HOSTED BY
ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jides

Evaluating the descriptive power of Instagram hashtags

*Stamatios Giannoulakis**, *Nicolas Tsapatsoulis*

Department of Communication and Internet Studies, Cyprus University of Technology, 30, Arch. Kyprianos str., CY-3036, Limassol, Cyprus

HIGHLIGHTS

- Instagram hashtags as annotation metadata is examined.
- Instagram photo-hastags as training sets for Automatic Image Annotation is proposed.
- Half of the chosen Instagram hashtags describe the visual content of an image.
- Instagram hashtags can be used as training examples for machine learning algorithms.

ARTICLE INFO

Article history:
Published online 11 November 2016

Keywords:
Instagram
Hashtags
Image tagging
Image retrieval
Machine learning

ABSTRACT

Image tagging is an essential step for developing Automatic Image Annotation (AIA) methods that are based on the learning by example paradigm. However, manual image annotation, even for creating training sets for machine learning algorithms, requires hard effort and contains human judgment errors and subjectivity. Thus, alternative ways for automatically creating training examples, i.e., pairs of images and tags, are pursued. In this work, we investigate whether tags accompanying photos in the Instagram can be considered as image annotation metadata. If such a claim is proved then Instagram could be used as a very rich, easy to collect automatically, source of training data for the development of AIA techniques. Our hypothesis is that Instagram hashtags, and especially those provided by the photo owner/creator, express more accurately the content of a photo compared to the tags assigned to a photo during explicit image annotation processes like crowdsourcing. In this context, we explore the descriptive power of hashtags by examining whether other users would use the same, with the owner, hashtags to annotate an image. For this purpose 1000 Instagram images were collected and one to four hashtags, considered as the most descriptive ones for the image in question, were chosen among the hashtags used by the photo owner. An online database was constructed to generate online questionnaires containing 20 images each, which were distributed to experiment participants so they can choose the best suitable hashtag for every image according to their interpretation. Results show that an average of 66% of the participants hashtag choices coincide with those suggested by the photo owners; thus, an initial evidence towards our hypothesis confirmation can be claimed.

Peer review under responsibility of Qassim University.

* *Corresponding author.*
E-mail addresses: s.giannoulakis@cut.ac.cy (S. Giannoulakis), nicolas.tsapatsoulis@cut.ac.cy (N. Tsapatsoulis).

http://dx.doi.org/10.1016/j.jides.2016.10.001

1. Introduction

On average 300 million photos are uploaded to Facebook per day [1] while an average of 80 million photos are shared every day in Instagram [2]. Locating and retrieving these and other images uploaded on the Web is very challenging not only in terms of effectiveness (retrieve the right image according to the user needs/queries) and efficiency (execution time) but also in terms of visibility (being locatable). Contemporary search engines retrieve images in a text-based manner since the majority of end users are familiar with text-based queries for retrieving web pages and digital documents. In text-based image retrieval images must be somehow related with specific keywords or textual description. This kind of textual description is, usually, obtained from the web page, or the document, containing the corresponding images and includes HTML alternative text, the file names of the images, captions, metadata tags and surrounding text [3,4]. However, images in social media, which constitute the great majority of Web images, cannot effectively indexed (extract relevant text description) with pure web-based techniques, mainly because the user pages in social media do not follow the classic web-page structure. As a result, the well-known content-based image retrieval field revitalized and a more specific research area, Automatic Image Annotation (AIA) [5] emerged. AIA refers to the process of extracting low-level features from an image and assigning one or more semantic concepts to it [6].

A large category of AIA involves machine learning techniques has its roots in the learning by example paradigm [7]. Training examples used for AIA are pairs of images and related tags. Many different models and machine learning techniques were developed to build the so-called 'visual models', that is, models that capture the correlation between image features and textual words from the training examples. Visual models are then fed with image features extracted from unseen images to predict their tagging [8]. Assuming that good visual models can be achieved, image retrieval using the training by example paradigm provides a promising alternative to text-based methods (since it does not require explicit annotation of all images in the collection, but only a small set of properly annotated images) [9]. Nevertheless, the first important step to create effective visual models is to use good training examples (pairs of images and annotations). In this context, automatic creation of training examples via crawling is highly desirable because it addresses the scalability (models for new concepts) and adaptability (modification of training models) issues.

According to a survey of Pew Research Internet Project,[1] the proportion of online American adults who use Instagram has doubled since 2012 showing the highest increase among all social media platforms [10]. Instagram is a free application for mobile devices, which offers a user the possibility to upload, edit and share with other Instagram users pictures and very short videos. The term Instagram is a combination of two words, from the word instant used to old market cameras and the gram comes from telegram from the snapshots people were taking.[2] Instagram launched on 6 October 2010 and rapidly gained popularity, managed to have 400 million active users on January 2016.[3] It is estimated that 80 million pictures are being shared per day [2] through Instagram.

In January 2011 Instagram added hashtags [11] and from 27 April 2015 users are able to use emoji as hashtags.[4] Hashtags are tags or words prepended with '#' used to indicate the content of the picture, allowing users to search for pictures and increase visibility. Photo owners sometimes want to connect pictures with emotions; in that case they use emoji which are pictograms that are connected with emotions.

Hashtags are not totally new in the web; users started to use them with IRC (Internet Relay Chat) in order to categorize items into groups. The first who used hashtags, in contemporary Social Media and especially inTwitter, was Chris Messina, a designer, who asked his followers how they felt about using the pound sign to group conversations [12]. Thus, a basic role of hashtags was traditionally to organize knowledge and facilitate access and enable retrieval of information (see also the work of Small [13] on this). Tapastreet, a search engine platform that offers users the opportunity to browse geo-located video and photos from social media such as Twitter, Facebook and Instagram, harvests location, time and hashtags [14] assuming that hashtags can be used in order to retrieve visual content. On the other hand, we know that users extend the function of hashtagging beyond findability and give hashtags a metacommunicative use. According to Daer et al. [15] the metacommunicative function can be split into four codes: 'emphasizing', 'iterating', 'critiquing', 'identifying', and 'rallying'. 'Emphasizing' is used to give emphasis or call attention; 'critiquing' expresses judgment or verdict; 'identifying' is used to refer to the author of the post; 'iterating' to expresses humor and 'rallying' brings awareness or support to a cause.

Several researchers suggest also that hashtags carry emotional information [16] which is not directly related with the context they appear [17]. In a research on the tags of a set of 2700 pictures, it was measured that approximately 10% of these photos were related with emotion words not directly related with their visual content [18]. A recent study, on gender difference in hashtag usage in Instagram for the hashtag 'Malaysianfood', revealed that women tend to use emotional hashtags while men hashtags are more informative [19]. Ferrara et al. [20] studied user behavior while they annotate their photos with hashtags. They found that users use quite a few hashtags in order to annotate an image.

It should be evident from the above that Instagram provides a rich forum for automatically creating training sets for AIA. It contains a huge amount of images which are commented through hashtags by their creators/owners and, despite that not all hashtags are actually related with the visual content of images, many of hashtags carry significant descriptive information of the visual content. Thus, if we assume that it is the owner who can better express the real visual content or meaning of an image then choosing among the hashtags for assigning tags to images is much safer than traditional text-based indexing approaches [21–23]. This is extremely important in training sets where pairs of images

[1] http://www.pewresearch.org/.
[2] Instagram: FAQ, https://instagram.com/about/faq/#.
[3] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.
[4] Instagram: Our Story, https://instagram.com/press/.

and tags have to be carefully selected because they affect the effectiveness of tag predicting models. However, Intsagram hashtags are used not only to describe the visual content of an image but also serve other functions falling under the metacommunicative use or expressing emotions. In this work, we are trying to check the extent to which hashtags are indeed related with the actual content of an image and the percentage of hashtags that are relevant to Instagram photo compared to those referring to metacommunicative use or carrying emotional information irrelevant to the visual content.

2. Related work

To the best of our knowledge, this is the first work that examines the appropriateness of Instagram photo-hashtag pairs for creating training sets for AIA. However, several approaches were proposed to (i) develop training datasets from the Web to be used for image classification/tagging [24], (ii) use the Flickr, a social network similar to Instagram, to construct image—tag pairs [25], (iii) get advantage of clickthrough data and search logs in search engines to form image-tag pairs [26], (iv) combining linguistic description with visual data in order to achieve automatic image annotation [27] and (v)investigate the quality of manual image annotation [28]. In the following, we examine the research in these areas in more detail.

2.1. *Developing image datasets by harvesting the Web*

The last decade research has moved towards automatically acquired (from the Web) data sources in order to be used for training AIA systems or concept detectors in general [29–31]. Such data sources include content that has been annotated by user-defined tags (e.g., Picasa, Flickr, Yahoo! Video, Youtube etc.) as well as images and videos annotated with keywords that have been automatically extracted from the surrounding text of the corresponding Web pages.

Schroff et al. [24] tried to automatically generate high-quality images for a specified object class. In order to achieve the aforementioned goal, they harvested images based on a text-based Web search on a specific object. Then they used a combination of text/metadata and visual features so to exclude irrelevant images and automatically rank the relevant ones.

Deng et al. [32] created one of the biggest image databases, ImageNet, a large-scale ontology of images. In order to collect the images the researchers submitted queries to several image search engines then selection of relevant images was achieved manually by humans who indexed images with the help of Amazon Mechanical Turk[5] a crowdsourcing Web service.

In an attempt to automate the image annotation process, NEIL (Never Ending Image Learner) [33], a computer program that aims to extract visual knowledge based on semi-supervised learning, collected, for each one of the concepts it models, images through Google Image Search and used them to construct the initial classifier. In the second step, NEIL, aims to extract concept relations while in the third step tries to find new instances from unlabeled data. The second and the third step are continuously repeated in order to improve the effectiveness of the initial classifier.

Do & Yanai [34] entered an automatic approach to build video datasets from the Web. They harvested videos and then segment them into shots; relative shots were grouped into clusters. Their goal was to identify shots to be used as training data for automatic detection of action concepts.

2.2. *Image tagging with the aid of Flickr*

According to Sigurbjörnsson & Zwol [25] research on Flickr about user annotation, users use only a few tags to annotate their photos and tend to annotate images according to their content. Ulges et al. [35] confirmed the results of Sigurbjörnsson & Zwol and proved also that users share, in the Web, images with specific structure and metadata.

Ntalianis et al. [36] developed a method for automatic annotation of image datasets based on implicit interaction and visual concept modeling using data collected from Flickr. They found that the manual annotation of Flickr is much more analytical and provides more keywords, compared to the typical usage of keywords by ordinary users in Web search environments. They also mention the difficulty to evaluate and weight the perception of users regarding the visual content of images they do not own.

Several approaches aiming at image clustering, making use of Flickr tags, were also explored. Cui et al. [37] combined tags and visual image features so to improve image clustering. Removal of irrelevant Flickr tags aiming at more effective image retrieval was proposed from Xia et al. [38]. Their approach is based on allocating content bi-layer clustering of similar images and dividing these images into groups. By grouping similar images based on the tags with a stronger relationship they could identify and remove irrelevant tags.

2.3. *Clickthrough approaches*

Joachims et al. [26] discovered that differences between implicit and explicit relevance judgments are not so far as they were thought to be. This innovative finding opened a new way, where implicit relevance judgments were considered as training data for various machine learning-based improvements to information retrieval [39,40]. Clickthrough data is a form of implicit judgment easily collectable and its collection introduces no additional cognitive burden on users performing the queries. Thus, it is not a surprise that they were used as training data in various tasks including the works of [41,42], where a Latent Semantic Analysis (LSA) algorithm was applied to search logs in order to build a semantic space for indexing images.

Tsikrika et al. [43] examined the quality of clickthrough data for training concept detectors in images. They showed that clickthrough data, if properly filtered, would be used for AIA. The problem with clickthrough data is that they express the interpretation of end users rather than the creators/owners, and, thus, they are highly subjective.

[5] https://www.mturk.com/mturk/welcome.

Despite that, the use of clickthrough data for developing AIA models is an attractive approach and Microsoft Research announced, for a third year in a row, a challenge based on data provided by Bing search engine.[6]

Sarafis et al. [44], based on clickthrough data harvested from professional image search engines, proved that a Fuzzy Support Vector Machine (FSVM) approach and calculation of weights from language models can lead to significant improvement in image retrieval, compared to concept detectors based on standard SVM and other machine learning approaches. In a further investigation [45], they pointed out that clickthrough data are valuable in constructing concepts which can help to image retrieval, but label noise (irrelevant tags) is a problem in machine learning approaches. So they extended their approach for automatic concept detection by incorporating a filter for label noise handling.

2.4. *Visual and language data alignment techniques*

A lot of work was, recently, devoted on aligning visual and language data for image retrieval. Conventional approaches use natural language processing techniques to automatically extract image tags from surrounding text/image context. For instance, Tsapatsoulis [46] finds the text that surrounds an image with the aid of an HML parser and then identifies keywords from this text with the aid of a learning free language model. Although his method for keyword extraction seems fast and effective, extraction of image context (WICE) in generic web pages remains a challenge.

On a radically different view, Farhadi et al. [47] coined the challenge of generating sentences from images. They proposed an architecture that learns an intermediate meaning space to project image and sentence features that allow retrieving text from images and vice versa. Based on this idea, Kiros et al. [48] built a log-bilinear model that generates phrase description from images. They introduced multimodal neural language models that can be conditioned on other modalities, such as the visual modality. Their model can be used to retrieve images given complex description queries, retrieve phrase descriptions given image queries, as well as generate text conditioned on images. They showed that, in the case of image-text modelling, joint learning of word representations and image features is feasible by training the models together with a convolutional network. According to the authors, their approach can generate sentence descriptions for images without the use of templates, structured prediction, and/or syntactic trees. Compared to our work, the work Kiros et al.. can be used to give us another, more objective, feedback regarding the appropriateness of Instagram hashtags as image annotation metadata for AIA purposes.

Karpathy & Fei-Fei [27], in a notable work, developed further the idea of Kiros et al. [48], and strived to generate dense descriptions of images (i.e., descriptions per image region) by designing a model that is rich enough to simultaneously reason about contents of images and their representation in the domain of natural language. Their model is based on a combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Their retrieval experiments were conducted in Flickr8K [49], Flickr30K [50] and MSCOCO datasets [51]. One of the practical challenges, the authors state, is that the available implicit descriptions of image datasets on the Internet multiplex mentions of several entities whose locations in the images are unknown. Although in our work we do not refer to image regions, rather we consider an image as a whole, we investigate an even more primary problem: which of these descriptions, in the case of Instagram the hashtags, are related with the image content.

Johnson et al. [52] introduced the dense captioning task, a computer vision system that both localises and describe salient regions in images in natural language. The dense captioning task generalises object detection based on single word descriptions, and Image Captioning when one predicted region covers the full image. The authors propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image in a single forward pass without requiring external regions proposals. The label sequences are generated through a Recurrent Neural Network language model. Our work can inform techniques like that of Johnson et al. [52]. Instagram hashtags are basically single word descriptions; thus, selecting the ones that really describe the image content could eliminate outlier training points and would lead to faster convergence and more effective image region annotation. In a recent work [53], we have tried to setup a practical framework for filtering out irrelevant, to the image content, hashtags so as to fulfil the previously mentioned task.

2.5. *Quality of manual image annotation*

Several approaches deal with the quality of manual image annotation, especially under a crowdsourcing setting. Nowak & Ruger [54] investigated the reliability of image annotation via crowdsourcing. They tried first to explore to which extent several sets of expert annotations differ from each other and then to investigate whether non-expert annotations are reliable. Their dataset consists of 99 images selected from the MIR Flickr Image Dataset and was annotated by 11 expert annotators from the Fraunhofer IDMT research staff using 53 concepts. The same set of images was distributed over the online marketplace Amazon Mechanical Turk in order get non-expert annotations. The consistency among expert annotators proved to be very high. The same also proved between the expert and non-expert groups. Thus, the conclusion was that crowdsourcing annotation is as accurate as experts' annotation.

Wang and Zhou, on an analysis about the crowdsourcing label quality, argue that crowdsourcing data improve the quality of image annotation and the error rate decreases as a function of the number of people selected for annotation [28]. In order to examine the image retrieval from social media and especially the diversification of image retrieval results, Ionescu et al. [55] compared experts and crowdsourcing annotation. The results showed that in the

[6] http://research.microsoft.com/en-us/projects/irc/.

crowdsourcing annotation the inter-rater agreement was a slightly lower than expert annotators. Veloso et al. [56] designed an algorithm aimed to automatically annotate clothes in photos users upload in social media such as Facebook and Instagram. They observed that user comments accompanying images in these media contain similar terms, depicting common garment items. As a part of their research regarding diversification of image retrieval results in the environment of social media, they examined the differences between expert and non-expert annotators. They found that expert annotators perform a bit better than non-experts for the aforementioned classification task. Comparison between expert annotation and crowdsourced annotation was also examined in the framework of automatic genre identification. Asheghi et al. [57] proposed crowdsourced annotation as a way to produce reliable web genre corpus with high interannotator consistency. For this purpose they used crowdsourcing and they calculated an agreement between annotators reaching 88.2%. However, annotation was performed on a web page level and not on photos. Nevertheless, this work provides another indication showing that crowdsourcing annotations can be used as a replacement of expert annotation in image tagging. Crowdsourcing annotation was also used for video annotation. In an investigation regarding the accuracy of crowdsourced video labeling, Di Salvo et al. [58], found that the aforementioned annotation method generates reliable results.

Since crowdsourcing annotation is far more cheaper and efficient than experts' annotation the conclusions of the works described earlier opened up new ways in application requiring training corpora, and towards AIA as well [55].

The importance of crowdsourcing annotation leads to several research efforts which further examine the quality of crowdsourced data. In crowdsourcing annotation, the participants expose different behavior during the annotation task. There are many reasons for the aforementioned behavior including the level of expertise, low-attention/low-concentration when they perform the task and there is always the bad intent of the annotators. Annotators with bad intention might be spammers, dishonest users or users trying to manipulate the system by answering in an unrelated or nonsense way [59]. In a research about crowdsourcing annotators' consistency Theodosiou et al. [60] used both vocabulary keywords and free keywords to check whether guided annotation (as assumed by the use of structured vocabulary) would increase annotation consistency. The researchers concluded that, indeed, by combing free keywords and vocabulary keywords annotation consistency increases compared to the use of free keywords alone. Baba & Kashima [61] suggested a two-stage procedure in order to evaluate the quality of crowdsourcing work. In the first stage, the crowd performs the annotation and next the results are reviewed. In order to control the quality of annotations, unsupervised statistical methods are involved including a parameter accounting for the reviewers' bias. Li et al. [62] developed a framework, called Requallo, in order to keep a balance between quality and quantity of annotated data. They aimed to optimize the 'value for money' of annotation tasks in commercial crowdsourcing platforms given a limited budget. They use annotators consistency, named as 'confidence', as a measurement of quality; thus, annotation results having high quality are those with high confidence. Hu et al. [63] tried to overcome the problem of low-quality annotations in crowdsourcing services by introducing a model which combines expert annotation with crowd annotation. They managed to achieve better performance in crowdsourcing learning tasks with the least possible number of expert labels.

This paper extends our previous work [64] by increasing both the number of images used for annotation, from 30 to 1000, as well as the number of participants, from 39 to 362, in order to verify the validity of conclusions drawn from that study. In addition, and in order to generate an online questionnaire for our research with a random selection of the image subset presented to each participant, we redesigned our database schema (see Fig. 1) and implement it in MySQL. Also, the results are automatically analysed with the aid of PHP code and presented online in a web page.[7] Because, a few of the participants in the previous study were also participated in the current study we chose to use a totally different image dataset; thus, none of the 30 images of the previous study was included in the new image dataset.

As in [64] the purpose of the current study is to examine if participants would choose the owner hashtags to annotate the image rather than random hashtags. We assume that owners' annotation data (in our case Instagram hashtags) are more close to experts' annotation compared to that of crowdsourcing since the latter expresses the end-users' perspective. Furthermore, web-crawled data are far more easier to collect than crowdsourcing ones. Among the web-crawled data, the ones collected from Instagram are much more accurate (in terms of descriptive value) compared to those used in traditional web-document indexing (keyword extraction from web-pages) while they are richer than those collected via clickthroughs or other forms of implicit judgement.

3. Methodology

In order to derive concrete results, in our study we followed a hybrid methodology combining a set up from social science research with a strict mathematical framework which is common in natural sciences. We decided to define clear research questions and properly select the participants of the experiment rather than randomly choosing among ordinary users of social media. We consider that in order to assess the descriptive value of Instagram hashtags of the photo owners/creators we need users that are familiar both with the social media and the use of metadata in digital content. Librarians would be ideal for this purpose. They use social networks daily and one of their main tasks is to organise knowledge and annotate electronic resources, so we can say they are, in some respect, experts in image annotation. Moreover, undergraduate and postgraduate university students are also good candidates for the population group because social media are highly popular

[7] http://cis.cut.ac.cy/nicolas.tsapatsoulis/aiai2015/code/user_annot.php.

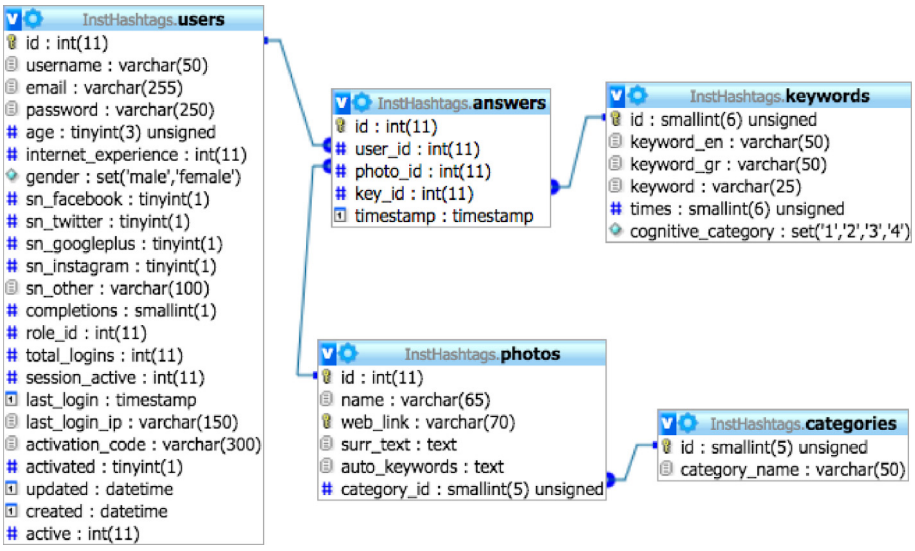

Fig. 1 – The database schema used to store annotation data and questionnaire results.

Table 1 – User demographics.

# participants	Female	Male	Average age (±Std)
295	227 (76.9%)	68 (23.1%)	33.2 ± 11.2

Table 2 – Social media usage of users participated in this study.

Internet exp. (years)	Facebook	Twitter	Google+	Instagram	Other
13.6 ± 5.8	81.7%	37.3%	35.3%	32.2%	22%

among students as we can conclude from the survey of Pew Research Internet Project [10]. Details of users' demographics and social media usage are shown in Tables 1 and 2.

3.1. *Aim of the research*

The present study has two goals. The first is to investigate whether Instagram hashtags accompanying images can be used as image tags so as to create image-tag pairs for training machine learning approaches for AIA. The second is to provide a rough estimation on the percentage of Instagram hashtags that describe the visual content of accompanying images. Towards this end, we had to select pictures from Instagram and to design an online questionnaire. We decided to create a set of 1000 images which were selected from 100 different subjects/hashtags (10 relevant images per subject/hashtag). Those images were uploaded to Instagram by 970 different Instagram users. Owners' hashtags surrounding these images were automatically crawled using the Beautiful Soup[8] library of Python. Then we chose, manually, 1–4 hashtags for each picture, which, according to our interpretation, better describe its visual content since, as mentioned previously, not all hashtags are intended to describe an image to increase its findability. Images, along with the corresponding hashtags and owner's nickname were stored in a database created using MySQL. The schema of this database is shown in Fig. 1. The aforementioned process of manually choosing images and manually entering the appropriate data in the online database took place between 2 June 2015 and 12 August 2015.

An online questionnaire (see http://cismir.ymdweb.com/) was designed based on the data stored in the database aiming to evaluate the descriptive power of chosen hashtags with respect to the corresponding images. Owner's hashtags along with irrelevant ones are presented beside each picture, and the participants are asked to choose among them the ones that better describe the shown photo. Fig. 2 presents an example. Among the eight choices given only two are hashtags the owner of the photo used in the Instagram. If participants' choices coincide with the hashtags the owner gave, we have a good indication that these hashtags are, indeed, related with the visual content of the picture (since what the participants see is the context-free picture without any sort of metadata).

3.2. *Data collection*

As mentioned before the data for this study gathered using an online questionnaire. The aim was to increase the number of participants, reduce the time required to fill in the questionnaire and avoid fatigue effects. For the latter, each participant was asked to 'annotate' only 20, randomly selected from the database, images in each session. However, users are allowed to repeat the process through another session as many times as they wish. Furthermore, with the online questionnaire we have the possibility to automate the result extraction process.[9] The choices given for each picture

[8] http://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[9] See http://cis.cut.ac.cy/~nicolas.tsapatsoulis/aiai2015/code/.

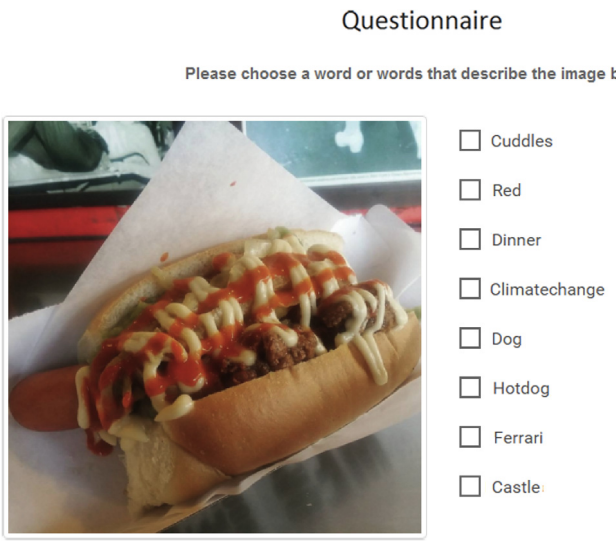

Fig. 2 – An example of an image interpretation multiple choice question.

are either four or eight depending on the number of hashtags the owner used. If only one hashtag of the owner was present then the choices given to the participants are four (including the hashtag of the owner); otherwise, the participants are given eight options to select from. This rule was applied in order to keep a minimum chance level higher than or equal to 25%. The 'wrong' hashtags are randomly selected among the hashtags given to other images stored in the database. In any case, participants are not aware that any of the given choices are related in any respect with the picture; thus, they are free to select as many of them as they wish according to their interpretation of the shown photo.

In order to reduce bad annotators we asked the participants to register, using username and password, so to complete the questionnaire. However, we have to note that participants had to provide a username and not their email, so that we can ensure the anonymity of the questionnaire. On a voluntarily basis users are also asked to fill in information about their age, gender, years of internet experience and social media usage (see Tables 1 and 2).[10]

Initially, four online questionnaires were distributed by electronic mail to four experts in order to evaluate them. The results of the evaluation assisted the creation of a more appropriate version, which was, then, distributed by electronic mail to librarians in Cyprus and Greece, to undergraduate and postgraduate students of the department of Communication & Internet Studies of the Cyprus University of Technology and to students of the Open University of Cyprus. A total of 362 users were registered; however, only 295 of them filled in the questionnaire at least once. 349 questionnaires were collected since some of the users took more than one session.

Students is one of the most active group in visual based social media such as Instagram and Facebook. Thus, they are familiar with both online images and their implicit annotation through comments, legends and hashtags. Librarians, on the other hand, provide a more professional view of image annotation. Dealing with image (and multimedia in general) tagging and social network media is a part of their everyday work. The contemporary digital libraries contain enormous amounts of digitised content, most of which is in the form of pictures. Librarians are a special group who frequently interact with this type of content aiming to provide annotation that fulfil the search needs of everyday people as well as the requirements of specific groups such as researchers working on the fields of image and multimedia retrieval. The fact that the librarians come from two different countries reduces the cultural bias of image interpretation. However, we cannot claim that this bias is fully eliminated since the way people evaluate the content of an image varies significantly across the Globe.

3.3. *Mathematical formulation*

Let us denote by P^i the ith participant ($i = 1, \ldots, N_P$) of a total of N_P participants ($N_P = 295$ in this study as already mentioned above). We also denote with I^j the jth image ($j = 1, \ldots, N_I$) in the image dataset where N_I is the total number of images annotated at least from one user (in our case $N_I = 955$). By set $H = \{h_1, h_2, \ldots, h_{N_H}\}$ we define the set of hashtags the owners/creators used to tag the images in set I while N_H is the total number of tags (in this study $N_H = 557$) used for this purpose. We should note here that the number of hashtags is smaller than the number of images due to the methodology we followed to collect images. As mentioned before the set of 1000 images were retrieved from 100 different subjects/hashtags (10 relevant images per subject/hashtag) so there was at least one common hashtag in each category (10 images). Moreover, common hashtags were located between the categories as well.

In order be able to conclude on the research questions defined earlier we must use some effectiveness measures. For

[10] See also: http://cis.cut.ac.cy/~nicolas.tsapatsoulis/aiai2015/code/user_dem.php.

this purpose we modified the well known Recall, Precision and F-measures [65] to fit in with the current experiment. In particular we define the participant's P^i recall value, R_{ij}, for image I^j as the proportion of owner's hashtags, for this image, that were selected by P^i in the questionnaire. In a mathematically formal way this is given by:

$$R_{ij} = \frac{\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\|}{\|\mathbf{T}_{jc}\|} \tag{1}$$

where $\mathbf{T}_{jc}$ is the set of distinct hashtags assigned to image I^j by the image owner, $\mathbf{T}_{ji}$ is the set of distinct hashtags the participant P^i assigned to image I^j (based on the choices presented to him/her in the questionnaire), $\cap$ is the set intersection operation and $\|\Omega\|$ denotes the cardinality of set Ω.

Extending Eq. (1) across all images participant P^i annotated we get the overall per participant recall value:

$$R_i = \frac{\sum_{j=1}^{N_I} \left\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\right\|}{\sum_{j=1, T_{ji}\neq\emptyset}^{N_I} \left\|\mathbf{T}_{jc}\right\|} \tag{2}$$

where the constraint $T_{ji} \neq \emptyset$ indicates that summation refers only to the images participant P^i annotated.

The overall per image recall value is computed with the aid of Eq. (3):

$$R_j = \frac{\sum_{i=1}^{N_P} \left\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\right\|}{N_P^j \cdot \left\|\mathbf{T}_{jc}\right\|} \tag{3}$$

where N_P^j is the number of participants who annotated image I^j.

In a similar manner we define per image (see Eq. (5)) and per participant precision (see Eq. (6)), i.e., the proportion of a participant's choices that coincide with owner's hashtags, and F-measure (harmonic mean of recall and precision) as follows:

$$P_{ij} = \frac{\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\|}{\|\mathbf{T}_{ji}\|} \tag{4}$$

(precision of participant's P^i choices for jth image)

$$P_j = \frac{\sum_{i=1}^{N_P} \left\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\right\|}{\sum_{i=1}^{N_P} \left\|\mathbf{T}_{ji}\right\|} \tag{5}$$

$$P_i = \frac{\sum_{j=1}^{N_I} \left\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\right\|}{\sum_{j=1}^{N_I} \left\|\mathbf{T}_{ji}\right\|} \tag{6}$$

$$F_j = \frac{2 \cdot R_j \cdot P_j}{R_j + P_j} \tag{7}$$

$$F_i = \frac{2 \cdot R_i \cdot P_i}{R_i + P_i}. \tag{8}$$

Let us now assume an index of hashtags $\vec{V}$ in which all the hashtag choices presented to the participants though the questionnaire images are concatenated. That is, if in the questionnaire the participants are asked to choose between 8 hashtags in the first image then these hashtags are the first 8 entries of vector $\vec{V}$. The available hashtag choices for the second image of the questionnaire will follow, then that of the third image and so on. Note that in index V the same hashtag may appear more than once and in different position indicating a particular choice for a specific image.

If we denote with '1' the hashtags chosen by a specific participant and with '0' the hashtags not chosen then a participant P^i can be represented by a binary vector $\vec{P^i}$, with length equal to that of index $\vec{V}$, denoting his/her 'profile'. In a similar way we can define the creators/owners vector, say $\vec{C}$ in which the hashtags used by the photo owners are represented with ones and hashtags not used by zeros. Obviously, the vector $\vec{C}$ does not correspond to a specific user profile but to the aggregated profile of all photo owners. The similarity of images' interpretation between photo owners / creators and each one of the participants can be, then, estimated by any vector comparison metric. Because both vectors $\vec{C}$ and $\vec{P^i}$ are binary ones the choose of Hamming distance [66] is evident. The aforementioned distance was introduced by Richard Hamming, is implied only at two equal strings and gives the number of positions at which corresponding symbols differ [66].

Thus, the similarity $S(C, P^i)$ between the choices a participant P^i made in order to characterize the images in the questionnaire with the actual hashtags the owners used, is given by:

$$S(C, P^i) = 1 - \frac{h(\vec{C}, \vec{P^i})}{L} \tag{9}$$

where $h(\vec{C}, \vec{P^i})$ is the Hamming distance of vectors $\vec{C}$ and $\vec{P^i}$ and L is the corresponding vector space dimension (i.e., the length of vectors $\vec{C}$ and $\vec{P^i}$ and index $\vec{V}$).

4. Experimental results and discussion

The data of the 295 filled in questionnaires were analyzed with aid of SPSS,[11] MS Excel[12] and the MATLAB[13] platform using the metrics defined in the previous section. Three users were identified as outliers, due to extremely low F-measure (users with ids 145 and 212) or unexpectedly a high number of keywords per image (user with id 203[14]), and their answers were ignored. Fig. 3 shows the per participants' Recall (Eq. (2)), Precision (Eq. (6)) and F-measure (Eq. (8)) of the pictures each participant had to interpret in the experiment (in the diagram we show the metrics for the 40 participants having the more extreme F-measure scores). As already explained not all participants evaluated all images; thus, the

[11] http://www-01.ibm.com/software/analytics/spss/.
[12] https://products.office.com/en-us/excel.
[13] http://www.mathworks.com/products/matlab/.
[14] For more details see http://cis.cut.ac.cy/~nicolas.tsapatsoulis/aiai2015/code/user_annot.php.

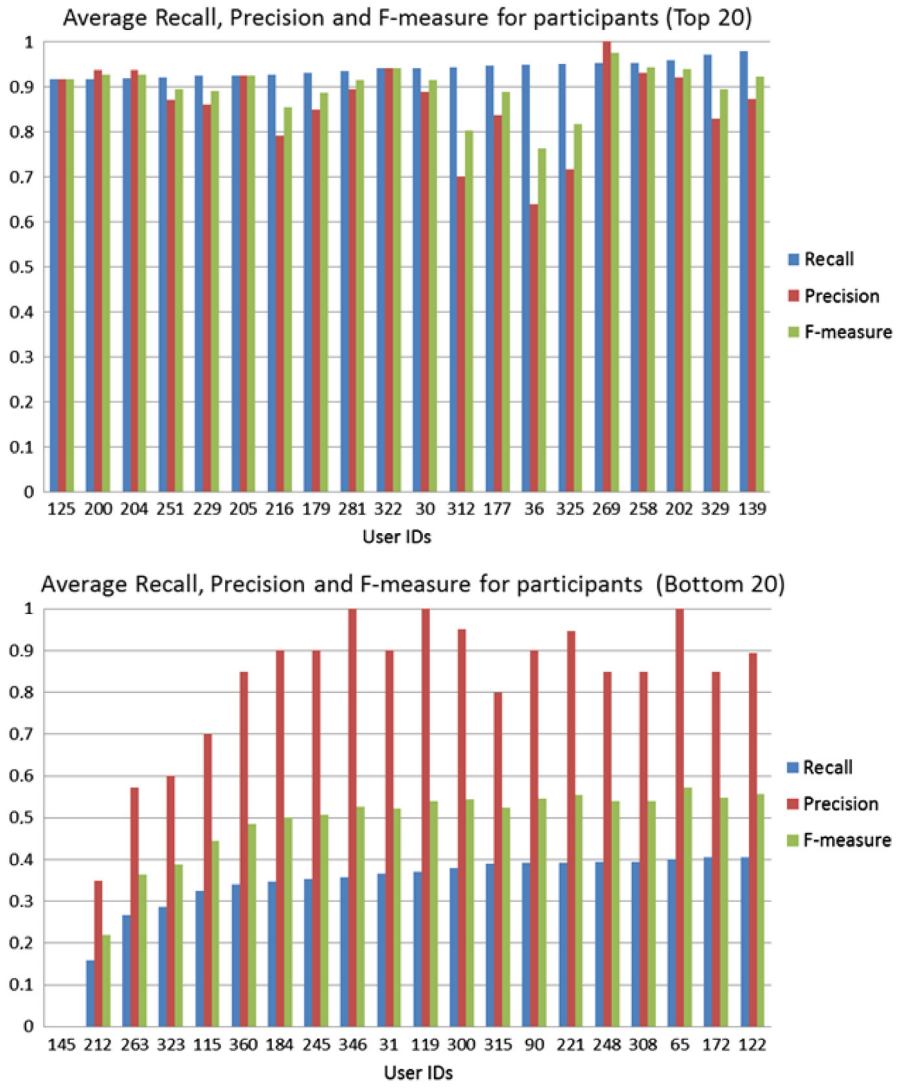

Fig. 3 – Average hashtags' recall, precision and *F*-measure per participant. The top 20 and bottom 20 performing participants are shown.

Table 3 – Per participant recall, precision and *F*-measure value statistics.

	Mean	St. Dev.	Minimum	Maximum
Recall	0.661	0.182	0.267	0.980
Precision	0.919	0.079	0.571	1.000
F-measure	0.751	0.122	0.364	0.976

computations were done using the subsets of images shown to the participants according to the questionnaire they were given. Fig. 4 shows the average hashtags' recall, precision and *F*-measure for all participants. For ease interpretation, the user ids were sorted based on *F*-measure (top diagram) and recall(bottom diagram) from lowest to highest values.

Some basic statistics of the per participant Recall, Precision and *F*-measure are shown in Table 3. We see there that the recall performance per participant is 0.661 ± 0.182 with the extreme values being 0.267 (minimum) and 0.980 (maximum). Thus, the conclusion is that at least two out three hashtags used by the owner in Instagram images is relevant to image content since other users consider it descriptive as well. The variation in performance, among users, is rather low indicating that in the experiment there were no spammers or users with dishonest behavior (excluding the three users mentioned earlier). The per participant precision is significantly higher (0.919 ± 0.079) than recall, showing the tendency of people to use as few as possible keywords to describe an image. This is in agreement with the generic behavior of Web users who use, on average, one to three keywords when searching for information through search engines [67] as well as with similar findings regarding the number of hashtags accompanying Instagram images [20]. Of course, we do not know whether this is an intrinsic human tendency or a behavior cultivated by the way search engines work (the fewer the keywords given the more the results presented to the user). Furthermore, the high precision values indicate also that the participants did not answer (chose hashtags for the shown images) randomly.

Overall, with the aid of Fig. 3 and Table 3 we can conclude on both research questions set in this study. Given that

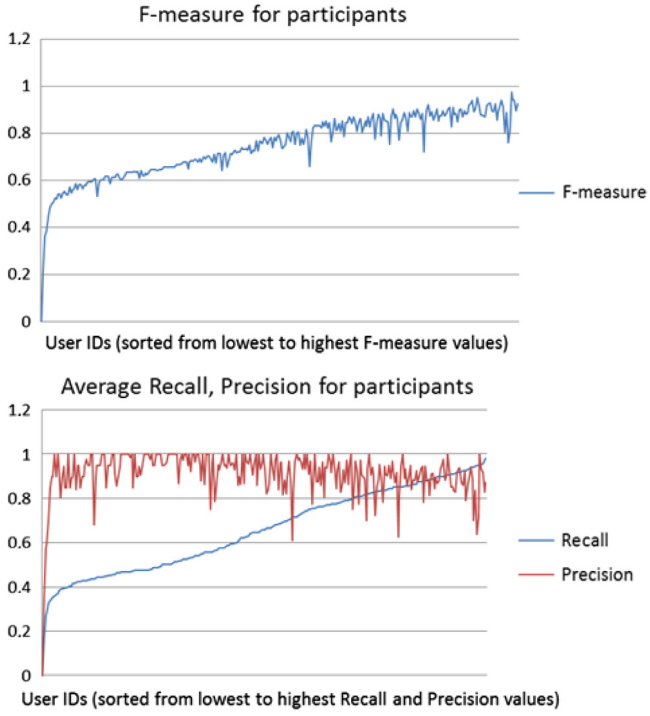

Fig. 4 – Average hashtags' recall, precision and *F*-measure per participant. For ease interpretation the user ids were sorted based on *F*-measure (top) and recall (bottom) from lowest to highest values.

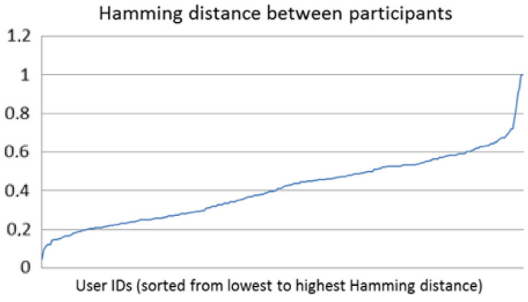

Fig. 5 – The hamming distance between participants and image owners/creators.

the participants in our experiments can be seen as experts (librarians and students having high experience with visual content tagging in the Internet and social media) we can claim that around 66% of the Instagram hashtags, that accompany images, are relevant to the actual content of the images and can be used for training purposes in an AIA context. The results confirm also the findings in our previous study [64] where 55% Instagram hashtags that accompany images proved to be relevant to the actual content of the images. The difference in scores (66% in the current study vs. 55% in the previous one) recall is higher, from 0.55 to 0.66 than in the previous one, can be attributed to increased participation and larger dataset.

By pointing out that on average only 30% of the (owner's) Instagram image hashtags are relevant to the images close to which they appear we can state that on average 20% (0.66·0.3) of Instagram hashtags are related with the visual content of Instagram images.

Table 4 – Statistics of normalized hamming distance between participants and photo owners in image interpretation.

Mean	St. Dev.	Minimum	Maximum
0.408	0.170	0.048	1.000

Fig. 5 shows the dissimilarity of image interpretation between each one of the participants and the photo owners with the aid of (normalized) Hamming distance and the mathematical formulation presented in the previous section. By normalized we mean that the Hamming distance is divided by the length of the strings compared (in our case total number of choices presented to the users - accompanying the photos- in the questionnaire session(s) they took). As we see in Table 4 the average normalized Hamming distance between the photo owners and the participants is 0.408 ± 0.170. This means that there is on average 40% disagreement (only two out of five hashtag choices/non-choices between image owners and participants differ); thus, we can confirm, once again, that the participants do not answer at random or in any dishonest manner. By looking at the extreme values in Fig. 6 we see that two users (those with ids 212 and 145) filled in the questionnaire in a clearly unfair way (total dissimilarity with hashtag choices/non choices of owners/creators) while another four (those with ids 263, 323, 115, 137) show unexpectedly low performance (high dissimilarity with the interpretation of picture owners) and could be easily filtered out. We should mention here that the users with ids 212 and 145 had been already identified as 'spammers' due to very low *F*-measure score. On the other hand, the user with id 269 presents

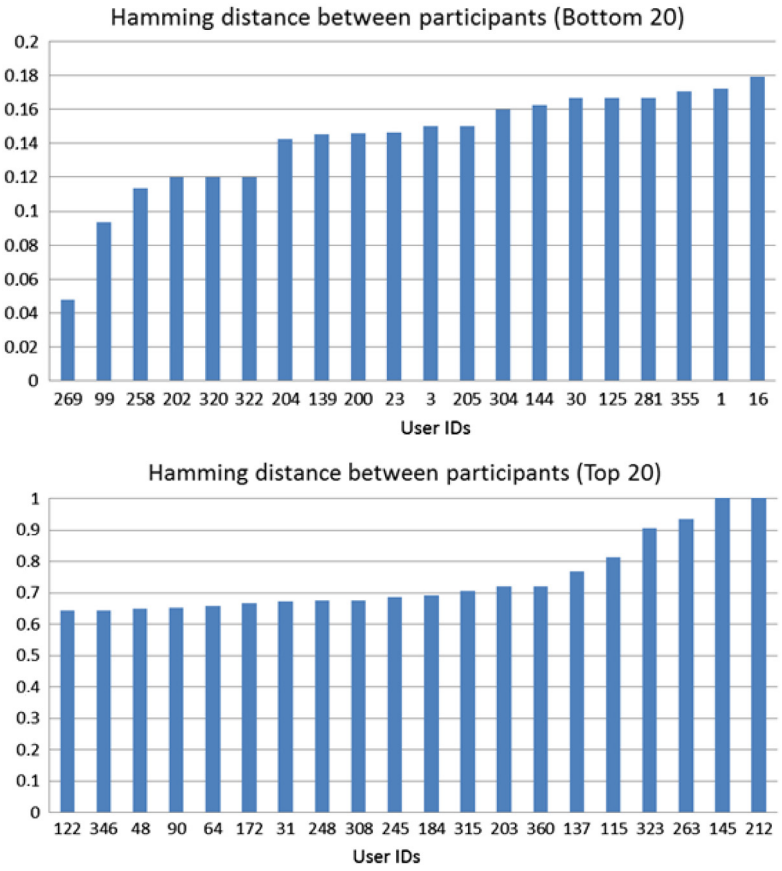

Fig. 6 – Hamming distance between participants and image owners/creators. In the top diagram we show the normalized distances for the 20 users with the best performance (lowest Hamming distance) while in the bottom diagram we show the distances of the 20 least performing users.

Table 5 – Per image Recall, Precision and *F*-measure value statistics.

	Mean	St. Dev.	Minimum	Maximum
Recall	0.655	0.251	0.111	1.000
Precision	0.988	0.112	0.167	1.000
F-measure	0.814	0.185	0.200	1.000

an excellent performance which indicates that even perfect matching between owners and participants is not impossible; this means that the hashtags given by the owners to the photos are indeed related with the visual content of images (i.e., what the images actually show and not, for instance, context or emotional information).

In Fig. 7 we present the per image Recall (Eq. (3)), Precision (Eq. (5)), and *F*-measure (Eq. (7)) values while in Table 5 are shown summary statistics for those values. We should mention here that keywords selected by only one user for images that received more than two annotations were considered 'noise' and were excluded from the calculations.[15] The basic aim of this analysis is to check whether the difficulty of interpreting images depends on their visual content. Comparing Tables 3 and 5 we observe that the variation of Recall, Precision, and *F*-measure across images is higher than that across participants. The same also holds for the extreme values. Thus, we can conclude that image content affects interpretability.

In Fig. 8 we show the images, annotated by at least 35 users each, with the lowest recall scores (from left to right images with ids 1366, 1677 and 1256). In the first case (photo annotated by 40 users, recall = 0.4, precision = 1.0) the owner gave the hashtags #dog, #bathtime, #bubbles but, probably due to photo resolution, only 10 out of the 40 users that annotated this photo selected the hashtag #bathtime and none of them selected the hashtag #bubbles. Similarly, for the photo with id 1677 (photo annotated by 35 users, recall = 0.44, precision = 0.94) the owner gave the hashtags #plate and #porcelain but only 11 out of 35 users selected the first while 20 out of 35 users selected the latter (#porcelain). It seems that, probably due to the angle this photo was taken, it is difficult for the users to interpret it. Finally, the photo with id 1256 (photo annotated by 38 users, recall = 0.46, precision = 1.00) was assigned by the owner the hashstags #flowers, #spring, #summer. While the first hashtag was easily recognized by the users, none of them selected the hashtag #summer and only 20 out of 38 selected the hashtag #spring. Both #spring and #summer can be considered as abstract concepts in terms of visual identification. However, flowers are strongly correlated with

[15] Per image annotation results can be seen at: http://cis.cut.ac.cy/~nicolas.tsapatsoulis/aiai2015/code/photo_annot.php.

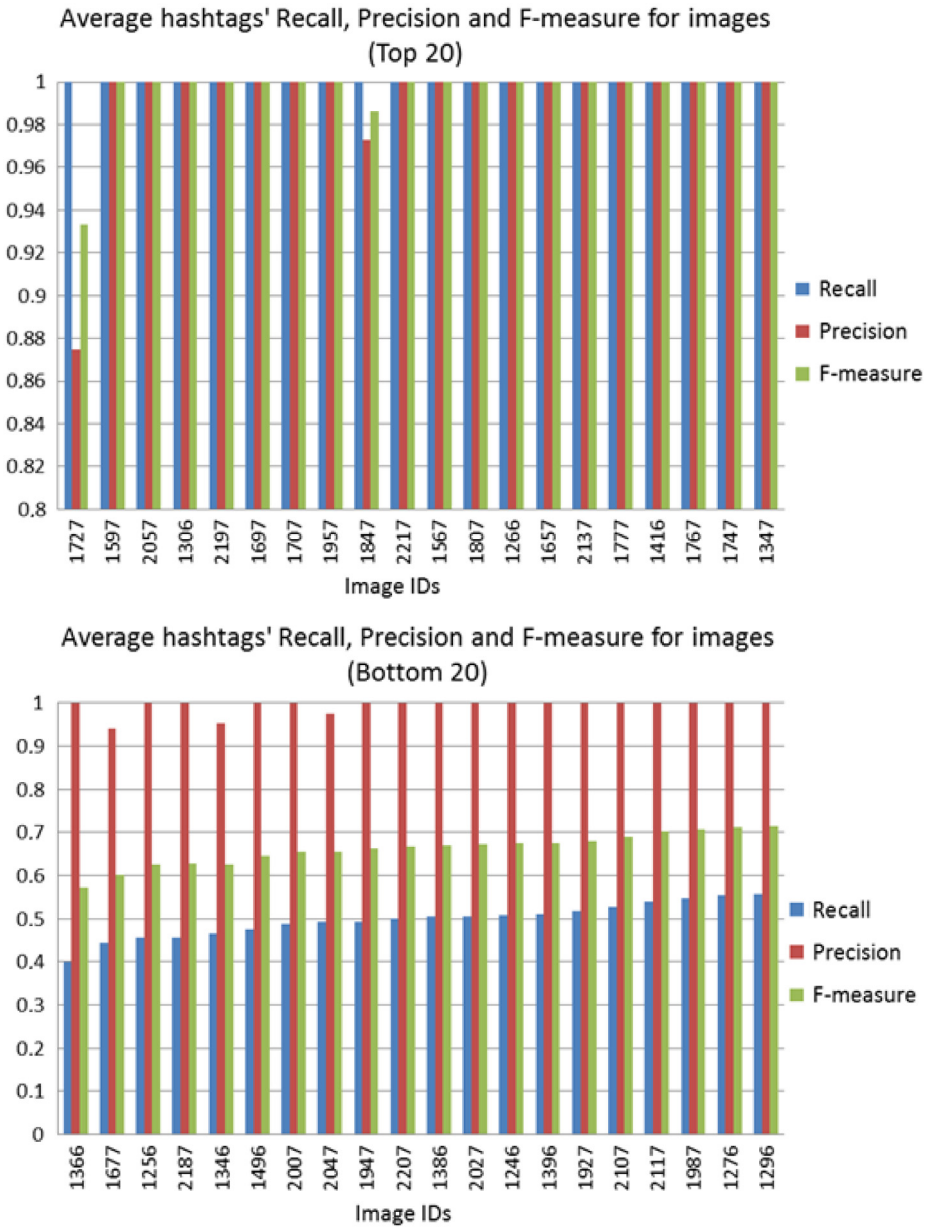

Fig. 7 – Average hashtags' recall, precision and *F*-measure per image. We show the scores for the 20 most easy and most difficult (in terms of recall) to interpret photos.

Fig. 8 – Difficult to interpret images.

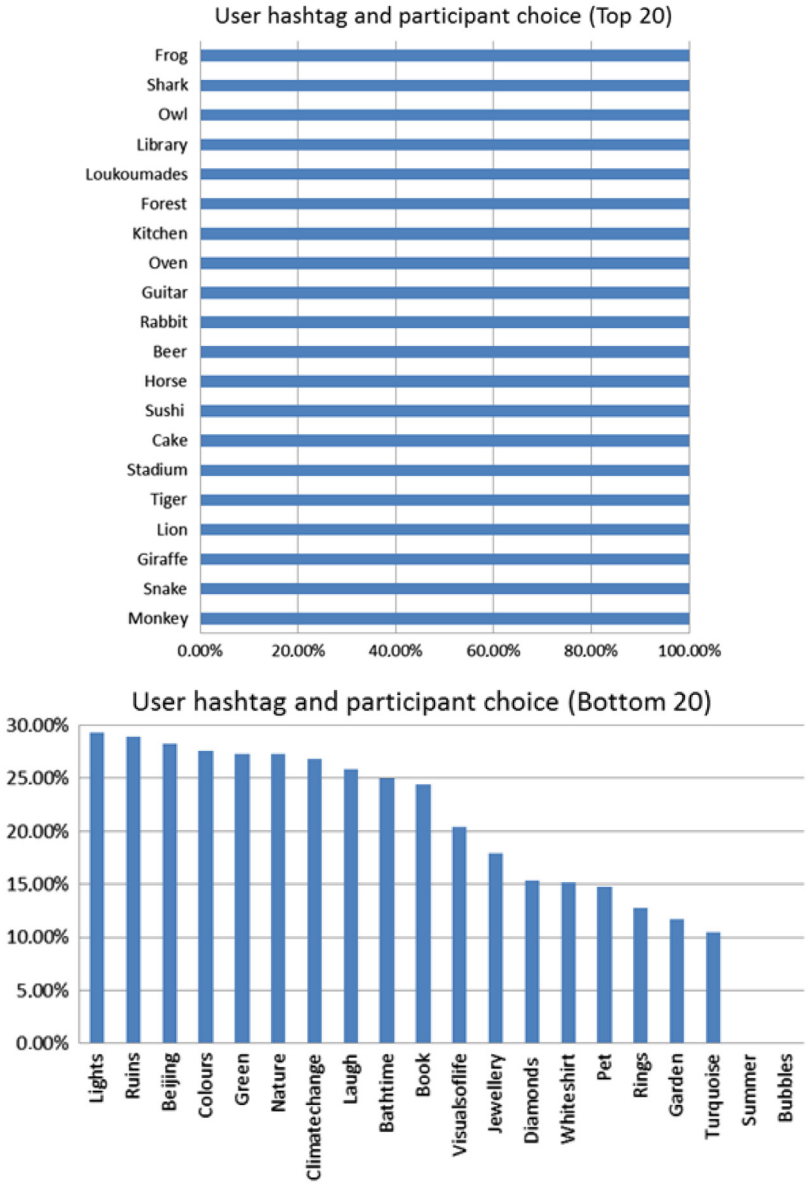

Fig. 9 – Percentage of the participants that chose each of the owner/creator hashtags. The 20 most easily (top) and hardest (bottom) to retrieve hashtags are shown.

spring; thus, half of the users made the association and selected spring as a keyword for this particular photo. We should mention here that the set of difficult to interpret images differs from our previous study [64] since the images used in the experiments also differ (see the discussion in Section 3.2). However, the methodology used to identify these images remains same: The difficult to interpret images are those having the lowest hashtag recall score among the participants.

In the last part of our analysis, we deal with the recall values of the hashtags. Our assumption is that abstract concepts should have lower recall values than concepts referring to tangible objects. Fig. 9 presents the recall values for the 20 most easily and the 20 hardest to retrieve owners' hashtags. It is clear that abstract concepts tend to have low recall values, as expected (see for instance 'Climatechange', 'Visualsoflife', 'Ruins'), however, there still many hashtags referring to non-abstract concepts that have low recall values as well (e.g. 'Beijing', 'Pet', 'Rings', 'Book'). This lead us to the conclusion that out of context interpretation of images is, in some cases, problematic. Nevertheless, the difficulty of interpretation in this case does not necessarily mean that the hashtag used by the owner is inappropriate for characterizing the particular image. By saying so we mean that the pair image-hashtag is still a good training example. Finally, the two hashtags with zero recall values ('Summer', 'Bubbles') had been already identified (see the earlier discussion on the difficult to interpret images) as problematic cases due to irrelevant use by the owner ('Summer') and low photo resolution in the questionnaire ('Bubbles').

5. Conclusion

In this paper we have presented our study about the descriptive value of Instagram hashtags as metadata for the images they accompany. By measuring if the participants would choose the same hashtags with the image creator / owner we found that in the 66% of the chosen hashtags participants and owners agree that the suggested hashtags can describe the visual content of an image. Moreover, we have an indication that approximately 20% of the Instagram hashtag datasets are appropriate for use in training examples (image—tag pairs) for machine learning algorithms. The results show also that an important portion of image hashtags in Instagram are not directly related with the concept depicted by the image. We have found also that both the image content and the context in which an image resides affect its interpretability. However, as we explained, this does not necessarily imply that the pairs images - difficult to interpret tags are invalid for training purposes.

In order to achieve Automatic Image Annotation is necessary to create good training examples, i.e. pairs of images and relevant tags. From the results we can conclude that large part of hashtags are not directly related with image's visual content. So research has to be done in order to locate and remove stophashtags and fully automated the hashtag selection procedure. As stophashtags we can define meaningless hashtags that frequently occur in different categories and these hashtags represent noise [68]. In a recent work [53] we propose a theoretical and empirical framework through which stophashtags can be identified. Also, techniques that create textual descriptions from images (such as those of Farhadi et al. [47], Kiros et al. [48], Karpathy & Fei-Fei [27] and Johnson et al. [52]) with aid of deep learning or similar approaches [69] can be used as an objective, cross-checking, mechanism for identifying Instagram hashtags that are relevant to image content.

Another action that can be taken in the future is to check the validity of image-hashtag pairs for training visual concept models (see [7]) in practice. This will lead us to a second, practical, stage of investigation and will allow comparison of the theoretical findings of this study with practical issues faced during training. However, we must be aware that, in machine learning, good training examples must be properly processed to extract appropriate, for learning, low level features; this is by no means an easy task.

Acknowledgment

The authors would like to thank Ioannis Despotis for his valuable help in the design of the online questionnaire and the technical support.

REFERENCES

[1] Zephoria, The top 20 valuable facebook statistics, updated December 2015. Online at: https://zephoria.com/top-15-valuable-facebook-statistics.

[2] Instagram, Our story: A quick walk through our history as a company, 2016. Online at: https://www.instagram.com/press/?hl=en.

[3] T.A. Souza Coelho, P.P. Calado, L.V. Souza, B. Ribeiro-Neto, R. Munt, Image retrieval using multiple evidence ranking, IEEE Trans. Knowl. Data Eng. 16 (4) (2004) 408–417.

[4] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2) (2008) 5:1–5:60.

[5] A. Hanbury, A survey of methods for image annotation, J. Vis. Lang. Comput. 19 (5) (2008) 617–627.

[6] C. Jin, S.-W. Jin, Automatic image annotation using feature selection based on improving quantum particle swarm optimization, Signal Process. 109 (C) (2015) 172–181.

[7] Z. Theodosiou, N. Tsapatsoulis, Image retrieval using keywords: The machine learning perspective, in: Semantic Multimedia Analysis and Processing, CRC press, 2014, pp. 3–30.

[8] C.G.M. Snoek, M. Worring, Concept-based video retrieval, Found. Trends Inf. Retr. 2 (4) (2009) 215–322.

[9] Z. Theodosiou, N. Tsapatsoulis, Crowdsourcing annotation: Modelling keywords using low level features, in: Proc. of the 5th International Conference on Internet Multimedia Systems Architecture and Application, IEEE, 2011, pp. 1–4.

[10] M. Duggan, Mobile messaging and social media 2015, in: Pew Research Center, 2015. Online at: http://www.pewinternet.org/files/2015/08/Social-Media-Update-2015-FINAL2.pdf/.

[11] M. Baranovic, What #hashtags mean to mobile photography, in: Digital Photography Review, 2013. Online at: http://connect.dpreview.com/post/1256293279/hastag-photography.

[12] L. Kolowich, The history of hashtags [infographic], in: Hubspot, 2014. Online at: http://blog.hubspot.com/marketing/history-of-hashtags.

[13] T.A. Small, What the hashtag? Inf. Commun. Soc. 14 (6) (2011) http://dx.doi.org/10.1080/1369118X.2011.554572.

[14] Z. Zdziarski, C. Bourges, J. Mitchell, P. Houdyer, D. Johnson, R. Dahyot, On summarising the 'here and now' of social videos for smart mobile browsing, in: Proceedings of the 2014 International Workshop on Computational Intelligence for Multimedia Understanding, IWCIM'14, IEEE, 2014, pp. 1–5.

[15] A.R. Daer, R.F. Hoffman, S. Goodman, Rhetorical functions of hashtag forms across social media applications, Commun. Des. Quart. Rev. 3 (1) (2014) 12–16.

[16] S.M. Mohammad, S. Kiritchenko, Using hashtags to capture fine emotion categories from tweets, Comput. Intelligence 31 (2) (2015) 301–326.

[17] F. Kunneman, C. Liebrecht, A. van den Bosch, The (un)predictability of emotional hashtags in twitter, in: Proceedings of the 5th Workshop on Language Analysis for Social Media, (LASM), Association for Computational Linguistics, 2014, pp. 26–34.

[18] D.M. Carmean, M.E. Morris, Selfie examinations: Applying computer vision, hashtag scraping and sentiment analysis to finding and interpreting selfies, Tech. Rep., Intel Labs, Hillsboro, OR, 2015, URL http://nebula.wsimg.com/27bab6eda0e75b69fcab8a5cdc4e22af?AccessKeyId=A6A4DAF733A0F616E396&disposition=0.

[19] Y. Zhang, F. Baghirov, H. Hashim, J. Murphy, Gender and instagram hashtags: A study of #malaysianfood, in: Proceedings of the 'eTourism: Empowering Places', ENTER'16, 2016. URL http://agrilife.org/ertr/files/2016/01/ENTER2016_submission_118_.pdf.

[20] E. Ferrara, R. Interdonato, A. Tagarelli, Online popularity and topical interests through the lens of instagram, in: Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT'14, ACM, 2014, pp. 24–34.

[21] F. Fauzi, J.-L. Hong, M. Belkhatir, Webpage segmentation for extracting images and their surrounding contextual information, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM, 2009, pp. 649–652.
[22] P.M. Joshi, S. Liu, Web document text and images extraction using dom analysis and natural language processing, in: Proceedings of the 9th ACM Symposium on Document Engineering, ACM, 2009, pp. 218–221.
[23] G. Tryfou, Z. Theodosiou, N. Tsapatsoulis, Web image context extraction based on semantic representation of web page visual segments, in: Proc. of International Workshop on Semantic and Social Media Adaptation and Personalization, IEEE, 2012, pp. 63–67.
[24] F. Schroff, A. Criminisi, A. Zisserman, Harvesting image databases from the web, IEEE Trans. Pattern Anal. Mach. Intell. 33 (4) (2011) 754–766.
[25] B. Sigurbjörnsson, R. van Zwol, Flickr tag recommendation based on collective knowledge, in: Proceedings of the 17th International Conference on World Wide Web, WWW'08, ACM, 2008, pp. 327–336.
[26] T. Joachims, L. Granka, B. Pan, H. Hembrooke, G. Gay, Accurately interpreting clickthrough data as implicit feedback, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'05, ACM, 2005, pp. 154–161.
[27] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'15, IEEE Computer Society, 2015, pp. 3128–3137. http://dx.doi.org/10.1109/CVPR.2015.7298932.
[28] Z. Wang, Z.-H. Zhou, Crowdsourcing label quality: a theoretical analysis, Sci. China Inf. Sci. 58 (11) (2015) 1–12. http://dx.doi.org/10.1007/s11432-015-5391-x.
[29] A. Ulges, C. Schulze, M. Koch, T.M. Breuel, Learning automatic concept detectors from online video, Comput. Vis. Image Underst. 114 (4) (2010) 429–438. http://dx.doi.org/10.1016/j.cviu.2009.08.002.
[30] X.-J. Wang, W.-Y. Ma, X. Li, Exploring statistical correlations for image retrieval, Multimedia Syst. 11 (4) (2006) 340–351.
[31] A. Ulges, C. Schulze, M. Koch, T.M. Breuel, Content analysis meets viewers: linking concept detection with demographics on youtube, Int. J. Multimedia Inf. Retr. 2 (2) (2013) 145–157.
[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the 2009 IEEE International Conference on Computer Vision and Pattern Recognition, CVPR09, IEEE, 2009, pp. 248–255.
[33] X. Chen, A. Shrivastava, A. Gupta, Imagenet: A large-scale hierarchical image database, in: Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV'13, IEEE, 2013, pp. 1409–1416.
[34] N.H. Do, K. Yanai, Automatic construction of action datasets using web videos with density-based cluster analysis and outlier detection, in: Proceedings of the 7th Pacific-Rim Symposium, PSIVT 2015, 2015, pp. 160–172.
[35] A. Ulges, M. Worring, T. Breuel, Learning visual contexts for image annotation from flickr groups, IEEE Trans. Multimedia 13 (2) (2011) 330–341.
[36] K. Ntalianis, N. Tsapatsoulis, A. Doulamis, N. Matsatsinis, Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution, Multimedia Tools Appl. 69 (2) (2014) 397–421.
[37] J. Cui, L. Liu, H. Wang, C. Du, W. Song, Tagged image clustering via topic models, in: Proceedings of the 27th Chinese Control and Decision Conference, CCDC'15, Microsoft, IEEE, 2015, pp. 4424–4429. http://dx.doi.org/10.1109/CCDC.2015.7162653.
[38] Z. Xia, X. Feng, J. Peng, J. Fan, Content-irrelevant tag cleansing via bi-layer clustering and peer cooperation, J. Signal Process. Syst. 81 (1) (2015) 29–44. http://dx.doi.org/10.1007/s11265-014-0895-y.
[39] T. Tsikrika, C. Diou, A.P. de Vries, A. Delopoulos, Image annotation using clickthrough data, in: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR'09, ACM, 2009, pp. 14:1–14:8.
[40] C. Macdonald, I. Ounis, Usefulness of quality click-through data for training, in: Proceedings of the 2009 Workshop on Web Search Click Data, WSCD'09, ACM, 2009, pp. 75–79.
[41] X. He, O. King, W.-Y. Ma, M. Li, H.-J. Zhang, Hidden annotation for image retrieval with long-term relevance feedback learning, IEEE Trans. Circuits Syst. Video Technol. 13 (1) (2003) 39–48.
[42] W. Jiang, G. Er, Q. Dai, J. Gu, Hidden annotation for image retrieval with long-term relevance feedback learning, Pattern Recognit. 38 (11) (2005) 2007–2021.
[43] T. Tsikrika, C. Diou, A.P. de Vries, A. Delopoulos, Crowdsourcing label quality: a theoretical analysis, Multimedia Tools Appl. 55 (1) (2011) 27–52.
[44] I. Sarafis, C. Diou, T. Tsikrika, A. Delopoulos, Weighted svm from clickthrough data for image retrieval, in: 2014 IEEE International Conference on Image Processing, (ICIP), IEEE, 2014, pp. 3013–3017.
[45] I. Sarafis, C. Diou, A. Delopoulos, Building effective svm concept detectors from clickthrough data for large-scale image retrieval, Int. J. Multimedia Inf. Retr. 4 (2015) 129–142. http://dx.doi.org/10.1007/s13735-015-0080-5.
[46] N. Tsapatsoulis, Web image indexing using WICE and a learning-free language model, in: 12th IFIP WG 12.5 International Conference on Artificial Intelligence Applications and Innovations, in: IFIP Advances in Information and Communication Technology, Springer International Publishing, 2016, pp. 131–140. http://dx.doi.org/10.1007/978-3-319-44944-9_12.
[47] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, Springer-Verlag, 2010, pp. 15–29.
[48] R. Kiros, R. Salakhutdinov, R. Zemel, Multimodal neural language models, in: Proceedings of the 31st International Conference on Machine Learning, ICML 2014, International Machine Learning Society, 2014, pp. 2012–2025.
[49] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, J. Artificial Intelligence Res. 47 (1) (2013) 853–899.
[50] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.
[51] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, C.L. Zitnick, Microsoft coco: Common objects in context, in: Proceedings of the 13th European Conference on Computer Vision, ECCV'14, Springer International Publishing, 2014, pp. 740–755.
[52] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'16, IEEE Computer Society, 2016, pp. 4565–4574.
[53] S. Giannoulakis, N. Tsapatsoulis, Defining and identifying stophashtags in instagram, in: Proceedings of the 2nd INNS Conference on Big Data, in: Advances in Intelligent Systems and Computing, Springer International Publishing, 2016, http://dx.doi.org/10.1007/978-3-319-47898-2.
[54] S. Nowak, S. Rüger, How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation, in: Proceedings of the International Conference on Multimedia Information Retrieval, MIR'10, ACM, 2010, pp. 557–566.

[55] B. Ionescu, A. Popescu, A.-L. Radu, H. Mller, Result diversification in social image retrieval: a benchmarking framework, Multimedia Tools Appl. 75 (2) (2016) 1301–1331. http://dx.doi.org/10.1007/s11042-014-2369-4.

[56] A.A. Veloso, J.A.d. Santos, K. Nogueira, Learning to annotate clothes in everyday photos: Multi-modal, multi-label, multi-instance approach, in: Proceedings of the 27th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI'14, IEEE Computer Society, 2014, pp. 327–334.

[57] N.R. Asheghi, S. Sharoff, K. Markert, Crowdsourcing for web genre annotationt, Lang. Resour. Eval. (2016) 1–39. http://dx.doi.org/10.1007/s10579-015-9331-6.

[58] R. Di Salvo, D. Giordano, I. Kavasidis, A crowdsourcing approach to support video annotation, in: Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications, VIGTA'13, ACM, 2013, pp. 8:1–8:6. http://dx.doi.org/10.1145/2501105.2501113.

[59] Y.E. Kara, G. Genc, O. Aran, L. Akarun, Modeling annotator behaviors for crowd labeling, Neurocomputing 160 (C) (2015) 141–156.

[60] Z. Theodosiou, O. Georgiou, N. Tsapatsoulis, Evaluating annotators consistency with the aid of an innovative database schema, in: Proceedings of the 6th International Workshop on Semantic Media Adaptation and Personalization, SMAP 2011, IEEE, 2011, pp. 74–78.

[61] Y. Baba, H. Kashima, Statistical quality estimation for general crowdsourcing tasks, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'13, ACM, 2013, pp. 554–562.

[62] Q. Li, F. Ma, J. Gao, L. Su, C.J. Quinn, Crowdsourcing high quality labels with a tight budget, in: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM'16, ACM, 2016, pp. 237–246.

[63] Q. Hu, Q. He, H. Huang, K. Chiew, Z. Liu, A formalized framework for incorporating expert labels in crowdsourcing environment, J. Intell. Inf. Syst. 7 (2015) 1–23. http://dx.doi.org/10.1007/s10844-015-0371-6.

[64] S. Giannoulakis, N. Tsapatsoulis, Instagram hashtags as image annotation metadata, in: 11th IFIPWG 12.5 International Conference on Artificial Intelligence Applications and Innovations, in: IFIP Advances in Information and Communication Technology, Springer International Publishing, 2015, pp. 206–220. http://dx.doi.org/10.1007/978-3-319-23868-5_15.

[65] W. Hersh, Terms, models, resources, and evaluation, in: Information Retrieval, Health Informatics, Springer, New York, 2009, pp. 3–39. http://dx.doi.org/10.1007/978-0-387-78703-9.

[66] A. Kulshrestha, On the hamming distance between base-n representations of whole numbers, Can. Young Sci. J. (2012) 14–17.

[67] K. Bradley, B. Smyth, Personalized information ordering: a case study in online recruitment, Knowl.-Based Syst. 16 (5–6) (2003) 269–275.

[68] G. Armano, F. Fanni, A. Giulian, Stopwords identification by means of characteristic and discriminant analysis, in: Proceedings of the International Conference on Agents Artificial IntellIgence, ICAART'15, SCITEPRESS Digital Library, 2015, pp. 353–360. http://dx.doi.org/10.5220/0005194303530360.

[69] R. Socher, C.-Y. Lin, A. Ng, C. Manning, Parsing natural scenes and natural language with recursive neural networks, in: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, International Machine Learning Society, 2011, pp. 129–136.