Predicting digital engagement on Instagram

Lorenzo Vecchi
Pontifical Catholic University of Paraná, Curitiba Campus (PUCPR)
Eliane Francisco-Maffezzolli
Pontifical Catholic University of Paraná, Curitiba Campus (PUCPR)

Cite as:
Vecchi Lorenzo, Francisco-Maffezzolli Eliane (2020), Predicting digital engagement on Instagram. *Proceedings of the European Marketing Academy*, 49th, (63299)

Paper from the 49th Annual EMAC Conference, Budapest, May 26-29, 2020.

Predicting digital engagement on Instagram

This work aimed to develop a method to predict the level of engagement of HEIs on Instagram with artificial intelligence. The final product, however, in addition to foresight, permeated ways to classify Instagram posts, analyze the relevance of using people for such feature extraction, and description of the patterns found by the program, due to the characteristic of being an interpretable machine-learning algorithm. The database used was collected from public images provided by one private Instagram HEI. The work has an exploratory character, despite having a step that aims to determine a possible causal relationship between the independent and dependent variables, whose final premise was the description and understanding of the patterns found. The results demonstrate the contribution that variables extracted by human perception may have, in addition to the different patterns learned for different engagement metrics.

Keywords: *Engagement prediction; Artificial intelligence; Instagram.*

Track: *Digital Marketing & Social Media*

1. Introduction

Research shows that we are living an increasingly connected reality, and each year the number of active users on the Internet keeps growing. According to We Are Social (2018), at the time of the survey, of the 7.59 billion people in the world, about 3.19 billion were users of some kind of social media. The attempt to attract consumers' attention in this scenario is such, especially with regard to social media, that they end up filtering what really interests them, due to the amount of information they are daily fed into (Lee, Hosanagar and Nair, 2018). The big issue, understood by this scenario, was how to predict the level of engagement in Instagram HEI posts, based on historical performance, given the relevance that knowledge of this data, and the types of posts that most tend to engage organically , would have for these companies. Understanding the user preferences and creating custom content for the audience is an essential function of the marketer nowadays, therefore, the main goal of this paper is to develop a method to predict the users' level of engagement considering the post content on pages of HEIs in Instagram.

2. Theoretical Framework

2.1 Digital Engagement

Because most corporate digital media efforts are business-critical, they need to have well-defined expectations for expected results. One of the companies' goals is to understand what makes customers even more willing to buy (Barger, 2013), therefore, one can understand the relevance of using traditional metrics: number of likes, comments and shares as a way to understand reactions and public tastes in social media (Poecze, Ebster, Strauss, 2018).

Given the existence of various forms and metrics of engagement, Visser and Richardson (2013) described a model of engagement phases. The first phase has its focus on reaching people. With some effort, one rises to the phase of interest, being exemplified by the like metric. Based on their interest, one can involve them, resulting, for example, in the comment metric. The audience involved is finally close to being activated to help create more value for the brand, being quantified, for example, by the sharing metric.

The engagement metrics present on Instagram for public consultation were the number of likes and number of comments, referring to the phases of interest and involvement, respectively. The two metrics were used as dependent variables of this work. Importantly,

although the semantics of the comments can also be analyzed, only their numerical form was used.

2.2 Persuasion Elements in Posts

To have a better understanding of Instagram posts as a whole, in order to extract structured information from them, the posts were analyzed in various ways, indenting to extract info from different sources of each post.

For the task of analyzing different points of information in each post, this paper selected three distinct extraction lines, considering explicit objects found in the images, color-related calculations and the strategic purpose intended for the post. This session will present the three papers that served as the basis for the independent variables used to classify posts during the paper. The reason for using these three works in collaboration is that, in addition to the academic relevance of all being extracted from international newspapers and fairs, each has the ability to extract a different characteristic from the posts to be analyzed, thus increasing the structured understanding of them, both in the general strategic aspect and in the explicit visual aspect.

The research that underpinned the independent variables that contributed to extracting the strategic purpose of the post was: the framework for categorizing social media posts. These categories were created around the theoretical framework on types of posts, along with an empirical analysis of which categories are most frequent and useful for a social media post classification process. Possible analogous descriptions of post types, derived from the theoretical framework and empirical test, were concatenated. These variables are important because they add a deeper level of interpretability, and, because of this, were extracted by human evaluators, given the level of subjectivity that their analysis requires.

Box 1 shows each of the twenty-three independent variables used by this study, along with the studies on which they were based. In addition to the variables, there is also the manner in which they were extracted, number of variables used from each work, and focus of extraction from each study.

	Reference works		
	A First Analysis of Instagram Photo Content and User Types	Ranking and Classifying Attractiveness of Photos in Folksonomies	A framework for categorizing social media posts
variable focus	explicit content (obj. detection)	color characteristics	strategic purpose
number of variables used	8	3	12
abstract	This work segmented the visual content of images, resulting in variables that address the explicit content of photos	Here the attractiveness is related to the human vision system, and therefore the variables were related to color	This paper has its variables built around the posting strategy, thus providing insight into the purpose of the posting
independent variables	Bench, Car, Person, People (+1), Dog, Umbrella, Plant Pot, Tie	Saturation, Saturation Variation, RGB Contrast	Emotional, Functional, Educational, Resonance, Experience, Event, Personal, Employees, Community, Relationship, Cause, Promotion
adequacy	Limitation of the technology used	Suggestion from the study itself	-
mode of extraction	automatic		human evaluation

Box 1 – Description of the independent variables used by this study

The adequacy or filtering process of the independent variables from the base works was for the following reasons: the work has already suggested which variables would tend to help the prediction process more, or by some limitation of the automated detection technologies used by this study.

It was noticed, from the description of the work: A framework for categorizing social media posts, that the variables: emotional, experience and personal do not have a punctuality as clear as the objectivity required for its use, having in its description words such as: evoke and awaken, thus reinforcing its subjective character. This will imply the availability of different scales for participants to evaluate these variables, a five-point Likert scale for the most subjective ones and a dichotomous for the more specifically described ones.

2.3 Artificial Intelligence

As stated earlier, machine learning has been used in the present work, and within the algorithms that fall into this branch, the decision tree is one of the simplest and most successful forms. The decision tree, more specifically, represents a function that takes as input a vector with values and returns a decision. Input and output values can be discrete or continuous (Russell; Norvig, 2009). In most decision trees applied to classification, the notion of information gain, which is defined by the entropy calculation, is used to divide each tree node. To obtain continuous results in a regression format, the implementation of the trees

continues with the same logic of selecting criteria to section the branches, but instead of using entropy, that takes into account possible result classes, the mean squared error calculation is used.

3. Methodology

The methodology of this study was performed in three parts, being the first step to validate the independent variables, the second to calibrate the human evaluators and the third as a way to analyze the algorithm results. Two algorithms were created, one to predict the number of likes and the other to predict the number of comments. Both algorithms were created using the implementation of the Scikit Learn library (Pedregosa, 2012). It is important to highlight that although this work has a stage that aims to determine a causal relationship, the essence of it still remains exploratory.

The images used to extract the independent variables were collected by the web scarping technique. These images were collected between June 20, 2018 and August 14, 2019 on Instagram from a Brazilian higher education institution. The final image database contained 265 examples.

To perform a prior analysis of the independent variables in order to better delineate which ones are most relevant to finding causal relationships with the dependent variables, a focus group was conducted with a group of nineteen marketing analysts from the same university whose Instagram was analyzed.

As the independent variables related to the strategic objective of the positions were analyzed by human evaluators, at this stage the previous leveling of these people occurred. The calibration of the research instrument aimed to answer any questions of the three evaluators regarding the technical definitions of each strategic variable of the posts. This process took place in two steps to compare and measure performance improvement, each with three different posts selected for convenience.

Extractions related to object detection were performed by the YOLOv3 deep learning algorithm (Redmon, 2018). The color calculations used as independent variables were performed with the OpenCv computer vision librarian (Bradski, 2000).

4. Results

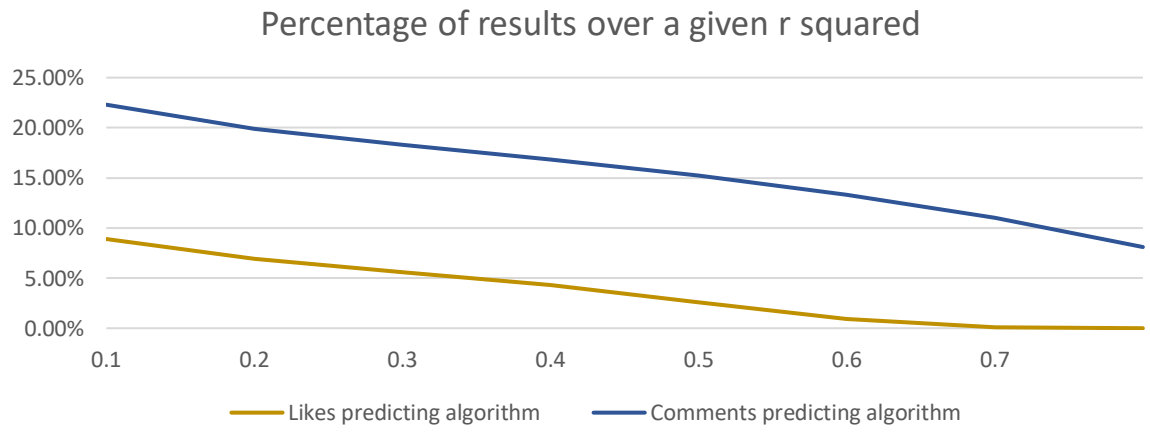
In the focus group analysis, which allowed the independent variables to be validated before being incorporated into the algorithm, it was concluded that the task of organizing posts characteristics without the aid of any automatic mechanism, or more detailed

specification of the cutout or target audience, is an extremely vague. In addition, the perceived difficulty and thoroughness in performing such a task reiterated the relevance and necessity of the type of analysis addressed in this paper.

Considering the calibration step, the agreement measurement of two participants, performed from the variables collected through the Likert scale, obtained an improvement of 14.29% compared to the first moment. Moreover, the measure of full agreement of the participants, made from the variables collected through the dichotomous responses, obtained an improvement of 18.75% compared to the first moment. This confirmed an improvement in the similarity of responses between participants, which enabled the final collection containing the 265 from the Instagram of a Brazilian higher education institution.

4.1 Algorithm Performance Tendency

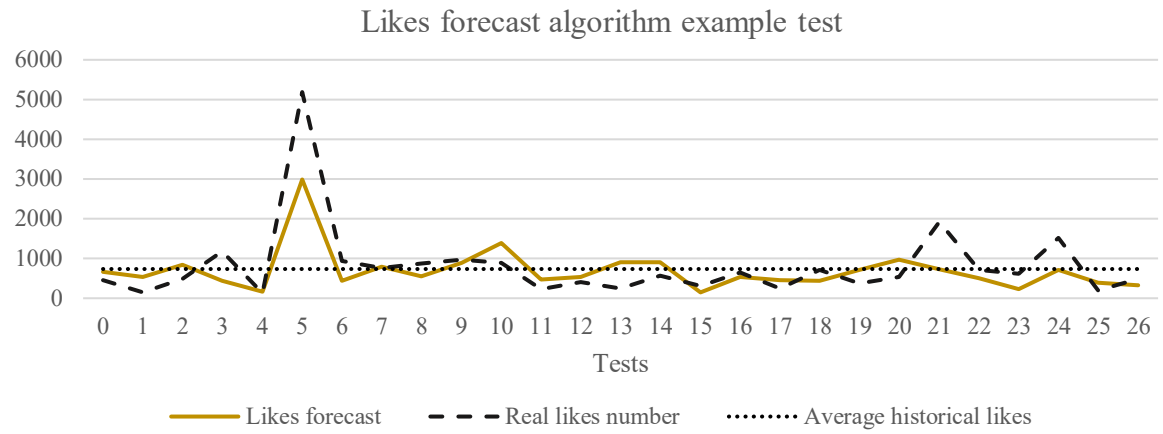
Decision trees are based on a heuristic algorithm and, therefore, their division rules for each node are based on local optimization, not guaranteeing a better overall error (Pedregosa, 2011). This randomness occurs so that the algorithm can find the best overall error eventually. So before producing the final tree for further analysis, fifty thousand tests were performed. This number was deliberately chosen to measure tree performance trends, which would not be possible with a single analysis. It is important to highlight that the training of the algorithms had as criterion of division of the nodes the mean squared error metric. The division between the training base and the test base changed randomly with each pass, but always followed the ratio of 9/1, respectively.



Graph 1 – Number of results (out of fifty thousand) that are above a given r squared

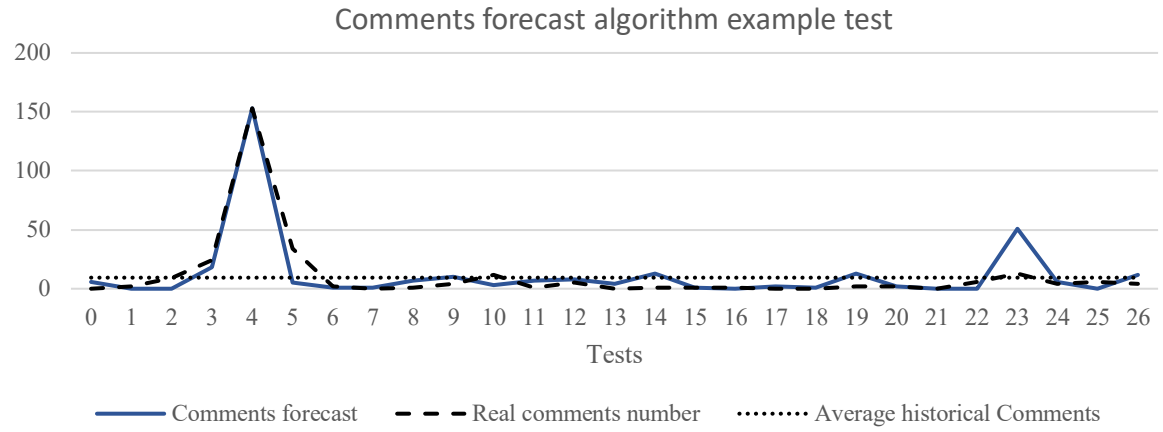
During the training of both trees to be analyzed, we selected the ones that would be in the best performance areas according to the tests, that is, for the tanking prediction algorithm an r ^ 2 above 0.6, and for the prediction algorithm. of comments, an r ^ 2 above 0.85.

4.2 Final Algorithms Analysis



Graph 2 – Likes prediction algorithm test

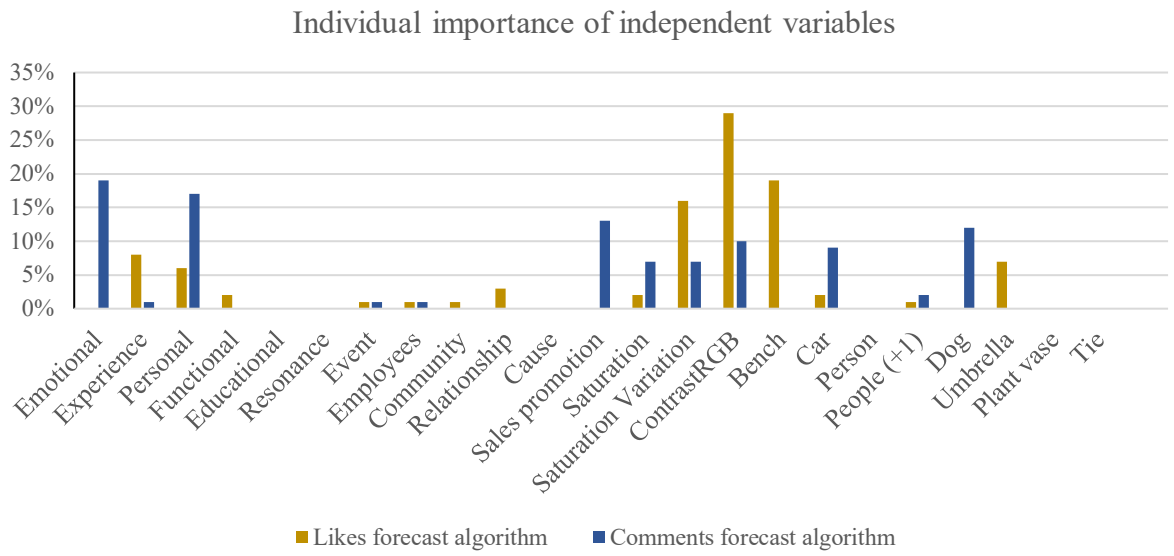
Percentage of times the algorithm indicates the correct direction in relation to the average likes was 70.37%. The r squared was equal to 0.7.



Graph 3 – Comments prediction algorithm test

Percentage of times the algorithm indicates the correct direction in relation to the average comments was 77.78%. The r squared was equal to 0.87.

In addition to the tests with some examples, both algorithms were tested using cross-validation, maintaining the same performance as in the test blocks presented in Graph 1 and Graph 2. A final analysis was performed to describe the most important independent variables for the decision process.



Graph 4 – Individual importance of independent variables

The general insight provided by the individual analysis of the independent variables is that the prediction of likes has a greater dependence, considering the explanation of their oscillation, in visual elements of the image, while the prediction of comments, of strategic elements. As described by Visser and Richardson (2013), the likes are an initial level of consumer interaction with brands, being very pertinent because they have a higher relationship with elements of rapid assimilation and less personal depth. Comments require a higher level of involvement than likes, and it is very pertinent that they have a greater relationship with more complex assimilation elements, with greater personal significance.

Both algorithms generate visualizable trees, and whose first nodes, used by this study to generate managerial suggestions for marketing analysts, were as follows for the likes and comments prediction algorithms. The first node generated by the likes prediction tree used as a condition the presence or absence of a dog in the image. If so, the average likes the post could have is: 2070, otherwise: 716. It is important to note that the Brazilian higher education institution used has a dog as a mascot of certain actions, so the presence of other dogs in photos could generate other results. When it comes to the comment prediction tree, its initial condition for testing was whether or not to qualify as a sale promotion. If this characteristic were verified, the average comment could be 54 if not found: 8.

5. Conclusion

The aim of this study was to develop a method to predict the level of user engagement in image content on Instagram HEI pages. The process to achieve this goal contributed beyond predictability, adding knowledge about post-classification topics, capacity and limitations of using people as a source of structured information and approximation of technical knowledge of artificial intelligence to the artistic and subjective context contemplated in social media. The main addition brought by this work was the relevance of using humanly collected independent variables. These have proven to be a complex and interesting source for extracting information from more subjective contexts in order to have better decision-making capabilities in this competitive social media landscape. Finally, among the suggestions brought by this work, the gathering of more technical and interpretative information, brought by the human interpretation in this paper, has shown to have great potential to be taken into consideration in future research.

6. References

Barger, C. (2012). *The social media strategist: build a successful program from the inside out*. New York: McGraw Hill.

Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Dubras, R., Underwood, L., Shuman, C., & Lore Oxford. (2018, January 30). Digital in 2018: World's internet users pass the 4 billion mark. Retrieved May 20, 2019, from https://wearesocial.com/blog/2018/01/global-digital-report-2018.

Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What We Instagram: A First Analysis of Instagram Photo Content and User Types. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8118/8087

Lee, D., Hosanagar, K., & Nair, H. S. (2018). Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science*, *64*(11), 5105–5131. doi: 10.1287/mnsc.2017.2902

Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing Research an applied approach*. Harlow: Pearson Education Limited.

Pedregosa et al. (2012). Scikit-learn: Machine Learning in Python. *JMLR 12*, 2825–2830. Retrieved from https://scikit-learn.org/stable/

Pedro, J. S., & Siersdorfer, S. (2009). Ranking and classifying attractiveness of photos in folksonomies. *Proceedings of the 18th International Conference on World Wide Web - WWW 09*. doi: 10.1145/1526709.1526813

Poecze, F., Ebster, C., & Strauss, C. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Procedia Computer Science*, *130*, 660–666. doi: 10.1016/j.procs.2018.04.117

Redmon, Joseph, Farhadi, & Ali. (2018, April 8). YOLOv3: An Incremental Improvement. Retrieved June 15, 2019, from https://arxiv.org/abs/1804.02767.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: a modern approach*. Upper Saddle River: Prentice-Hall.

Tafesse, W., & Wien, A. (2017). A framework for categorizing social media posts. *Cogent Business & Management*, *4*(1). doi: 10.1080/23311975.2017.1284390

Visser, J., & Richardson, J. (2013). Create value with digital engagement. Retrieved from https://digitalengagementframework.com/.