Sofia Strukova and Ruipérez-Valiente, J. A. Using Online Digital Data to Infer Valuable Skills for the Modern Workforce. In Handbook of Research on New Media, Training, and Skill Development for the Modern Workforce. (pp. 89 - 109). IGI Global. ISBN 9781668439968

DOI: 10.4018/978-1-6684-3996-8.ch005

https://doi.org/10.4018/978-1-6684-3996-8.ch005

Using Online Digital Data to Infer Valuable Skills for the Modern Workforce

Sofia Strukova and José A. Ruipérez-Valiente
University of Murcia, Spain

ABSTRACT

This chapter uncovers the opportunities that online media portals like content sharing and consumption sites or photography sites have for informal learning. The authors explored online portals that can provide evidence of evaluating, inferring, measuring skills, and/or contributing to the development of competencies and capabilities of the 21st century with two case studies. The first one is focused on identifying data science topical experts across the Reddit community. The second one uses online Flickr data to apply a model on the photographs to rate them as aesthetically attractive and technically sound, which can serve as a base for measuring the photography capabilities of Flickr users. The presented results can serve as a base to replicate these methodologies to infer other competencies and capabilities across online portals. This novel approach can be an effective alternative evaluation of key 21st century skills for the modern workforce with multiple applications.

Keywords: Computational Social Science, Data-Driven Evaluation, Data Science, Expert Identification, Reddit, Flickr, Competencies, Capabilities, Informal Learning

INTRODUCTION

Every day there are more people that strongly believe that a significant proportion of learning, both intentional and unintentional, is happening online due to two different reasons: the ever-growing popularity of internet usage by all segments of the population and the appereance of more and more socio-political, climate and health emergencies like COVID, that are demanding remote work and education modalities. According to the UNESCO report regarding the impact of COVID-19 on education (UNESCO, 2021), almost half of the world's students are still affected by partial or complete school closures. In this sense, the education and training stakeholders have to adapt to students' new needs and abilities by creating modern training techniques and providing flexible and personalized educational approaches. Moreover, the pandemic has caused traumatic circumstances to many people, making it harder for learners under duress to focus on learning (Carter et al., 2020). One of the solutions to this problem is to create easy-to-access, engaging, online educational opportunities, including resources like electronic books, recorded lectures, quizzes, podcasts, discussion forums, or interviews.

Due to this context, many universities around the world are integrating online learning into their courses, starting from formal and structured learning management systems (LMSs), such as Moodle[1] or Sakai[2], to informal social networking sites (SNSs), such as Facebook[3] or Flickr[4] (Ulla & Perales, 2021), among many others. Nowadays, the use of these environments is becoming more common as it can help improve the missing social aspects of remote learning in order to avoid disrupting the system when climate, health,

and socio-political crises emerge. However, LMSs do not provide the functionality to facilitate engaging ways to interact with peers or exchange feedback. On the other hand, with the use of SNSs, students can present themselves, articulate their thoughts to their social networks, develop or preserve connections with peers and share information, knowledge, and artifacts within a community. Moreover, SNSs promote the creation of shared interest groups that make users feel a sense of community engagement (McCarthy, 2017).

Therefore, if research thus far has proved that students benefit from learning methods that involve the use of SNSs as part of their curriculum (Shih, 2012), could users learn in these and similar sites without being part of a formal educational program? In this chapter, the researchers will focus their attention on a trending and controversial issue of particular importance related to the informal learning happening across various online portals capable of holding evidence of users' competencies and capabilities of different nature. The authors will discuss the opportunities that digital media online provide to evaluate and generate feedback on learners' skills. The conclusions of this study, in return, will help to discuss the affordances of new online media to infer 21st-century skills and how the modern workforce can benefit from these approaches.

Thus, the objectives of this chapter are as follows:

1. To explore online portals that can provide evidence of accessing, inferring, measuring skills, and/or contributing to the development of competencies and capabilities of the 21st century.
2. To analyze if it is possible to identify topical experts across the Reddit[5] community by examining comments of the selected subreddit.
3. To predict which images a typical user would rate aesthetically attractive and technically good in Flickr, which will serve as a base for measuring the photography capabilities of Flickr users.

The remainder of this chapter is structured as follows. First, the researchers will focus on reviewing related works of this study that have explored online portals, which have already shown evidence of inferring users' competencies and capabilities of different nature. Next, the authors will describe two case studies that they have performed: a case study on data science competencies in Reddit and a case study focused on measuring photography capabilities through image quality and aesthetics in Flickr. Then, the researchers will present a detailed section discussing the possibilities that some online portals have to enhance education and learning, as well as depicting potential applications of the data from online digital media to evaluate and provide feedback on 21st-century skills. Finally, the authors will draw their conclusions and future research directions.

BACKGROUND

An ever-changing society and technology require a disruptive transformation of the 21st workforce in terms of the needs and capabilities to adapt to. Now more than ever, society needs to re-envision new ways of training and evaluating the skills required for the population to function in the modern world successfully. Luckily, multiple new approaches and media have emerged over the last decade, generating opportunities for novel teaching and learning approaches. These have empowered traditional educational environments and strengthened knowledge acquisition through informal learning with innovative technologies and services in the new digital world (Strukova et al., 2022). In this way, this section will provide a discussion of diverse online digital media that have already proved to be able to generate rich informal learning experiences for their users. In the scope of this study, the authors are especially interested in internet portals where users can interact with content and other peers. Naturally, the users leave traces of their behavior within these digital environments, and the researchers argue that these

interactions hold the potential to help to evaluate the aforementioned critical capabilities for the modern society.

Within the myriad of internet websites that the researchers find online, they are specifically interested in photo and video sharing platforms (e.g., Pinterest[6] and YouTube[7]), as well as question and answer (Q&A) portals which can be represented, for example, by forums (e.g., Tripadvisor forum[8]), an access point to news (e.g., Financial Times[9]), or content sharing and consumption portals (e.g., Reddit). They all can provide valuable information about the proficiency of users in specific skills and knowledge using the traces of the interactions performed by the users online. In portals such as the ones previously mentioned, the authors detected the following commonalities: the users are responsible for both consuming, generating, and sharing the content; they can interact with other peers, follow existing topics of interest, or create communities around niche-based topics, connecting with professionals in the selected industry, broadcasting live videos and even finding inspiration. Most importantly, by completing the actions mentioned earlier, the users can learn a wide range of valuable skills and competencies, from soft skills to knowledge of a particular topic.

The authors found in the literature several studies that focused on the approach of analyzing the competencies of users by processing their traces online in a way that could provide evidence of their knowledge and skills, and/or contribute to the development of competencies and capabilities of the 21st century. For example, Packiam and Geoffrey performed research aiming to investigate the effect of SNSs engagement on cognitive and social skills. They examined the use of three SNSs, namely Facebook, Twitter[10], and YouTube, in a group of young adults testing their working memory, attentional skills, and levels of social connectedness. The authors concluded that certain activities on Facebook, such as checking friends' status updates and on YouTube, such as sharing a video with a friend, positively influenced working memory test performance. On the other hand, the results indicated that active and passive SNS users had qualitatively different profiles of attentional control (Packiam & Geoffrey, 2012).

Alternatively, Pal et al. presented an approach to find topical authorities in Instagram[11]. Their method relied on the self-described interests of the list of users following popular accounts. The authors inferred regular users' interests from their self-reported biographies that are publicly available and used Wikipedia[12] pages to ground these interests as fine-grained, disambiguated concepts. The authors concluded that individual user biography-based interests provided strong evidence to infer the topical authorities (Pal et al., 2016).

Also, Vesselinov and Grego studied the effectiveness of Duolingo[13], an online gamified platform for learning languages (Vesselinov & Grego, 2012). As participants of the Duolingo effectiveness study announced on its web, the authors analyzed a random representation of native English users who studied Spanish. The participants took a placement Spanish language test at the beginning of the study and one test at the end of the study. The authors concluded that after using Duolingo, language proficiency improved significantly. Also, this study estimated that a person with no knowledge of Spanish would need around 34 hours on average to cover the material for the first college semester of the Spanish course. Moreover, the main factor for higher effectiveness was the participants' motivation, which highlighted the strong interest of users to study for traveling. Another catalyst of the higher effectiveness was the initial language knowledge level, indicating that beginners were the ones that improved the most.

The final example is the work done by Yan et al. focused on the professional social network LinkedIn[14] (Yan et al., 2019). In their paper, the authors developed a framework for collecting validations for members' skill expertise at the scale of billions of member-skill pairs and a machine learning (ML) model to make suggestions to collect validations more efficiently. They discovered insights on how users evaluated their mates in professional social networks. For example, juniors often were getting better evaluations than the members with higher seniority. The authors evaluated their model by predicting who

was hired for a job requiring a particular skill. After all, the experiments estimated the members' skill expertise accurately at a large scale and offered a benchmark to validate social theories on peer evaluation.

All of the examples presented above have several common characteristics:

- All the previously mentioned portals (language learning portal, professional social network, photo, and video sharing platform, SNS) in the web generate large amounts of data stemming from the interactions carried out by their users in such contexts.
- All these studies performed a measurement of essential competencies or capabilities necessary in the 21st century (language proficiency, social and cognitive skills, topical expertise).
- The users whose data was analyzed across these studies were responsible for consuming and generating the content.

After analyzing the existing studies, the authors saw that the current academic works do not discuss the evidence of the online digital portals to measure modern up-to-the-minute competencies essential for the unknown future of the 21st century. These portals hold the potential to do it since they contain a large load of data that can be easily accessed through the Application Programming Interface (API), by scraping the public domain, or by having direct access to the databases. Therefore, in this investigation, the researchers will show two case studies representing a couple of these portals - the photography social network and the Q&A portal, which will be described next.

CASE STUDIES

Identifying Experts in Question & Answer Portals: A Case Study on Data Science Competencies in Reddit

Context and Motivation

With the growing popularity of Q&A portals, they have become the leading platform for all kinds of users to share information, ask questions, and comment on the doubts of others. Users help each other by providing altruistic responses, making the resolution of doubts across many topics much more straightforward. However, many questions are never answered because the question statement lacks a good structure, or an appropriate expert did not see it. Users can read and reread their questions before posting them on the platform, which could help turn a poor-quality question into a good quality question helping not only the user but also the community. Unfortunately, in many Q&A portals, there is not a metric that suggests whether the questions or answers are poor-quality or high-quality. This is due to the fact that it is not a trivial task to interpret the human language. On these grounds, natural language processing (NLP) is trying to fill this gap by focusing on creating models that can understand and interpret human language.

On the other hand, data science is the study of the generalizable extraction of knowledge from data. It is an emerging discipline that combines expertise across various domains, including software development, data management, and statistics (Saltz et al., 2017). A data scientist requires an integrated skill set spanning mathematics, ML, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions (Dhar, 2013). In various fields such as computer science, computer engineering, business administration, and business management, a subset of courses focus on skills that are essential for being a data scientist. However, still, there is a need to introduce the integration of other skills needed to function as a data scientist.

On this wise, there is an enormous interest in identifying experts in the data science field, and the Q&A portals can hold evidence to do it. Therefore, the authors will focus this case study on finding data science experts across one of them. Next, they will describe several Q&A portals in order to choose the most appropriate one for developing a case study within this gap.

Question & Answer Portals

Q&A portals can be defined as online discussion sites where users can post messages asking questions or replying to others' questions. In other words, they are collective knowledge creation sites where users contribute to generating the content while many users access these questions and responses by searching online. All users' activities are usually incentivized by reward or penalty with gamification features (like points or badges) that are triggered based on their actions or on the reviews given by other users through voting or choosing favorites (Raj et al., 2011). Next, the authors will explore the most popular Q&A portals.

- Answers.com[15], formerly known as WikiAnswers, states as its goal to help students of all ages learn, study, and connect in an online community. Therefore, Answers.com can be considered as an encyclopedia-type Q&A portal. Across it, anyone can pose an inquiry, answer a question, or improve a previous response to a question.
- AskFm[16], unlike the rest of the websites, is a social networking platform, which uses a Q&A format to connect users through conversational exchanges (Farrugia et al., 2019). Questions and answers become public on a user's profile only once they are answered. Also, users can configure the settings in such a way that they do not allow anonymous questions. Moreover, users can refuse to reply or delete questions they do not like.
- Quora[17] is an innovative Q&A site with a rapidly growing user community that differs from its competitors by integrating a social network into its basic structure (Wang et al., 2013). Users sign in and can immediately start searching for answers to specific questions or topics, which are subject headings assigned by users. Each Quora question has its own page, which includes a list of its answers and a list of related questions. Users can add new answers and comment, edit, and vote on existing answers. Users can follow topics that allow them to see all questions from a given topic in their feed. Also, users can have contacts, and their questions will appear on the Quora feed. Unlike in other Q&A portals, Quora allows users to follow each other (social connections in Quora are directional) to form a social network. On the other hand, Quora focuses on leveraging social connections to get questions answered by encouraging users to post Quora questions to SNSs, giving the question more exposure (Ovadia, 2011).
- Reddit reports more than 52 million daily active users who contribute to more than 100 thousand communities of an enormous variety of topics. One of the portal's strengths is its sense of community. There are already many specific online communities in Reddit (called subreddits) for a diverse range of topics, and otherwise, it can be easily created. Within each of these subreddits exists a unique community with a distinct sub-culture. Users are attracted by the sense of belonging, as well as a feeling of validation when a submission or a comment is upvoted. There is a hierarchical nature in the portal due to the earning of karma points. Reddit created its etiquette, motivating community members to respond to rudeness, trolling, or spam with downvotes and reporting of the guilty parties.
- StackOverflow[18] is a portal mainly for computer programmers' and developers' questions. The users who ask the questions can add unique tags to help others discover what the question is about. The questioner can select one answer as the most helpful one (called the accepted answer). At the same time, other registered members can vote on questions and answers. The positive and negative votes (called upvote and downvote, respectively) show how helpful that question/answer

was for the audience. Each site member has a reputation score which changes as users participate in different activities across the website, such as posting questions, answers, or comments. A higher reputation value also means more privileges for a user member, such as the ability to edit questions or answers or close a Q&A thread. The interactive nature of StackOverflow makes it possible for both questioners and responders to clarify the vagueness in a question or answer. Moreover, StackOverflow has three properties of new social learning technologies. It supports learners to find the right content using natural language, not just relying on keywords; users can connect with the right people. It motivates them to learn by encouraging them in the question/answer game with the reputation incentive gained from votes.

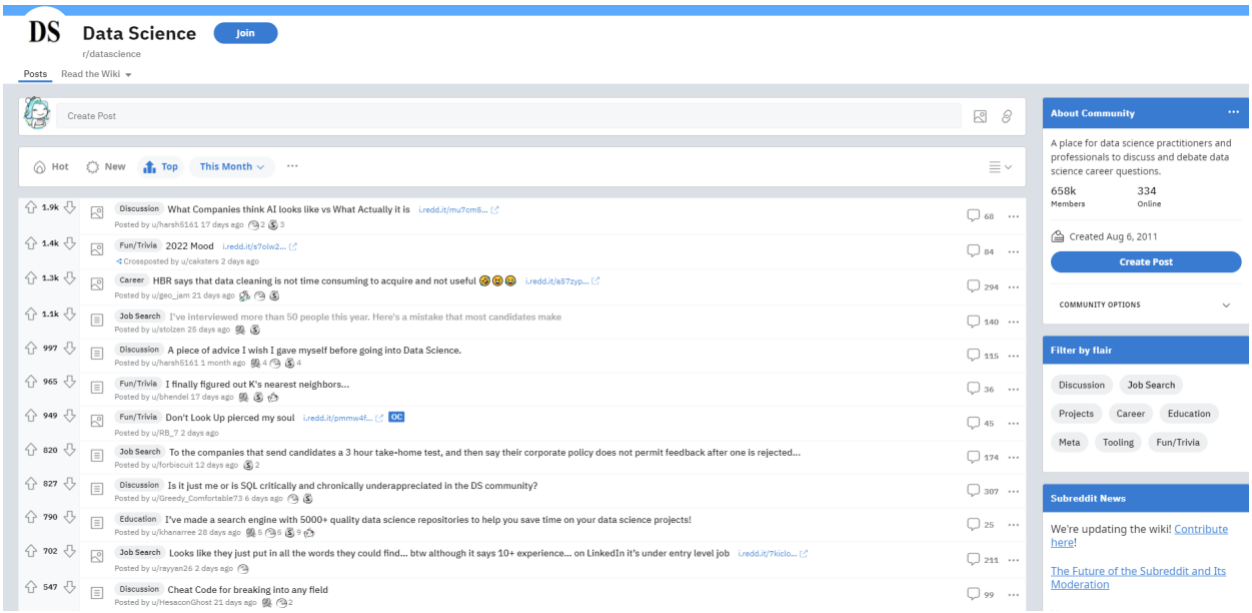
- Yahoo! Answers[19] has shut down as of May 4, 2021[20]. However, it was one of the most important Q&A portals facilitating the preservation and retrieval of answered questions aimed at building an online knowledge base meaning to encompass any topic (Gyongyi et al., 2008). In order to post a question, a registered user had to provide a short question statement, an optional longer description, and had to select a category to which the question belongs. A question could be answered over a seven-day open period, and then the user who asked the question could pick the best answer or put the answers for a community vote. Questions that remained unanswered or ended up with the majority vote on "No Best Answer" were deleted from the system. When the best answer was selected, the question became resolved and was stored permanently in the system. However, users of Yahoo! Answers could not answer their own questions; thus, an honest discussion could not emerge. Yahoo! Answers had no support for threads that would be essential for diverging discussions (Gyongyi et al., 2008). Finally, it is essential to mention that users were accumulating points through their interactions with the system what was explicitly demonstrated in the leaderboard.

Summary. After exploring the mentioned above portals, the authors concluded that they would focus on working with the most active Q&A portal, namely Reddit, being visited by 430 million active users every month. Across Reddit, there are many subreddits focused on diverse topics of interest. This makes Reddit more attractive for research because the results can serve as a base for identifying topical experts in other subreddits. Moreover, the portal provides an API[21] that can facilitate the step of downloading the required data.

Data Science Experts and Methodology for their Identification

Target. Figure 1 shows an example of the "Data Science" subreddit describing itself as "A place for data science practitioners and professionals to discuss and debate data science career questions," counting up to 658,000 members. The posts are chosen for the current month (January 2022) and sorted by the "Top" criteria, which takes into consideration the most upvotes regardless of downvotes, while "Hot" counts the most upvotes recently, meaning that the content is rapidly becoming in demand and "New" shows the most recent posts. Next to each post, there can be seen its score, which is counted as the number of upvotes minus the number of downvotes, along with the arrows for upvoting or downvoting, the number of comments, the nickname of the author, and his awards. It is also feasible to filter posts by flair suggesting the following keywords: "Discussion," "Job Search," "Projects," "Career," "Education," "Meta," "Tooling," "Fun/Trivia." Most importantly, at the top of the page, every user can ask a question by clicking on "Create Post," where it is possible to provide a title of the question, elaborate on a more detailed text, attach an image, a link, or a poll, and choose a flair.

Figure 1. Data Science thread on Reddit



Data collection. The authors downloaded all the posts and the respective comments and subcomments (comments to comments) of the "Data science" thread in Reddit from May 2020 to April 2021, a representative data set for identifying experts in the data science field. They decided to use Python's Python Reddit API Wrapper (PRAW)[22] package because it eases Reddit's official API access. The total number of posts that the authors downloaded is 16,436, and the total number of comments is 100,052 (44,029 comments and 56,023 subcomments). On average, every post has six comments, including subcomments. The semi-supervised ML method started with unlabeled data and adopted a data-driven approach to label expert, non-expert, and out-of-scope comments. Two data scientists with relevant experience following the methodology described by Gobert et al. (Gobert et al., 2013) manually labeled approximately 1,000 random comments into these three categories. The labeled data will be used as ground truth and will be needed for ML models to learn the underlying patterns.

An example of one thread that has at least one comment of each type is as follows (the punctuation and spelling of the authors are retained):

> Post: "*Best software for large tables?*"
>
> Expert comment: "*Pandas will work as long as you have more RAM than data which if it's 2 only gb shouldn't be a problem. If you get too big for that you'll have to switch to sql.*"
>
> Non-expert comment: "*SQL database*"
>
> Out-of-scope comment: "*Now that I'm asking, what software would be best for analyzing adjacent tiles on a 2d grid? I have a dataset with x and y coordinates, and I feel like there should be a more efficient way to do this than editing and merging tables using a single processor core.*"

Model training. The authors trained the following multi-class supervised models based on the created labels: Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). At each step, the researchers performed 10-fold cross-validation on the test set. The features that they used for training the ML models are divided into the following three groups:

- **NLP features.** These are the ones that can be extracted from words, sentences, and phrases. This group of features includes comment length, word count, character count, sentence count, average word length, average sentence length, the average number of subjective words in posted answers,

several readability scores, readability time, and term frequency-inverse document frequency derived features.

- **Crowdsourced features.** These involve information or opinions from a group of people who submit their views via the Reddit site. These features include the karma of the comment and its score (the number of upvotes minus the number of downvotes).
- **User features.** These aim to gauge the activity level of the users who wrote the initial comment on Reddit. This group of features is limited compared to other Q&A portals because the API of Reddit allows getting only the awardees' karma, the awarded karma, a verified account boolean, a verified email boolean, total karma, and comment karma. Recently, Reddit introduced the ability for users to see their followers instead of the raw count. However, the list of users' followers is not publicly available. Thus, the authors could not include the number of followers or the percentage of bidirectional friends. On the other hand, the researchers decided to make use of the rest of the comments that were not used for manual labeling. For each user whose comments happened to be labeled, the authors computed the number of comments and the number of posts throughout the stated timeline. Also, they computed the average score per author, the average number of words in posted questions and posts, account age, average response time, and average upvotes.

The authors used several feature selection methods to be able to compare results across various groups of features.

Model evaluation. ML models are parameterized so that their behavior can be tuned for a given problem. Supervised learning is a family of ML algorithms that focus on finding patterns between input variables and their respective labels. In this way, a supervised learning algorithm analyzes the labeled data and produces an inferred function, which can be used for mapping new cases. An optimal scenario will allow the algorithm to determine the class labels for unseen instances correctly. There is a wide range of supervised learning algorithms, each having its strengths and weaknesses. These models can have many parameters, and finding the best combination of parameters is a separate task. To do so, the authors measured the performance of the models using two standard measures: accuracy and Area Under Curve (AUC). Accuracy is the total percentage of correctly classified elements. AUC is a more comprehensive measure of how well the classifier can distinguish between classes. In other words, AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example.

Preliminary Results and Future Steps

According to these metrics, the RF is the best algorithm (Table 1) to predict if any given comment is an expert comment, a non-expert comment, or an out-of-scope comment. The two other models, namely LG and DT, also show decent results displaying accuracy scores as 0.78 and 0.74 and AUC scores as 0.91 and 0.87, respectively. RF shows an accuracy equal to 0.83 and AUC equal to 0.93, which proves its better performance and means that it has a very high precision. To obtain these results, the following features were selected as the best features: the age of the account of the user who left the comment, the average word length and the number of difficult words of the comment, the average number of words per comment, and posts per author, comment karma, the number of data science terms, the polarity of the comment, several readability scores - Coleman-Liau index (Coleman & Liau, 1975), Dale-Chall readability score (Edgar & Jeanne, 1948) and Spache readability formula (Spache, 1953), and the term frequency-inverse document frequency derived features. These exact features are important to detect experts because they represent essential characteristics of the user who performed the comment and the comment itself. As said, the account age and the average number of words and posts of the author show the activity that this concrete user performed on Reddit and how detailed their answers were. On the other

hand, the rest of the best features pertain to the NLP group, emphasizing the fact that the comment content itself is quite important.

Table 1. Comparison of LR, RF, and DT algorithms by accuracy and AUC metrics

	Logistic Regression	Random Forest	Decision Tree
Accuracy	0.78	0.83	0.74
AUC	0.91	0.93	0.87

The authors believe that the obtained results prove the potential that Q&A portals open data have to support finding topical experts as it has been done in Reddit. By estimating the user expertise, it is also possible to infer the quality of the content because expert users have a higher probability of producing better quality content. Moreover, through the detection of user expertise of content authors, it is feasible to predict the information quality of the content despite the lack of user votes as well as match open questions to potential experts. On the other hand, Q&A sites are a primary source of news and information that can be steered, distorted, and influenced. In this way, potential malicious or unreliable users can be detected, and their influence can be reduced. From another perspective, the prediction of content information quality could then be applied to improve the curriculum and formative evaluation of academic programs. Finally, effectively identifying the expert in each domain is a key to better understanding user engagement.

Identifying Experts in Photo & Video Sharing Portals: A Case Study on Photography Capabilities in Flickr

Context and Motivation

Digital photography technology has advanced substantially during the last decade. With this progression, the photography capabilities of people are also being developed and enhanced. Concretely, nowadays, people are more willing to improve their photography capabilities and publicly share their pictures. In this way, the authors investigate the possibility to measure the photography capabilities of photo and video sharing platform users through the guiding principles of photography, namely, image aesthetics and its technical quality.

Photo & video sharing platforms are widely used nowadays. Users can upload their video or photographs, discover the content of others, evaluate, share and save it, be members of interest groups and provide valuable feedback to other users. However, content with many comments or likes does not necessarily mean that it is high quality content. Therefore, much research focuses on identifying technically good and aesthetically attractive pictures. Unfortunately, it is not a trivial task because photography skills are subjective. Thus, the authors will aim to fill this gap by making an exploratory analysis of how deep learning models can help identify professionals in this field through photo sharing platforms.

Photo & Video Sharing Portals

This section will review several photo and video sharing sites.

- Flickr is a cloud storage portal not only for storing but also for sharing photographs. Its audience includes diverse profiles of both professional and amateur photographers who want to share their portfolios. Flickr is also geared toward beginners and enables them to edit the photos directly on the platform, such as adjusting brightness and contrast and applying various filters. Flickr provides a unique cloud storage feature to share photos with the public or in subject groups. Flickr users can upload their own photos, post them in galleries and groups, add photos to favorites, write comments, or follow other users. Moreover, the portal offers a Pro account that

provides the ability of an unlimited number of uploads, while the free version of Flickr allows 1000 photos. Flickr is different from other photo-sharing websites due to its large community.

- 500px[23] is an appropriate portal for serious photographers offering an image-focused design and providing a clean and elegant way to display their best shots. It is a portal where photographers can gain exposure, connect with other professionals or find inspiration. Users can organize their pictures into sets of a particular theme or stories of a particular event. The free version of the service allows to upload up to seven photos per week. Once the user uploads a photograph, it will have a 24 hours lifespan in order to gain popularity measured in likes and comments. If it does not get enough reputation, the photo will have no chance of appearing in the popular feed.
- Instagram is a favorable place to edit, showcase, and share photos with friends or the world. It offers a surprisingly good set of photo storage, sharing, and editing tools. This portal provides filters that anyone is able to edit their photos instantly. To utilize the power of the Instagram community, its users can use hashtags to get more exposure. However, there is no opportunity to share the original-size photo.
- 1x.com[24] is curated by a group of professionals. As said, the uploaded photograph is going to be approved or disapproved by a group of professional photographers. Since only a small percentage of pictures is authorized, getting pictures published on this portal can play a good role for the resume. Other valuable features of this platform include critique, tutorials, and right-click download prevention.
- SmugMug[25] is a design-focused photo storage site with a custom homepage and many design templates allowing to build an online gallery for professional users. This portal is designed to meet the portfolio needs of photographers wanting customizability, presentation, and e-commerce. It follows the fact that there is no free version of the subscription. Regardless of the plan, users get unlimited storage of photos up to 500MB in size and a wide range of editing tools.

Summary. According to Rahayuningsih and Yuniarti (Rahayuningsih & Yuniarti, 2017), professional photographers and photography aficionados with a deep interest in photography prefer to upload their shots to professional portals such as 1x.com or Flickr in order to obtain criticism and advice and encouragements from fellow photographers about the pictures they take so they can improve their knowledge and quality of the photographs. After exploring and testing several photo and video sharing platforms presented above, the authors decided to work with the data from Flickr users. Since it was important to obtain photos from both professional and beginner photographers for this study, the authors selected Flickr because it provides features for all types of users, and the authors were able to accomplish their primary goal of getting diverse types of pictures for further evaluation. The researchers believe that the user data from Flickr has a great potential to infer the photography of its users.

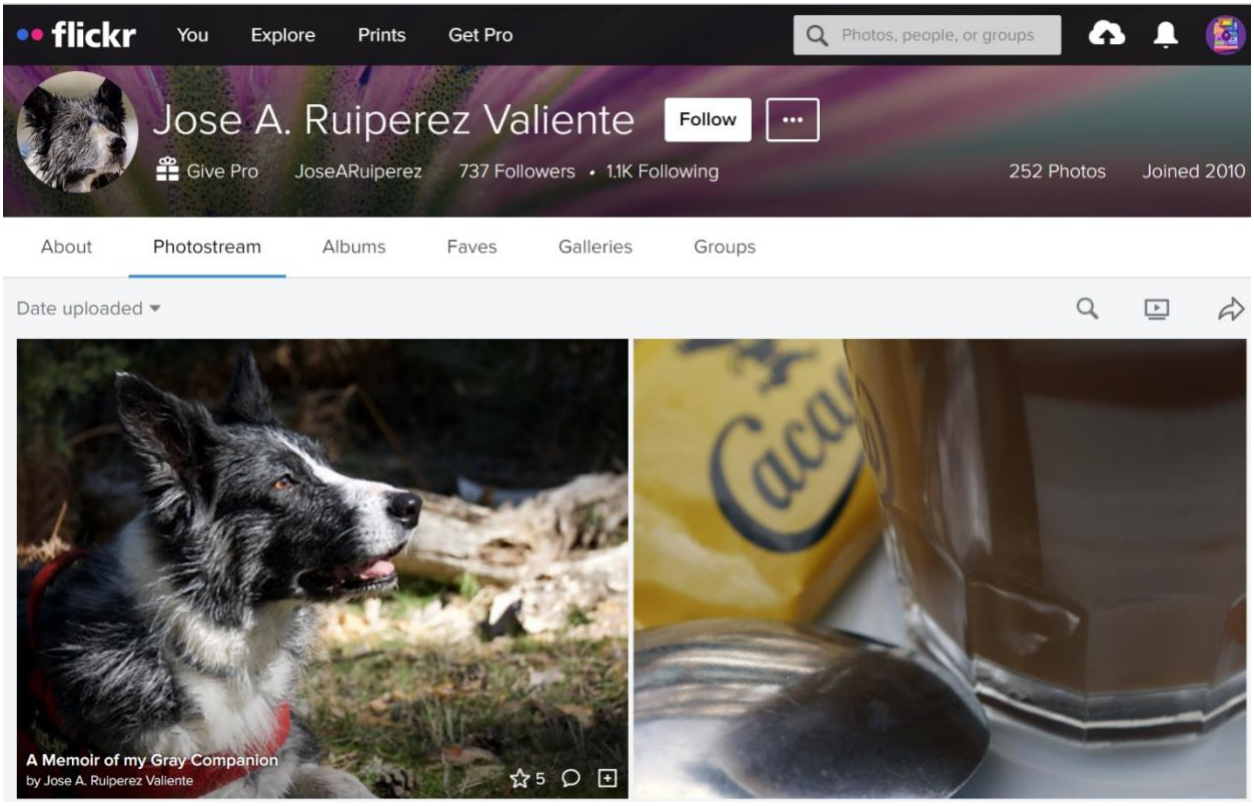
Professional Photographers and Methodology for their Identification

Target. This case study aims to detect which images a typical user would rate as looking aesthetically attractive and technically good, which will serve as a base for measuring the photography capabilities of users. Applying Convolutional Neural Networks (CNN) to the aesthetic quality categorization task is not straightforward since it is challenging to perform automatic feature learning with regard to both the global and the local views of the input images (Lu et al., 2014). By good fortune, there are several already existing approaches of deep CNN trained with human-labeled data, which will serve as a base to measure the photography competencies of users. One can argue that photography skills are something subjective. However, an expert is pre-planning the shooting, all possible lighting approaches, composing the scene, forecasting the character modeling, and all this determines the final effect of a photo (Zhao, 2017).

An example of a users' profile photostream is represented in Figure 2. There also can be observed basic information of the user profile, including the joining year, the number of photos, followers, and

following. In the “About” section, the users can provide their occupation, hometown, current city, country, and website. It also shows the number of views, tags, geotags, how many users added their pictures to favorites, and how many groups the user is a member of. The “Photostream” section shows all the user photos sorted by the date uploaded or date taken, while the “Albums” section represents how the pictures are distributed by particular topics. Moreover, users can mark content of other users as favorites and organize photographs to galleries that are publicly available as well. Finally, Flickr gives its users an opportunity to be members of groups to share photographs with other like-minded members, which can be seen in the “Groups” section. Accordingly, there are communities for professionals and amateurs; some of them are moderated, meaning that there is a need for the approval of every picture. These groups will serve as a base for the data collection.

Figure 2. Photostream of the user JoseARuiperez on Flickr



Data collection. To accomplish the goal of this case study of detecting which images a typical user would rate as looking technically good or aesthetically attractive in Flickr, the researchers downloaded 1,900 images of the Flickr professional group and 1,900 images of the Flickr amateur group.

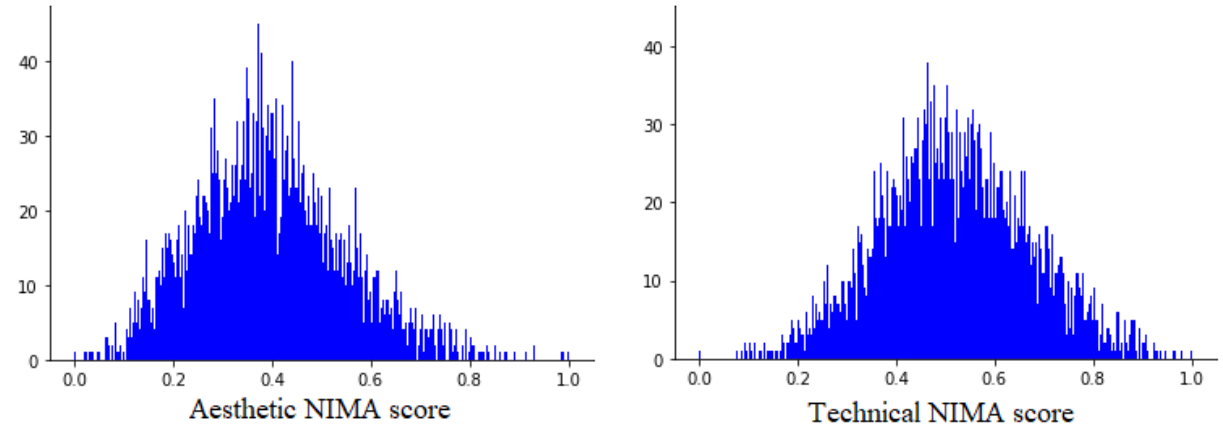
Selected model. From the deep learning models, the authors will use Neural Image Assessment (NIMA) which is a deep CNN that is trained to predict which images a typical user would rate as looking good (technically) or attractive (aesthetically) (Talebi & Milanfar, 2018). NIMA authors explored several classifier architectures and replaced the last layer of the baseline CNN with a fully-connected layer with ten neurons followed by soft-max activations. Baseline CNN weights were initialized by training on the ImageNet dataset (Krizhevsky et al., 2012). Unlike other models, this one produces a distribution of ratings for any given image - on a scale of 1 to 10, assigning likelihoods to each of the possible scores. In this way, the authors can use NIMA to rank photos technically and aesthetically through various functions of the NIMA vector score (such as the mean), which can serve as a base for training a ML model to identify professional photographers in Flickr.

Model evaluation. The authors applied the previously explained deep learning NIMA model to solve the regression problem of predicting the technical quality and aesthetic attractiveness in Flickr that serve as features. The authors normalized the computed features between 0 and 1 in order not to have distorting differences in the ranges.

Preliminary Results and the Following Steps

In Figure 3, there are shown normalized distributions of the aesthetic and technical scores across the obtained data set. It can be seen that they have Gaussian shapes with a mean of 0.4 for the aesthetic score and 0.52 for the technical score. Figure 4 represents two examples of the photos that the NIMA model ranked as most and least aesthetically attractive with scores of 0.84 and 0.12 points, respectively. Even though many photos get an average score, these can serve as important features for training the model because there are still outliers.

Figure 3. The distributions of aesthetic (a) and technical (b) NIMA scores across the data set of photos



(a) (b)

Figure 4. The examples of high (a) and low (b) ranked aesthetic scores by the NIMA model: (a) nima_score = 0.84 (picture taken from the user manu.palix[26] under CC BY-NC-SA 2.0 license), (b) nima_score = 0.12 (picture taken from the user Antony Petrushko[27] under CC BY-NC 2.0 license)

(a) (b)

On these grounds, NIMA can infer aesthetical and technical scores of photographies, meaning that it can recognize users capable of taking photos that a typical user would like. Moreover, the model can be applied to the entire portfolio of users' photographs. In this way, the authors believe that this exploratory

analysis has proved that Flickr online photos can serve as a base to measure the photography capabilities of users. Additionally, computing the following features with Flickr API[28] would be beneficial to train the final ML model for evaluating photography quality:

- **Photography features.** These are the ones that can be extracted from the photos. This group of features includes the output of the NIMA model (aesthetic and technical scores), Exchangeable Image File Format (EXIF) data which is specific information stored as part of the file when a photo is captured using a digital camera such as camera settings and adjustments, date/time of capture or copyright ownership.
- **Crowdsourced features.** These involve information or opinions from a group of people who submit their data via the Internet. These features include the number of users who added the photo of the user to their favorites, the number of views the photo had; the number of the comments which were written, the number of groups in which the photo appeared, the difference between the upload date and the last update of the metadata.
- **User features.** These aim to gauge users' activity level who wrote the initial comment on Reddit. This group of features includes occupation, the number of following users, join date, a profile description, and the number of groups of which the user is a member.

The authors are of the opinion that the initial investigation already proves the potential that photo and video sharing platforms have to support finding professional photographers. It can be beneficial to understand the proportions and distinguish photography experts and amateurs in order to maintain well the groups of interest and reduce the amount of work of group administrators approving pictures in moderated groups. Moreover, it can be possible to detect sensitive and inappropriate content.

DISCUSSION

The case studies presented have proved that digital media portals hold evidence of users' competencies and capabilities that are important in the 21st century. In this section, the researchers will question the underlying rationale why these platforms generate support for providing informal learning opportunities and enhancing the educational process. They will also explore different applications of the presented case studies.

Where Can Informal Learning Happen?

Nowadays, Q&A websites and SNSs have gained much interest among people due to their support in solving different kinds of problems. Over the recent years, swift growth in the number of users of these networks has been tracked. Moreover, with the growth of information availability, Q&A websites provide users with a valuable platform for information sharing and searching where they can contribute and interact by posting questions and answers, commenting, and voting. On the other hand, a large part of the population uses SNSs, across which behaviors, motivations, and user profiles differ (Hellemans et al., 2020). These sites attract diverse groups of people who want to belong to a specific group, consume, and share information. The authors believe that precisely these types of platforms, Q&A portals, and SNSs, can provide opportunities for informal learning to happen. The authors elaborate on the potential reasons next.

First of all, both Q&A websites and SNSs are widely used by many people every day. It is a reality of the 21st century that most people have become very dependent on their smartphones. Indeed, this fact can be seen negatively in certain aspects. However, here the authors want to focus on the potential affordances that this online activity can have, for example, in the context of learning and education. Secondly, the users are creating and consuming a lot of content. Evidently, it is the responsibility of every individual what content to get and produce, therefore requiring to filter out low-quality content or misinformation appropriately. However, there is no doubt that recently, many forums emerged that could be beneficial for learning, and the number of people that learn online keeps growing and growing. Finally, it is also

important to mention that Q&A portals and SNSs keep their users motivated to maintain an engaging atmosphere. This fact can be decisive for choosing informal learning opportunities.

Practical Applications

The authors are of the opinion that the case studies that they exemplified in this chapter can have various applications across several domains.

Firstly, the results presented above can be used as part of the hiring processes of companies. It is not a secret to anybody that from a long time ago, recruiters have been utilizing information online about the candidate as part of the screening process, not only by monitoring the behavior of the potential candidates across professional social networks like LinkedIn but also in other SNSs. There are several motivating reasons why employment services prefer online screening methods over traditional methods. Firstly, nowadays, it seems like a natural impulse to investigate more about the prospective candidates' online identity insights that could reflect some work qualities not presented in the resume. Moreover, online monitoring is cost-effective, does not require much time, and foresees easy data processing (Quarterly, 2013). That being in the case, a company could benefit from employing an expert who has proved to be active online in career-related portals by showing interest in getting or providing feedback to colleagues or displaying other necessary dynamic capabilities that exhibit the person's potential. Above all, socio-technological environments are able to contextualize the capabilities and efforts of users empowering prospective employers to avail themselves of vast amounts of intentionally and inadvertently disclosed applicant information (Karriker, J. H., & Hartman, 2018). However, job hunters should not be deceptive in advancing their social media employment profiles; instead, they should build their resumes matching the job offers and be cognizant of the influences of their contextual information reflecting their potential (Karriker, J. H., & Hartman, 2018).

On the other hand, the performed case studies can be used as a part of formative assessment, providing ongoing feedback to teachers to improve their teaching and to students to boost their self-awareness. This can help enhance their learning process by recognizing their strengths and weaknesses in order for learners to put the focus on areas that require more consideration and for instructors to identify the tangled topics and address the problems immediately. Formative assessment typically involves qualitative feedback focused on the details of the students' work, performance, and actions under students' control, rather than on the students themselves. It is vital that high-quality, consistent, and timely feedback is sufficient in frequency and detail in order to provide the foundations for learner autonomy and a framework for high achievement.

Content sharing and consumption portals facilitate a kind of collaborative learning and peer feedback where equal status users, namely peers, provide opinions and critiques on other users' submissions by giving upvotes and writing comments. Peer feedback has proved to be beneficial for student learning, including enhanced knowledge of the subject matter, feedback from a range of sources, constructive reflection, attention to detail, critical analysis, critical thinking, and improved quality of work area (C. Jacoby et al., 2014). To this extent, this kind of feedback can provide users with more informal learning opportunities.

Finally, the results that the researchers provided can contribute to the literature on self-regulated learning, influencing self-efficacy, motivation, and effort towards academic success. It is crucial because self-regulated learners are able to meet the demands of many situations, including academic tasks, by actively controlling their active, goal-directed, self-controlled behavior, motivation, affect, time, and cognition. Self-regulated learners view the acquisition of skills as a systematic and controllable process, and they take greater responsibility for their achievement outcomes. Self-regulation reflects the idea of being able to develop knowledge, skills, and attitudes that can be transferred from one learning context to another

(Pintrich, 1995). Along these lines, self-regulated learners solve educational tasks with confidence, diligence, and resourcefulness; moreover, they are aware when they know a fact or possess a skill and when they do not. On top of that, when they encounter obstacles such as poor study conditions, confusing teachers, or abstruse textbooks, they find a way to succeed (Zimmerman, 1990). The good side is that students can learn to self-regulate regardless of age, gender, ethnic background, actual ability level, prior knowledge, or motivation. Therefore, the case studies presented in this chapter hold the potential to be used for developing self-regulation.

CONCLUSION

Social media and content sharing and consumption portals are now so integrated into people's daily life that many possibilities have emerged for online learning across them. In this chapter, the authors discussed the characteristics of these online platforms and the reasons why informal learning is happening precisely across them. Moreover, the researchers presented two case studies, one focused on detecting data science experts across the Reddit community by examining comments of the respective subreddit thread and another one measuring the photography capabilities of Flickr users. The presented results proved that it is possible to identify experts in specific fields by analyzing users' open online traces. Moreover, these approaches can serve as a base to replicate these methodologies to infer other competencies and capabilities across online portals.

The authors believe that there are several promising applications of the case studies presented in this chapter, including the implementation of these models as part of the hiring processes to improve candidate selection, their application in the educational context of formative assessment, and to improve the self-regulated learning capacities of students to obtain higher performance and motivation. Also, there is no doubt that the last years' experience has proved that there is a need to reform the formal education approaches. In this way, the official syllabus of the courses must include and adopt diverse teaching strategies, for example, by making use of various online teaching platforms, promoting informal and self-regulated learning. With this in mind, teachers, learners, and education stakeholders, in general, will be better prepared for the times of crisis and the need for new competencies required to function in the 21st century successfully.

ACKNOWLEDGMENT

Blinded for review.

REFERENCES

C. Jacoby, J., Heugh, S., Bax, C., & Branford-White, C. (2014). Enhancing learning through formative assessment. *Innovations in Education and Teaching International*, *51*(1), 72–83. https://doi.org/10.1080/14703297.2013.771970

Carter, R. A., Rice, M., Yang, S., & Jackson, H. A. (2020). Self-regulated learning in online learning environments: strategies for remote learning. *Information and Learning Science*, *121*(5–6), 311–319. https://doi.org/10.1108/ILS-04-2020-0114

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284. https://doi.org/10.1037/h0076540

Dhar, B. V. (2013). Dhar_Data_Science_Prediction. *Communications of the ACM*, *56*(12), 64–73.

Edgar, D., & Jeanne, S. C. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11-20+28. https://www.jstor.org/stable/1473669

Farrugia, L., Lauri, M. A., Borg, J., & O'Neill, B. (2019). Have You Asked for It? An Exploratory Study About Maltese Adolescents' Use of Ask.fm. *Journal of Adolescent Research*, *34*(6), 738–756. https://doi.org/10.1177/0743558418775365

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining. *Journal of the Learning Sciences*, *22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391

Gyongyi, Z., Koutrika, G., & Pedersen, J. (2008). Questioning yahoo! answers. *Www2008*. http://ilpubs.stanford.edu:8090/819

Hellemans, J., Willems, K., & Brengman, M. (2020). *Daily Active Users of Social Network Sites: Facebook, Twitter, and Instagram-Use Compared to General Social Network Site Use*. *1*, 194–202. https://doi.org/10.1007/978-3-030-47595-6_24

Karriker, J. H., & Hartman, N. S. (2018). Social Media and Dynamic Capabilities: Mining Millennial Resources. *Journal of Organizational Psychology; West Palm Beach Tomo 18, N.º 4*, 43–56.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84–90.

Lu, X., Lin, Z., Yang, J., & Wang, J. Z. (2014). RAPID : Rating Pictorial Aesthetics using Deep Learning ∗ Categories and Subject Descriptors. *ACM Multimedia*, 457–466.

McCarthy, J. (2017). Enhancing feedback in higher education: Students' attitudes towards online and in-class formative assessment feedback models. *Active Learning in Higher Education*, *18*(2), 127–141. https://doi.org/10.1177/1469787417707615

Ovadia, S. (2011). Internet connection quora.com: Another place for users to ask questions. *Behavioral and Social Sciences Librarian*, *30*(3), 176–180. https://doi.org/10.1080/01639269.2011.591279

Packiam, T., & Geoffrey, R. (2012). The impact of engagement with social networking sites (SNSs) on cognitive skills. *Computers in Human Behavior*, *28*(5), 1748–1754. https://doi.org/10.1016/j.chb.2012.04.015

Pal, A., Herdagdelen, A., Chatterji, S., Taank, S., & Chakrabarti, D. (2016). Discovery of Topical Authorities in Instagram Categories and Subject Descriptors. *Www*, 1203–1213.

Pintrich, P. R. (1995). Understanding Self-Regulated Learning What Is Self-Regulated Learning? *New Directions for Teaching and Learning*, *63*, 3–12.

Quarterly, B. S. (2013). Social Networking Sites and Hiring: How Social Media Profiles

Influence Hiring Decisions Sara Sameen (Assistant Professor Human Resource, FAST-NUCES) and Samia Cornelius (Lecturer Marketing, FAST-NUCES). *Journal of Business Studies Quarterly*, *7*(1), 27–35.

Rahayuningsih, H., & Yuniarti, K. W. (2017). Balancing Development and Sustainability in Tourism Destinations. *Balancing Development and Sustainability in Tourism Destinations*. https://doi.org/10.1007/978-981-10-1718-6

Raj, N., Dey, L., & Gaonkar, B. (2011). Expertise prediction for social network platforms to encourage knowledge sharing. *Proceedings - 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011*, *1*(September), 380–383. https://doi.org/10.1109/WI-IAT.2011.93

Saltz, J. S., Shamshurin, I., & Crowston, K. (2017). Comparing data science project management methodologies via a controlled experiment. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2017-Janua*, 1013–1022. https://doi.org/10.24251/hicss.2017.120

Shih, R. (2012). Can Web 2.0 technology assist college students inlearning English writing? Integrating Facebookand peer assessment with blended learning. *Australasian Journal OfEducational Technology*, *37*(2), 36–39.

Spache, G. (1953). A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal*, *53*(7), 410–413. https://doi.org/10.1086/458513

Strukova, S., Ruipérez-Valiente, J. A., & Mármol, F. G. (2022). A Survey on Data-Driven Evaluation of Competencies and Capabilities across Multimedia Environments (in press). *International Journal of Interactive Multimedia & Artificial Intelligence*.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899

Ulla, M. B., & Perales, W. F. (2021). Facebook as an integrated online learning support application during the COVID19 pandemic: Thai university students' experiences and perspectives. *Heliyon*, *7*(11), e08317. https://doi.org/10.1016/j.heliyon.2021.e08317

UNESCO. (2021). *Education: From disruption to recovery*. https://en.unesco.org/covid19/educationresponse

Vesselinov, R., & Grego, J. (2012). Duolingo Effectiveness Study. *City University of New York, USA*, *28*(December 2012), 1–25.

Wang, G., Gill, K., Mohanlal, M., Zheng, H., & Zhao, B. Y. (2013). *Wisdom in the social crowd*. *August 2015*, 1341–1352. https://doi.org/10.1145/2488388.2488506

Yan, X., Yang, J., Obukhov, M., Zhu, L., Bai, J., Wu, S., & He, Q. (2019). Social skill validation at LinkedIn. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2943–2951. https://doi.org/10.1145/3292500.3330752

Zhao, Q. (2017). Summary of Commercial Photography Skills in the Context of Visual Communication. *DEStech Transactions on Social Science, Education and Human Science*, *ssme*, 102–107. https://doi.org/10.12783/dtssehs/ssme2017/12939

Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, *25*(1), 3–17. https://doi.org/10.1207/s15326985ep2501_2

ADDITIONAL READING

Durães, D., Toala, R., Gonçalves, F., & Novais, P. (2019). Intelligent tutoring system to improve learning outcomes. AI Communications, 32(3), 161-174.

Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS), 6(4), 1-19.

Patil, S., & Lee, K. (2016). Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors. Social network analysis and mining, 6(1), 5.

Pham, P., & Wang, J. (2015, June). AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In International conference on artificial intelligence in education (pp. 367-376). Springer, Cham.

Raikos, A., & Waidyasekara, P. (2014). How useful is YouTube in learning heart anatomy?. Anatomical sciences education, 7(1), 12-18.

Strukova, S., Ruipérez-Valiente, J. A., & Mármol, F. G. (2021, September). Data-Driven Performance Prediction in a Geometry Game Environment. In Proceedings of the Conference on Information Technology for Social Good (pp. 283-288).

Wong, M. M., & Csikszentmihalyi, M. (1991). Motivation and academic achievement: The effects of personality traits and the duality of experience. Journal of Personality, 59(3), 539-574.

Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., & Ogata, H. (2015, November). Identifying and analyzing the learning behaviors of students using e-books. In Proceedings of the 23rd International Conference on Computers in Education (pp. 118-120). Hangzhou, China: Asia-Pacific Society for Computers in Education.

KEY TERMS AND DEFINITIONS

Application Programming Interface (API): A set of functions that allows building and integrating applications' software.

Data Science: An interdisciplinary field whose objective is to extract and interpret knowledge and insights using scientific methods, processes, algorithms, and systems.

Formative Assessment: A wide variety of methods that teachers use to evaluate students' comprehension, learning needs, and academic progress during an educational process.

Informal Learning: Learning happening outside of a structured, formal classroom environment.

Information Retrieval: An automatic process, methods, and procedures of searching and obtaining data that are relevant to an information need.

Machine Learning (ML): A branch of computer science that uses data and specific algorithms to imitate how humans think and learn.

Self-regulated Learning: the ability of individuals to understand and control their learning environment.

ENDNOTES

[1] https://moodle.org/
[2] https://www.sakailms.org/
[3] https://www.facebook.com/
[4] https://www.flickr.com/
[5] https://www.reddit.com/
[6] https://www.pinterest.com/
[7] https://www.youtube.com/
[8] https://www.tripadvisor.com/ForumHome/
[9] https://www.ft.com/
[10] https://twitter.com/
[11] https://www.instagram.com/
[12] https://www.wikipedia.org/
[13] https://www.duolingo.com/
[14] https://www.linkedin.com/
[15] https://www.answers.com/
[16] https://ask.fm/
[17] https://www.quora.com/
[18] https://stackoverflow.com/
[19] https://answers.yahoo.com/
[20] https://help.yahoo.com/kb/SLN35642.html
[21] https://www.reddit.com/dev/api/
[22] https://github.com/praw-dev/praw/
[23] https://500px.com/
[24] https://1x.com/

[25] https://www.smugmug.com/
[26] https://www.flickr.com/photos/194391943@N08/51789666311
[27] https://www.flickr.com/photos/99614740@N07/20094364928
[28] https://www.flickr.com/services/api/