs: -1`. The chosen hyperparameters are designed to reduce overfitting by not allowing the decision trees to grow too complex. This is important in ensuring the model's generalisability to new, unseen data.

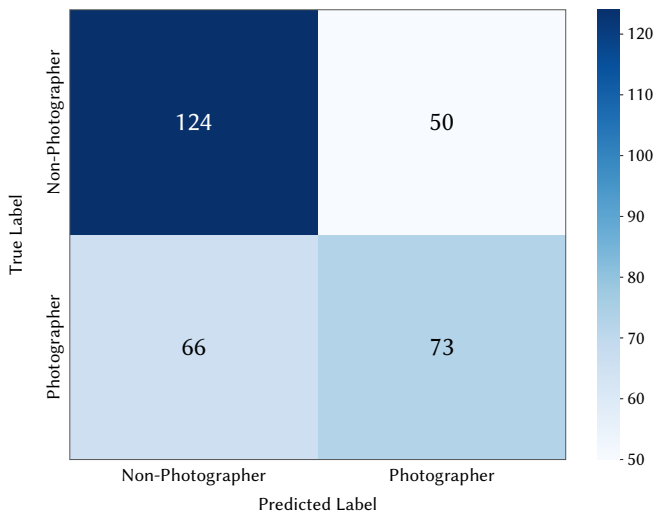


Fig. 2. RF Confusion matrix.

B. RQ2 -Contribution of the Features in the Final Prediction

We used RF to answer RQ2 due to its best predictive power and optimal resources. Fig. 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 Fig. 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.

Besides, `avgTechScore`, indicating the average technical quality of the photos, is a vital metric. High technical quality is often synonymous with professional work. Also, `avgLikes`, the average number of likes

per post, is also important, as it reflects audience engagement, which is typically higher for professional content.

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 `captionLenght`. We can also note that some features are practically useless, such as `isAccesible`, `captionPolarity` or `hasLink` indicating that they are not significant differentiators for professional photographers.

C. RQ3 -Differences Between Professional and Non-Professional Photographers

Fig. 4 shows a boxplot distribution for the 15 most important features described in subsection B. 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 differs statistically ($F = 6.26$, $p = 0.012$). 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`) differ statistically ($F = 10.19$, $p = 0.001$), 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. However, there are also some features, according to ANOVA results, that do not have a real impact on the classification, like `avgComments` ($F = 1.23$, $p = 0.27$), `varTechScore` ($F = 0.45$, $p = 0.5$), `captionPolarity` ($F = 2.03$, $p = 0.16$), etc.

V. DISCUSSION

In this section, we first discuss the obtained results of this work. Then, we focus on the benefits and applications of our study in real scenarios. We also talk about the limitations we faced, as well as our study's ground truth.

A. Obtained Results

As we can see in Table V, ML models provide meaningful results in the classification of professional photographers. Despite the rapid evolution of deep learning and multimodal architectures, in our experiments, classical ML models -particularly RF -consistently outperformed more recent neural approaches in terms of accuracy, interpretability, and robustness. This can be attributed to several factors. First, our dataset, while rich in features, contains only 1,042 labelled profiles -a size insufficient to fully exploit the representational power of deep neural networks without overfitting. Second, the dataset already includes well-crafted, domain-informed features that summarize complex signals (e.g., visual quality, linguistic complexity, tagging behaviour), reducing the need for end-to-end feature extraction. In this context, ensemble tree-based models like RF are well suited to capture non-linear interactions among heterogeneous variables, while remaining resilient to noise and irrelevant features. Lastly, classical models offer direct access to feature importance metrics, which aligned well with our interpretability goals and enabled the identification of latent competencies. Although recent multimodal transformers are promising, especially for raw image-text inputs, they require substantially larger training datasets or fine-tuning on task-specific benchmarks - conditions not met in our current study.

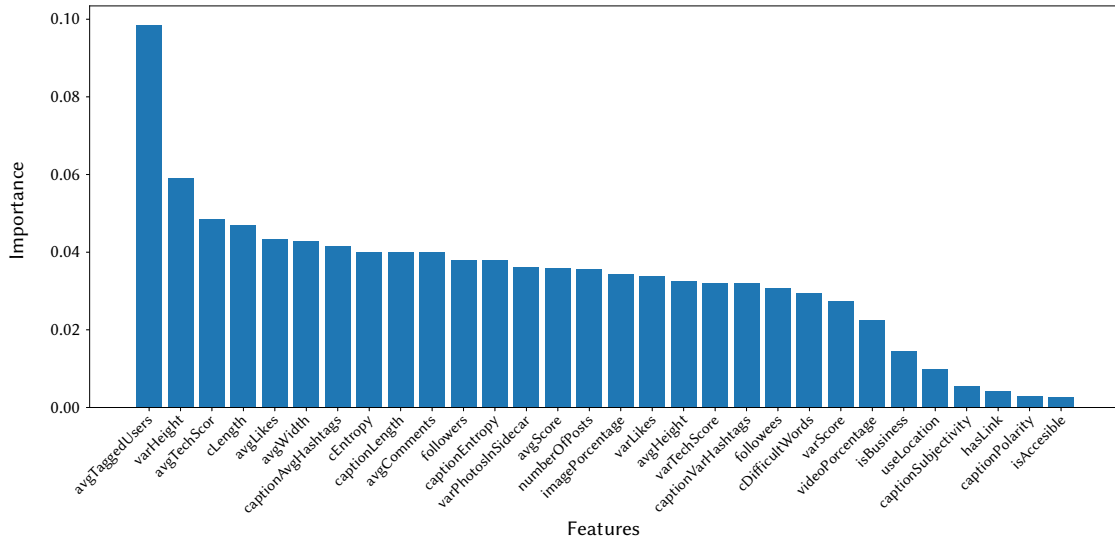


Fig. 3. Feature importance plot based on RF model.

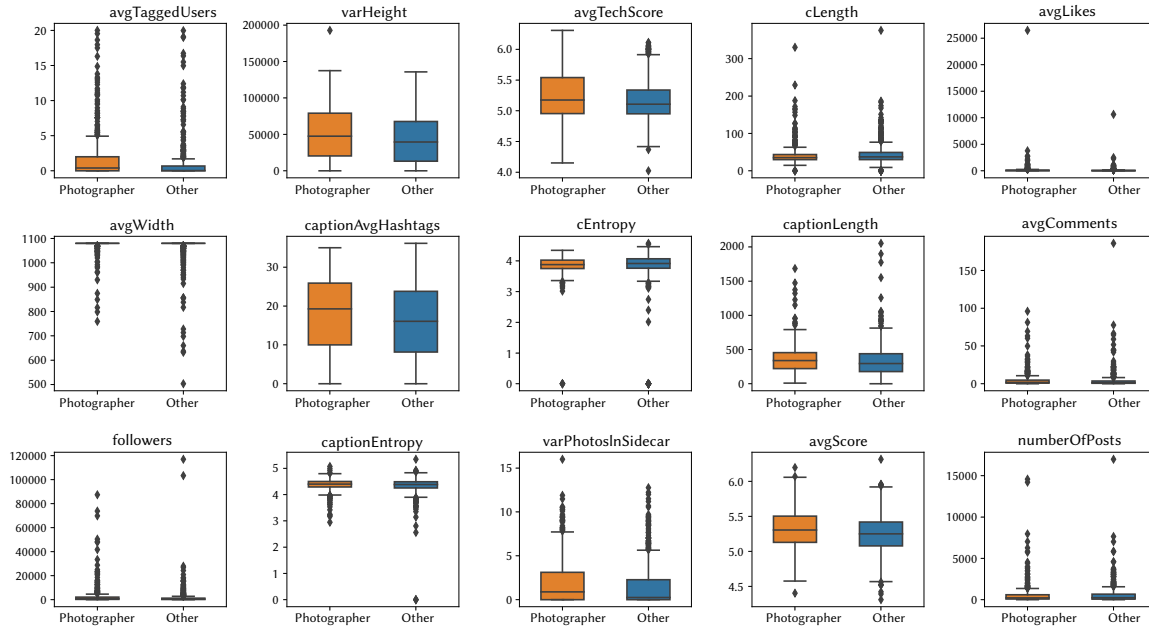


Fig. 4. Differences between classes.

TABLE V. MODEL PREDICTION RESULTS

Algorithm	AUC	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.654	0.629	0.594	0.525	0.557
Decision Trees	0.582	0.588	0.533	0.576	0.554
Random Forest	0.706	0.658	0.64	0.525	0.577
XGBoost	0.662	0.61	0.57	0.496	0.531
Gradient Boosting	0.66	0.613	0.576	0.489	0.529
CatBoost	0.702	0.665	0.655	0.518	0.578
TabNet	0.479	0.524	0.333	0.072	0.118
Stacking	0.68	0.633	0.603	0.504	0.549

If we compare our results to other expertise finding studies mentioned in subsection B, 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 there are plenty of photography styles and the interpretations of a picture can vary depending on the viewer. If we

compare our results with the previously mentioned Flickr study [6], 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, which features contribute the most to the prediction, each one of the three feature categories that we established is important. Looking at Fig. 3, the importance of avgTaggedUsers above the others stands out, allowing us 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. Also, the polarity of the captions is similar in both groups.

Finally, talking about RQ3, MANOVA results manifest that there is a statistical difference between both groups based on the features obtained. Fig. 4 and ANOVA results provide us with more detailed information, revealing that avgHeight, imagePercentage, 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.

B. Latent Competencies Reflected in Instagram Profiles

While the primary aim of this study was to classify professional photographers, the extracted features also reveal broader underlying competencies that characterize how users engage with visual creation and presentation on Instagram. Based on our results, the most predictive features can be meaningfully grouped into distinct domains of user capability:

- Visual-technical competence – a user's ability to produce technically high-quality images, as captured by features such as avgTechScore, varTechScore, avgScore, avgHeight, and avgWidth.
- Visual diversity and adaptability – the capacity to work across varied formats, compositions, or tools. It is reflected in the variability of image properties and formats, including varHeight, varPhotosInSidecar, varScore, videoPercentage, and sidecarPercentage.
- Platform navigation competence – the ability to strategically use Instagram's built-in tools and affordances. Relevant features include avgTaggedUsers, captionAvgHashtags, hasLink, and isBusiness. Users with higher values in these variables demonstrate fluency in leveraging the platform for visibility, discoverability, and possibly commercial purposes.
- Audience engagement competence – the skill of attracting, maintaining, and interacting with an audience. This is reflected in features such as followers, avgLikes, avgComments, and numberOfPosts.
- Narrative and expressive competence – the ability to convey meaning, identity, or emotion through text. Captured through features like captionLength, captionEntropy, captionDifficultWords, cLength, and cSubjectivity, reflecting how users complement visual storytelling with textual expression.
- Social and collaborative orientation – the degree to which a user engages with others or presents themselves as part of a

network. Key indicators include avgTaggedUsers, useLocation, isProfessional, and hasLink.

These domains suggest that the content and metadata associated with user profiles encode more than creative output - they also reflect communication, strategic, and identity-related competencies. Beyond artistic and communicative skills, Instagram profiles may also signal non-artistic competencies such as digital literacy, strategic self-presentation, entrepreneurial orientation, attention to detail, and time management. For instance, consistent use of platform features, curated content, business account markers, regular posting schedules, and the presence of structured or accessible captions can reveal planning, awareness, and goal-directed behaviour. Our findings indicate that such competencies can be inferred from indirect, observable features. This opens another future work direction which could explore competency modelling across creative domains, and building more nuanced user representations beyond binary classification.

C. 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.

From another perspective, there are several challenges with using Instagram as a data source. Firstly, data privacy concerns can arise as users become more concerned about their digital privacy. Therefore, platforms like Instagram might limit data accessibility, making it challenging to obtain comprehensive datasets. Secondly, Instagram's content delivery algorithm is continually evolving. As a result, the content a user sees and engages with might not necessarily be a direct reflection of their photographic skill but more of what the algorithm determines as relevant.

In order to mitigate these challenges, researchers can maintain transparency with users about data collection purposes, ensuring that the acquired data is anonymized and securely stored. They also could use a blend of active engagement metrics (e.g., likes, comments) and passive ones (e.g., view duration) and continually update the models to accommodate changes in the platform's content delivery algorithm.

D. 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.

E. 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.

VI. 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. We applied a feature engineering process to add NLP, IQA and other computed features to maximise the information of the dataset. Then, we undertook a data preprocessing and we trained some ML models to classify professional photographers’ profiles on Instagram. Utilising a RF classifier, the research successfully identifies a ubiquitous structural pattern among professional photographers’ Instagram profiles. Notably, professional photographers tend to include twice as many user tags within their posts compared to average users. They also differ more in the height of the pictures posted, also showed higher technical quality. 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.

CREDiT AUTHORSHIP CONTRIBUTION STATEMENT

Sofia Strukova: Conceptualization, Methodology, Investigation, Writing -Original draft preparation.

Daniel Sánchez-Rodríguez: Data curation, Visualization, Software, Writing -Original draft preparation.

José A. Ruipérez-Valiente: Conceptualization, Methodology, Validation, Writing -Review and Editing.

DATA STATEMENT

The data that support the findings of this study are available upon reasonable request.

DECLARATION OF CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

- [1] S. Strukova, J. A. Ruipérez-Valiente, A Framework for Data-Driven Computer-Based Diagnostics of Competencies and Capabilities Across Contexts, pp. 57–81. *Cham: Springer Nature Switzerland*, 2025, https://doi.org/10.1007/978-3-031-87740-7_4
- [2] J. J. Van Bavel, C. E. Robertson, K. Del Rosario, J. Rasmussen, S. Rathje, “Social media and morality,” *Annual Review of Psychology*, vol. 75, pp. 311–340, Jan. 2024, doi: <https://doi.org/10.1146/annurev-psych022123-110258>
- [3] A. Whiting, D. Williams, “Why people use social media: a uses and gratifications approach,” *Qualitative market research: an international journal*, vol. 16, no. 4, pp. 362–369, 2013, doi: <https://doi.org/10.1108/QMR06-2013-0041>
- [4] E. Lee, J.-A. Lee, J. H. Moon, Y. Sung, “Pictures speak louder than words: Motivations for using instagram,” *Cyberpsychology, behavior, and social networking*, vol. 18, no. 9, pp. 552–556, 2015, doi: <https://doi.org/10.1089/cyber.2015.0157>
- [5] S. Kemp, “Digital 2023 april global statshot report,” 2023. [Online]. Available: <https://datareportal.com/reports/digital-2023-aprilglobal-statshot>
- [6] S. Strukova, R. G. Marco, F. G. Mármol, J. A. Ruipérez-Valiente, “Identifying professional photographers through image quality and aesthetics in flickr,” *Expert Systems*, Dec. 2023, doi: <https://doi.org/10.1111/exsy.13526>
- [7] J. Kim, S. Lee, “Deep learning of human visual sensitivity in image quality assessment framework,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1676–1684, doi: <https://doi.org/10.1109/CVPR.2017.213>
- [8] A. Mittal, A. K. Moorthy, A. C. Bovik, “Noreference image quality assessment in the spatial domain,” *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012, doi: <https://doi.org/10.1109/TIP.2012.2214050>
- [9] D. Sánchez, S. Strukova, J. A. Ruipérez-Valiente, “Instagram profile database,” 2023. [Online]. Available: <https://github.com/strukovas/DatasetInstagramProfiles>
- [10] L. Chen, A. Roy, “Event detection from flickr data through wavelet-based spatial analysis,” in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 523–532, doi: <https://doi.org/10.1145/1645953.1646021>
- [11] Y. Hu, L. Manikonda, S. Kambhampati, “What we instagram: A first analysis of instagram photo content and user types,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, pp. 595–598, May 2014, doi: <https://doi.org/10.1609/icwsm.v8i1.14578>
- [12] D. Lekkas, R. J. Klein, N. C. Jacobson, “Predicting acute suicidal ideation on instagram using ensemble machine learning models,” *Internet Interventions*, vol. 25, p. 100424, 2021, doi: <https://doi.org/10.1016/j.invent.2021.100424>
- [13] A. Zohourian, H. Sajedi, A. Yavary, “Popularity prediction of images and videos on instagram,” in *2018 4th International Conference on Web Research (ICWR)*, 2018, pp. 111–117, IEEE, doi: <https://doi.org/10.1109/ICWR.2018.8387246>
- [14] W. H. Lim, M. J. Carman, S.-M. J. Wong, “Estimating relative user expertise for content quality prediction on reddit,” in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT ’17, 2017, p. 55–64, Association for Computing Machinery, doi: <https://doi.org/10.1145/3078714.3078720>
- [15] S. Patil, K. Lee, “Detecting experts on quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors,” *Social Network Analysis and Mining*, vol. 6, 12 2015, doi: <https://doi.org/10.1007/s13278-015-0313-x>
- [16] V. Ha-Thuc, G. Venkataraman, M. Rodriguez, S. Sinha, S. Sundaram, L. Guo, “Personalized expertise search at linkedin,” in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 1238–1247, IEEE, doi: <https://doi.org/10.1109/BigData.2015.7363878>
- [17] P. Wesołowski, “Enhancing architectural engineering students’ acquisition of artistic technical competences and soft skills,” *Cogent Arts*

& *Humanities*, vol. 9, no. 1, p. 2043997, 2022, doi: <https://doi.org/10.1080/23311983.2022.2043997>

- [18] V. S. Pagolu, K. N. Reddy, G. Panda, B. Majhi, "Sentiment analysis of twitter data for predicting stock market movements," in *2016 international conference on signal processing, communication, power and embedded system (SCOPEs)*, 2016, pp. 1345–1350, IEEE, doi: DOI:10.1109/SCOPEs.2016.7955659
- [19] S. M. Idrees, M. A. Alam, P. Agarwal, "A prediction approach for stock market volatility based on time series data," *IEEE Access*, vol. 7, pp. 17287–17298, 2019, doi: <https://doi.org/10.1109/ACCESS.2019.2895252>
- [20] D. van Dijk, M. Tsagkias, M. de Rijke, "Early detection of topical expertise in community question answering," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, New York, NY, USA, 2015, p. 995–998, Association for Computing Machinery, doi: <https://doi.org/10.1145/2766462.2767840>
- [21] M. Gil-Ramírez, R. Gómez-de TravesedoRojas, A. Almansa-Martínez, "Political debate on youtube: revitalization or degradation of democratic deliberation?," *Profesional de la información*, vol. 29, no. 6, 2020, doi: <https://doi.org/10.3145/epi.2020.nov.38>
- [22] P. P. Tricomi, S. Kumar, M. Conti, V. Subrahmanian, "Climbing the influence tiers on tiktok: A multimodal study," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, pp. 1503–1516, May 2024, doi: <https://doi.org/10.1609/icwsm.v18i1.31405>
- [23] M. Kostic, H. F. Witschel, K. Hinkelmann, M. Spahic-Bogdanovic, "Llms in automated essay evaluation: A case study," *Proceedings of the AAAI Symposium Series*, vol. 3, pp. 143–147, May 2024, doi: <https://doi.org/10.1609/aaais.v3i1.31193>
- [24] A. K.-K. Alexander Graf, "Instaloder: Instagram scraper repository," 2016. [Online]. Available: <https://github.com/althonos/InstaLooter>
- [25] K. Seshadrinathan, T. N. Pappas, R. J. Safranek, J. Chen, Z. Wang, H. R. Sheikh, A. C. Bovik, "Image quality assessment," in *The Essential Guide to Image Processing*, Boston: Academic Press, 2009, pp. 553–595, doi: <https://doi.org/10.1016/B978-0-12-374457-9.00021-4>
- [26] L. Kang, P. Ye, Y. Li, D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740, doi: <https://doi.org/10.1109/CVPR.2014.224>
- [27] H. Talebi, P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018, doi: <https://doi.org/10.1109/TIP.2018.2831899>
- [28] N. Murray, L. Marchesotti, F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE conference on computer vision and pattern recognition*, 2012, pp. 2408–2415, IEEE, doi: <https://doi.org/10.1109/CVPR.2012.6247954>
- [29] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Jay Kuo, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015, doi: <https://doi.org/10.1016/j.image.2014.10.009>
- [30] Z. Wang, J. Chen, S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020, doi: <https://doi.org/10.1109/TPAMI.2020.2982166>
- [31] A. Ly, B. Uthayasooryar, T. Wang, "A survey on natural language processing (nlp) and applications in insurance," 2020, doi: <https://doi.org/10.48550/arXiv.2010.00462>
- [32] S. Loria, "Textblob: Simplified text processing," 2013. [Online]. Available: <https://github.com/slوريا/textblob>
- [33] S. Bird, "Natural language toolkit (nltk)," 2006. [Online]. Available: <https://github.com/nltk/nltk>
- [34] S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st ed., 2009.
- [35] A. Ward, "Textstat: Nlp python package," 2014. [Online]. Available: <https://github.com/textstat/textstat>
- [36] Y. Karaca, M. Moonis, "Chapter 14 -shannon entropybased complexity quantification of nonlinear stochastic process: diagnostic and predictive spatiotemporal uncertainty of multiple sclerosis subgroups," in *MultiChaos, Fractal and Multi-Fractional Artificial Intelligence of Different Complex Systems*, Academic Press, 2022, pp. 231–245, doi: 10.1016/B978-0-323-90032-4.00018-3



Sofia Strukova

Sofia Strukova has an interdisciplinary background in computer science and Big Data. She earned her B.Sc. in computer science from Moscow Power Engineering Institute, Russia, and subsequently her M.Sc. in Big Data and Ph.D. in Computational Social Science from the University of Murcia, Spain. Her research interests revolve around computational social science, expertise finding, educational technology, data mining and data science in general. More info at <https://strukovas.github.io/>



Daniel Sánchez-Rodríguez

Daniel Sánchez-Rodríguez received his B.Sc. degree in computer science from the University of Murcia. His research interests include artificial intelligence, software development, data analysis, and computer science in general.



José A. Ruipérez-Valiente

José A. Ruipérez-Valiente received his B.Eng. degree in telecommunications from Universidad Católica de San Antonio de Murcia in 2011 and a M.Eng. degree in telecommunications in 2013, together with his M.Sc. and Ph.D. degrees (2014 and 2017) in telematics from Universidad Carlos III of Madrid while conducting research with Institute IMDEA Networks in the area of learning analytics and educational data mining. He was a postdoctoral associate at MIT. He has received more than 20 academic/research awards and fellowships, has published more than 130 scientific publications in high-impact venues, and participated in over 24 funded projects. He is currently an Associate Professor of Computer Science and Artificial Intelligence at the University of Murcia. More info at <https://webs.um.es/jruiperez>