**INTRODUCTION**

SOCIAL media are online communication channels dedicated to community-based input, interaction, contentsharing, and collaboration. These media give the users the opportunity to share their content such as text, video, and images. Users usually accompany the content they post with text such as comments or hashtags. This alternative text (comment, hashtags, etc.) provides valuable information about the user posts and other information. Preece et al. construct a Sentinel platform that can enhance social media data in order to understand different situations they based also in Youtube video comments. Sagduyu et al. present a novel system that can present large-scale synthetic data from social media. In their system, they use textual content (hashtags and hyperlinks in tweets) to produce topics and train the n-gram model. The users in several of those media, e.g. Twitter, Instagram, and Facebook, use hashtags to annotate the digital content they upload. Hahshtags are, usually, words or nonspaced phrases preceded by the symbol # that allow creators/content contributors to apply tagging that makes it easier for other users to locate their posts. A great portion of the digital content shared on social media platforms consists of images and short videos. Thus, effective retrieval of images from social media and the web, in general, becomes harder and more challenging day by day. Contemporary search engines are basically based on text descriptions to retrieve images; however, inaccurate text descriptions and the plethora of nontextually annotated images led to extended research for content-based image retrieval techniques.

The main problem of the content-based image retrieval is the so-called semantic gap: content-based retrieval is associated with low-level features while humans use high-level concepts for their search. To overcome this problem, automatic image annotation (AIA) methods were developed, that is, processes by which computing systems automatically assign metadata in the form of captions or keywords to images. Among the AIA methods, those based on the learning by example paradigm are probably the most common one. A small set of manually annotated training images are used to train models, which learn the correlation between image features and textual words (high-level concepts) and then allow automatic annotation of other (unseen) images. Obviously, good training examples, i.e., representative and accurate pairs of images and related tags are vital in this case. Social media, and especially the Instagram, provide a rich source of image–tag pairs. Mining the right ones, automatically or semiautomatically, so as to be used as training examples is extremely important.We have to consider, however, that, in many cases, hashtags that accompany images in social media are not related with the image’s content but serve several other purposes such as the expression of user’s emotional state, the increase in user’s clicks and findability, and the beginning of a new communication or discussion.

In our previous research, we have shown that the percentage of the Instagram hashtags that describe the visual content of the image they are associated with does not exceed 25% [12]. We have also noticed that many Instagram hashtags are used across images that have nothing in common, just for searchability enhancement. We named those hashtags as stop hashtags. Thus, filtering the Instagram hashtags in terms of the visual content of the image they accompany is required. Hyperlink-induced topic search (HITS) is a ranking algorithm than we could use to filter Instagram hashtags and locate the most relevant. The purpose of the HITS algorithm, developed by Jon Kleinberg, is to rate webpages. The basic idea is that a webpage can provide information about a topic and also relevant links for a topic. Thus, webpages belong to two groups: pages that provide good information about a topic (“authoritative”) and those that give to the user good links about a topic (“hubs”). The HITS algorithm gives to each webpage both a hub and an authoritative value. We have started experimenting with the HITS algorithm for mining informative Instagram hashtags in one of our previous works and we extend this paper here by considering the application of the HITS algorithm in a real crowdtagging environment facilitated by the Figure-eight, formerly known as Crowdflower, crowdsourcing platform. In addition, we have increased the number of annotations per image to 500, we formed the bipartite graphs for all images, and we calculated the performance of annotators across all those images. Moreover, FolkRank is used as a baseline to evaluate the performance of the proposed method.