Cyprus University of Technology | Department of Communication and Internet Studies

Doctoral Dissertation

Instagram hashtags as a source of semantic information for Automatic Image Annotation

Stamatios Giannoulakis

Advisor: Nicolas Tsapatsoulis

October 2021

CYPRUS UNIVERSITY OF TECHNOLOGY

FACULTY OF COMMUNICATION AND MEDIA STUDIES

DEPARTMENT OF COMMUNICATION AND INTERNET STUDIES

Doctoral Dissertation

Instagram hashtags as a source of semantic information for Automatic Image Annotation

by

Stamatios Giannoulakis

Advisor: Nicolas Tsapatsoulis

Limassol, October 2021

Approval Form

Doctoral Dissertation Proposal

Instagram hashtags as a source of semantic information for Automatic Image Annotation

Presented by

Stamatios Giannoulakis

Supervisor: Nicolas Tsapatsoulis, Professor,
Faculty of Communication and Media Studies, Cyprus University of Technology

Signature:

Committee Member: Nikos Grammalidis, Researcher B',
Center for Research & Technology Hellas

Signature:

Committee Member:Klimis Ntalianis, Professor,
Department of Business Administration, University of West Attica

Signature:

Cyprus University of Technology
Limassol, October 2021

Copyrights

The approval of the dissertation by the Department of Communication and Internet Studies does not imply necessarily the approval by the Department of the views of the writer.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Nicolas Tsapatsoulis. My supervisor's precious help and support were vital to fulfill my target and finish my Ph.D. thesis—also, the members of the advisory committee, Dr. Lambros Lambrinos, and Dr. Christos Papatheodorou. A special thank you to Eftychios Protopadakis for the valuable help and support. In addition, I wish to thank Christos Rodosthenous for the support and discussions we had. Moreover, I would like to thank Katerina Xenou, who urged me to start the Ph.D., and her support. A special thank you to Vasilis Manavopoulos, and Christos Panayiotou for the journey in the Ph.D.; thank you very much for your support and help all these years. I wish to thank Marilena Agathokleous, Savvas Zenonos, Zenonas Theodosiou, Argyris Argyrou, and all the other colleagues in the Ph.D., for their help and support. I want to thank Dr. Photis Psomopoulos for the help with the latex. Furthermore, I would like to thank my colleagues in the Library and Information Service of the Cyprus University of Technology for their help and support, especially the director Mr. Marios Zervas. I must thank everyone who took the time to fill out the questionnaires for different stages of the research. I also want to thank the examining committee members, Dr. Nikos Grammalidis and Dr. Klimis Ntalianis, for their time and precious comments. Of course, I wouldn't manage to succeed without the valuable assistance of my mother.

The thesis is dedicated to my father, who left us a few years ago.

Abstract

Billion digital images are uploaded every single day on the Internet and especially on social media. It is vital to develop effective and efficient methods that allow the retrieval of those images according to users' demands. Among the approaches that have been proposed for digital image retrieval is Automatic Image Annotation (AIA). AIA techniques automatically learn the visual representation of semantic concepts from a number of image samples, and use these concept models for tagging new images.

Learning good concept models requires representative pairs of image-tags. Manual annotation is a hard and time-consuming task since a large number of images are necessary to create effective concept models. Moreover, human judgment may contain errors and subjectivity. Therefore, it is highly desirable to find ways for automatically creating training examples, i.e., pairs of images and tags. Contemporary social media, such as Instagram, contain images and associated hashtags, providing a source of indirect annotation. Instagram is a photo-oriented social media platform where users upload images and describe them with hashtags; thus, it might be a rich source for automatically creating pairs of image-tags for AIA.

The thesis focuses on investigating Instagram images and hashtags as a field for AIA purposes. This primary research question is further analyzed through several studies: we define the portion of Instagram hashtags that are related to the visual content of images they accompany and we develop a methodology to locate stophashtags, i.e., common non-descriptive hashtags. We also employ the HITS algorithm in a crowdsourcing environment in order to filter Instagram hashtags and locate the ones that correspond to the visual content of Instagram images they accompany. Topic modelling of Instagram hashtags is introduced as a means for retrieving Instagram images in the traditional text-based information retrieval approach while transfer learning, utilizing filtered Instagram data (pairs of images and hashtags) is applied for a content-based image retrieval scenario.

Keywords: Machine Learning, Deep Learning, Automatic Image Annotation, Crowdtagging, Crowdsourcing, Instagram, Hashtags, HITS algorithm, Topic Modelling, Transfer Learning

Contents

List of Figures

Chapter 1

Introduction

Introduction contents

According to its inventor, the World Wide Web is a collection of documents where every user has access to the appropriate technology to search and locate documents or media. The Web also provides users with news and updates for anything they wish to know [8]. With Web 2.0, the next generation of Web, users can not only search for information but also provide their content, cooperate, communicate, and transfer knowledge. As a result of Web 2.0 technologies, digital data are growing every second, and the users face the so-called "information overload" problem. Intelligent search engines and recommender system techniques try to solve this problem. Contemporary Search Engines are pretty successful in text retrieval [9], but this is not the case for image (and multimedia information) retrieval where the availability of accurate and descriptive manually inserted metadata is of primary importance. So, image / multimedia retrieval is more complicated than text retrieval and still open for further research [10].

The traditional approach for image retrieval is text-based, and it was first introduced in the late 1970s when databases containing images with descriptive texts to perform effective image retrieval [11] were developed. Modern search engines still follow the text retrieval paradigm because they use techniques for document indexing. For this purpose, they have to relate images with specific keywords or textual description. Textual description of images is usually based on the web page, or the document, containing the corresponding photos and includes HTML alternative text, the file names of the pictures, captions, metadata tags, and surrounding text [12]. However, the assumption that the surrounding text is related to the nearby images and describes their visual content is not always valid. In addition, there are thousands of image collections / databases and millions of image files without related text (either container HTML file or surrounding text): photographs (old and new), medical images, etc. According to the known English proverb " a picture is worth a thousand words", but you need to locate this picture first to understand the words! And to locate this picture, you need words that describe its visual content! Manual annotation of these images is required to allow text-based retrieval. This, in turn, requires time, money, and huge effort in order for humans to annotate large image collections. Moreover, human perception of the visual content of an image, not owned by them, is highly subjective. The annotation of image content is also affected by people's background and mood [11].

Due to the previously mentioned difficulties on image retrieval, a new approach was introduced: Content-based image retrieval. In this approach, photos are retrieved based on visual based features such as colour, texture, shape, etc., and it is usually example based (query by example), that is an example of the picture we are looking for is given, and similar (in terms of low-level features) images are returned. However, this approach has two main drawbacks:

- Users are familiar with retrieving images based on text queries. Queries based on text describe the user needs at the semantic level instead of feature level.
- Queries based on text can better describe what the user wants to retrieve than content-based image retrieval. For example, imagine a user that is looking for images containing dogs. Suppose that the user submits, as an example for the system, a photo containing a dog and a tree. Based on that photo, the system can retrieve images containing only trees and not the dog, which is the subject the user wanted. Therefore, it is difficult for the system to understand what the user wants: a tree or a dog? In the case of text-based retrieval, a query such as "dogs" is clear, and the system can bring to the user the desired photos.

To overcome the gap between content-based retrieval, which is associated with low-level features, and humans that use high-level concepts for their search, Automatic Image Annotation (AIA) was proposed. In AIA, computer systems automatically assign tags in the form of captions or keywords to images [13].

1.1 Aims and Objectives

The current thesis focuses on the development of learning-based AIA techniques that use Instagram images and hashtags as training sets. It covers different aspects of current research in the area, including crowdsourcing, natural language processing, creation of training sets, and machine learning methodologies. It also proposes the idea of using a social media platform for automatic image annotation purposes: The purpose of AIA is to assign a few relevant words in a limited vocabulary to the images without labels [14]. Instagram is a photo-oriented social media platform where users share their images and videos and associate them with hashtags, words with hash symbol #, to describe their visual content but also their emotional state. So, in Instagram, we have images and annotations; this means that we can develop training sets to learn concept models, encoding visual representations, for AIA and then automatically use these concept models to assign keywords to other (unseen and non-annotated) photos. In summary: with the approach mentioned above, we can create databases with representative examples of image-tag pairs overcoming the problems (and need) of manual annotation.

1.2 Research Questions

The research conducted in the current thesis is mostly an empirical one and was formulated on the basis of seven axes expressed via the following research questions:

1. Is Instagram the best social media platform for research related to AIA? Why not, for example, Facebook or another social media platform?
2. Can we use Instagram hashtags to create image-tag pairs for training machine learning approaches for Automatic Image Annotation? What is the portion of Instagram hashtags that are related to the visual content of the associated images?
3. Can we identify Instagram hashtags that are common across images showing different and independent concepts, thus, being non-descriptive? Such hashtags should be filtered out for AIA purposes.
4. Can we identify descriptive Instagram hashtags with the aid of graph-based algorithms?
5. Can we find descriptive Instagram hashtags with the help of contemporary topic modeling techniques?
6. Is there any correlation between the color contents of Instagram images and their filtered hashtag sets?
7. Are the photo-tag pairs we managed to create after filtering out the irrelevant hashtags appropriate for creating training sets for AIA?

1.3 Contribution

A summary of the contributions of the current thesis, along with the related work published in scientific journals and in the proceedings of international conferences, is listed below:

1. We enter the idea of using social media platforms in AIA framework. In social media, users upload photos and annotate them with hashtags. So we can exploit it and create training data sets for AIA. The rational behind this is that the owner of a photo can describe it better than the experts. So we have quality data to use it for AIA.
2. An in-depth investigation of the hashttags that are related to the visual content of the images they accompany. Not all the hashtags the owners use for the images they post on Instagram are related to their visual content. A methodology to estimate the percentage of the hashtags that are related to the visual content of the accompanied images was proposed [15, 16].
3. Among the hashtags we noticed the same non-descriptive (i.e., #f4f, #instagood) in post from independent subjects/hashtags (i.e.#dog, #lion). We called these hashtags stophashtags, and we propose an innovative methodology to locate these hashtags that can be used not only on Instagram but in other social media like Twitter [17].
4. Using the graph theory, we propose a solid methodology for tag filtering, which applies not only in AIA but also in many other fields that pertain to direct or indirect interaction among two types of entities. Using the HITS algorithm in a crowdsourcing environment, we filtered irrelevant hashtags and kept only those related to the image's visual content. The experimental results showed high precision [18, 19].
5. Using the topic modelling approach for tag filtering. We propose a methodology that with the help of topic modeling we can identify the relevant hashtags for a category of Instagram photos [20–22].

1.4 Structure of the Thesis

In Chapter 2 we present the basic framework for the thesis, and we analyze the image retrieval methodologies focusing on automatic image retrieval. Chapter 3 provides a review of the literature covering the research questions of the thesis. In that chapter we give an overview of the existing literature and research related to hashtags, training datasets for automatic images annotation, common non-descriptive hashtags, HITS algorithm, topic modeling, image, and hashtag distance measure, and transfer learning. In Chapter 4 we analyze the methodology for the study, including a description of the research design for each research question. In Chapter 5 we describe the data collection and results for each research question. In Chapter 6 we summarize the significant findings of the thesis and propose future research directions.

1.5 Summary

In Chapter 1 we have detailed the aim and objectives of the thesis and we have introduced the seven research questions that guide the study. In Chapter 2 we analyse the contemporary image retrieval methodologies emphasizing on Automatic Image Annotation which is the central concept of the thesis. Moreover, we describe crowdsourcing, deep learning, and transfer learning concepts which are also related to our research. Finally, we analyze the reasons Instagram was chosen as a social media platform for Automatic Image Annotation.

Chapter 2

Theoretical background

Theoretical background contents

The World-Wide Web brings the global information universe to the user through the appropriate use of technology. The Web was invented in 1989 by Tim Berners - Lee, in an effort for people who worked in CERN to communicate and share information, equipment, and software across groups [23]. The easy use of the Web made it the most successful Internet service. The next generation of the Web, known as Web 2.0, provides the user with a dynamic environment. Web 2.0 allows users not only to search for information but also to create their content and interact with other users; that is, collaboration, communication, and information sharing.

The name Web 2.0 was first referred to by Darcy DiNucci in an article published in *Print Magazine*, but became popular from Tim O'Reilly in 2004 when he used it in a conference[1] and an article he pub-

[1] http://www.paulgraham.com/web20.html#f1n

lished [24]. In Web 2.0 we see the vision of McLuhan [25], Negroponte [26] and Dertouzos [27] for recreation of the world in a global village, becoming a reality [28].

Web 2.0 had, as a result, an impressive increase of online data and led to today's big data era where 2.5 Exabytes of data are produced per day [29]. Moreover, the advent of "Internet of Things"[2] contributed an even higher amount of data. During 2020 data production reached 44 zettabytes which is 10 times higher than the data (4.4 zettabytes) produced in 2013. These numbers show that every two years, the amount of data that is produced doubles. The most significant part of those data corresponds to visual information, including digital images.

2.1 Image retrieval

On the Web, the user can locate a variety of digital data forms: text, images, sounds, videos, and animations. The ability of current handheld devices, such as smartphones and digital cameras, to connect to the Internet and publish images and videos explains the fact that the amount of online digital images continuously increases. Users of social media such as Facebook, Instagram, and Twitter, upload and publish images at a breathtaking rate; the amount of images produced every day is inconceivable. As a result, effective and efficient image retrieval techniques and systems that respond to the user needs are of high importance, and the ongoing research in digital image retrieval is very active.

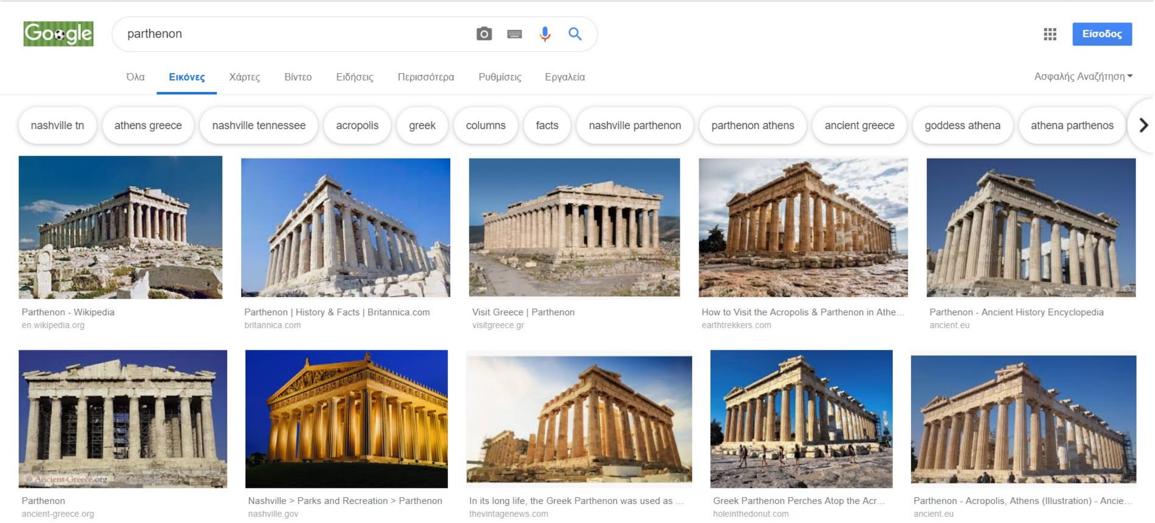


Figure 2.1: Text-based image retrieval. The query used for that example was *parthenon*

Manning *et al.* define information retrieval as: "Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" [30]. Image retrieval, on the other hand, is the process of trying to locate, from a repository, images that respond to the need of the user [2]. Image retrieval methods fall within two basic categories: text-based and content-based. The text-based techniques are inspired by document retrieval using keywords, while in content-based retrieval, an image is given as an example, and the system analyses the example (target) image and retrieves similar images according to its visual content. Fig. 2.1 shows an

[2] https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

example of text-based image retrieval with the term *parthenon*. Fig. 2.2 shows the target image we used for content-based image retrieval and Fig. 2.3 the result.

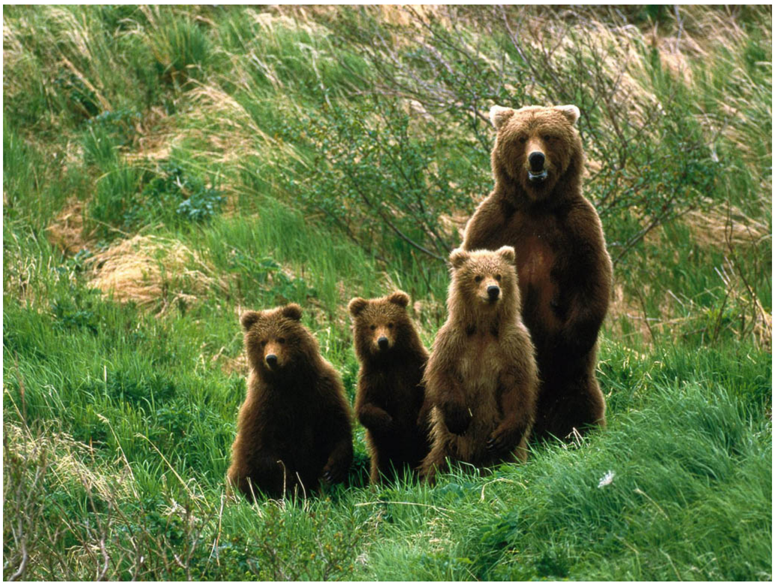
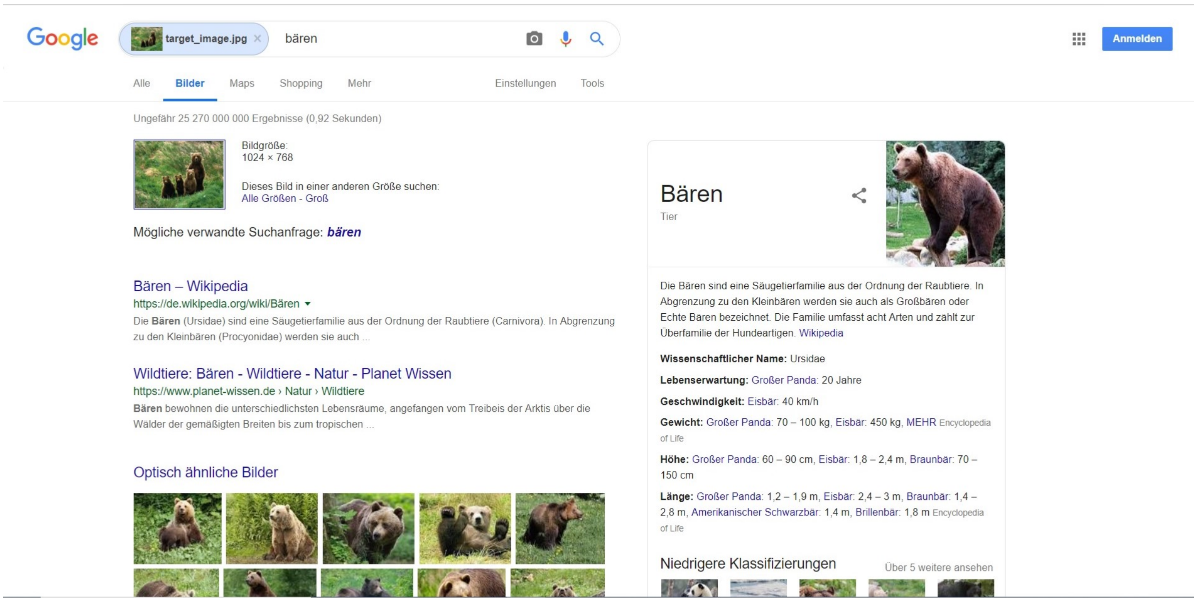


Figure 2.2: Target image for content based image retrieval



Ungefähr 25 270 000 000 Ergebnisse (0,92 Sekunden)

Bildgröße:
1024 × 768

Dieses Bild in einer anderen Größe suchen:
Alle Größen - Groß

Mögliche verwandte Suchanfrage: ***bären***

Bären – Wikipedia
https://de.wikipedia.org/wiki/Bären
Die **Bären** (Ursidae) sind eine Säugetierfamilie aus der Ordnung der Raubtiere (Carnivora). In Abgrenzung zu den Kleinbären (Procyonidae) werden sie auch ...

Wildtiere: Bären - Wildtiere - Natur - Planet Wissen
https://www.planet-wissen.de › Natur › Wildtiere
Bären bewohnen die unterschiedlichsten Lebensräume, angefangen vom Treibeis der Arktis über die Wälder der gemäßigten Breiten bis zum tropischen ...

Optisch ähnliche Bilder

Bären
Tier

Die Bären sind eine Säugetierfamilie aus der Ordnung der Raubtiere. In Abgrenzung zu den Kleinbären werden sie auch als Großbären oder Echte Bären bezeichnet. Die Familie umfasst acht Arten und zählt zur Überfamilie der Hundeartigen. Wikipedia

Wissenschaftlicher Name: Ursidae

Lebenserwartung: Großer Panda: 20 Jahre

Geschwindigkeit: Eisbär: 40 km/h

Gewicht: Großer Panda: 70 – 100 kg, Eisbär: 450 kg, MEHR Encyclopedia of Life

Höhe: Großer Panda: 60 – 90 cm, Eisbär: 1,8 – 2,4 m, Braunbär: 70 – 150 cm

Länge: Großer Panda: 1,2 – 1,9 m, Eisbär: 2,4 – 3 m, Braunbär: 1,4 – 2,8 m, Amerikanischer Schwarzbär: 1,4 m, Brillenbär: 1,8 m Encyclopedia of Life

Niedrigere Klassifizierungen
Über 5 weitere ansehen

Figure 2.3: Content based image retrieval based on Figure 2.2

.

Text-based systems provide quick response to users because they are based on string matching [31] even in the unlikely case that pre-computed indexing terms for the image do not exist. The text-based image retrieval approach is influenced by document retrieval, but, in contrary to documents/texts, the indexed terms are not part of the content, as in the text documents, but part of the annotation information/metadata. It is necessary, therefore, in order for the text-based approach to work, all images to be assigned metadata / tagging. The simplest way to do that is to let people manually annotate (add tagging information) the images, which means expense of time and effort. Manual annotation is influenced by the annotators and their perspective, background, and mood. In order to overcome the problems of manual annotation, researchers in the early 1990s introduced the idea of content-based image retrieval. In content-based techniques, images are indexed and retrieved by their visual content, usually based on low-level characteristics such as colour, texture, and shape [32]. As we can see in the example of Fig. 2.3 the results are indeed similar to the picture, but, assuming that the real interest is to find images of bears, the green background of the target image Fig 2.2 influences the results. This means that the system will not probably return images showing bears in a zoo, or images from bears that live as pets. On the other hand, if we search bear in text-based retrieval, it is more likely to retrieve bears in a variety of contexts, as we mentioned before.

Nevertheless, the text-based approach is the approach that users and search engines prefer because they are familiar with it. Users retrieve images and documents with text queries, while in content-based image retrieval, there is a lack of semantic meaning both in the query image but also in the indexing method. Furthermore, image examples are not always available, and in addition, as we show before, the examples may be interpreted differently by the system and the user. Search engines also prefer the text-based approach because it is based on well-established document retrieval techniques and has been achieving satisfactory results so far. In Fig. 2.4 we see the overall architecture of a text-based image retrieval system: the user enters a textual query, and the system tries to retrieve all images that are indexed (based on their metadata / tags through text assignment) with the terms used in the query.

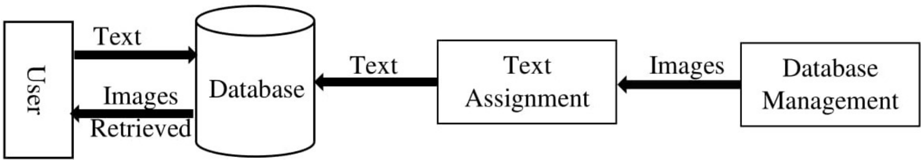


Figure 2.4: A typical text-based Image Retrieval system [1]

.

As already mentioned, in the text-based method it is necessary to relate images with keywords, annotation, or textual description. In the absence of manual annotation, search engines in order to locate text that describes an image are based on the web page of the document the photo is contained in. HTML alternative text, the file name of the image, caption, surrounding text, metadata tags, or words that are indexed from the web page [33–35] are used for this purpose. The assumption is that images are directly related to the document or web page they it appears in, which in some cases is not a valid assumption. Moreover, while this method applies to images in web pages or text documents, it does not apply to image collection databases. In the case of image collections, we can either employ content-based methods or manually annotate the images to allow text-based retrieval.

2.2 Content-based Image Retrieval

Content-based image retrieval (CBIR) or Query By Image Content (QBIC) refers to systems that take as a query an image and return to the user images that resemble the query (target) image. Thus, in the case of CBIR, instead of text, we use visual content (query image) to search for similar images. In order to describe (create "indexed terms") an image low-level features based on colour, texture, shape, and spatial location, are used. Each image is converted into a feature vector, which acts as an index for a stored image and as the query terms for the target image. Then, the system calculates the similarity (using distance metrics) between the feature vector of the query image and the feature vectors stored in the database images. In the last step, the system retrieves the relevant images using a similar ranking process in traditional search engines. In Fig. 2.5 the generic architecture of a CBIR system is presented.

Visual content search can be more effective in locating relevant images than text-based retrieval because it is more close to human perception of visual data. In addition to this, CBIR is not based on text, so we overcome the problem of image annotation or locating the relevant textual description in the web page or the document the image appears in. In several cases, to enhance the results of CBIR systems and achieve some kind of personalization, the systems exploit user's relevance feedback [31].

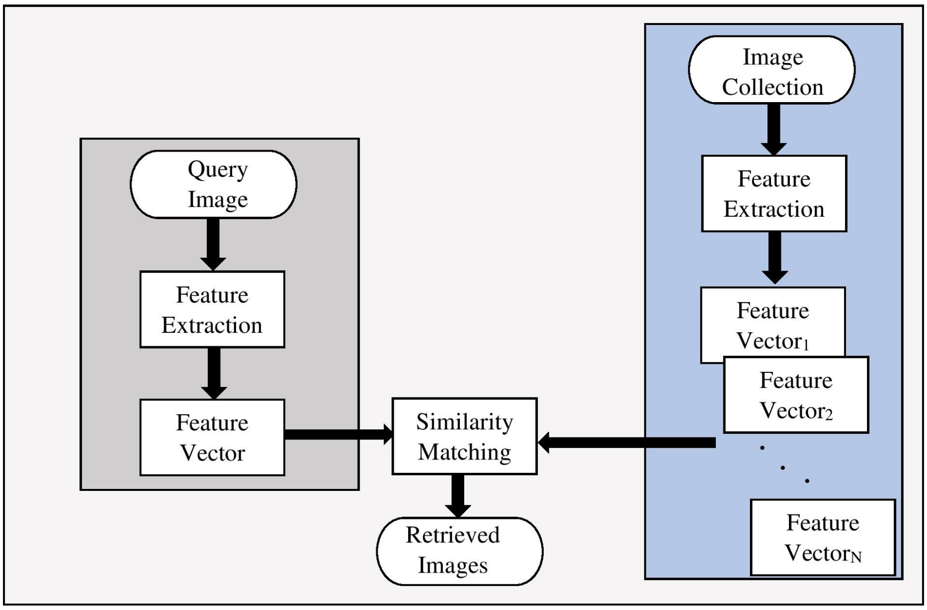


Figure 2.5: Content-based Image Retrieval [2].

2.2.1 Image segmentation

In text -based retrieval we do not search by providing a whole document as an example. Instead, we provide a query composed of specific terms and we are looking for documents containing or being relevant with that terms. In CBIR the query image may correspond to many abstract and non-abstract terms. For instance, a photo of a person in a beach may correspond to abstract terms such as "summer", "vacation", "relax", etc, and to non-abstract terms such as "beach", "human", "sea", "sand", "sky", etc. If such a query image is provided how the CBIR system will be able to correctly interpret the real need of the user? In an effort to tackle this problem, at least as far as the non-abstract terms is concerned, the researchers proposed to represent an image with several feature vectors corresponding to the homogeneous areas of the image [36]. The problem of partitioning an image into homogeneous (usually in terms of colour or texture) areas is known as *image segmentation* and the various areas are called *segments* [37].

According to Vartak and Mankar [38] "*Image segmentation is a process of extracting from the image domain one or more connected regions satisfying a uniformity (homogeneity) criterion which is based on feature(s) derived from spectral components*". The actual purpose of image segmentation is to detect objects in an image assuming that objects are univocally related with specific concepts [39]. However, even with the most advanced image segmentation techniques image segments rarely match to real objects.

Dey *et al.* [5] classifies image segmentation techniques into three categories: edge-based techniques which try to identify the borders between image areas based on intensity differences [40], region-based method that focus on the uniformity within a sub-region based on a specific feature like intensity, colour and texture [41,42], and the grid-based approaches, in which the image is divided into regular rectangular blocks. That method is independent of features like colour or texture and it is based only on the size of the block [43].

With the previous discussion we can conclude that content-based image retrieval has two main drawbacks:

1. a target image may not be available to the user or may not be suitable to express his/her needs,
2. locating relevant images is difficult because the system may interpret differently from the user the target image and thus, fail to understand what to search. For instance, if we enter in a system a photo showing a dog by the sea, the system could not easily understand if we want to locate images of dogs or sea.

2.3 Automatic Image Annotation

The aim of Automatic Image Annotation (AIA) methods is to automatically extract the visual content of pictures and then assign metadata / tags in the form of captioning or keywords to digital images. AIA approaches use learning algorithms along with pairs of image-tags to create the so-called “concept models” (i.e., visual representations of semantic terms). Then the trained concept models are used to assign keywords to unseen images through object detection/recognition and image classification methods [44]. Figure 2.6 summarizes the AIA process: In the offline stage images are assembled and annotated, feature vectors are extracted from them and used to train concept models. During the online stage, we enter a target image, extract the appropriate features and we locate similar images on the basis of the outputs of concept models. The target image is annotated on the basis of existing annotation of the retrieved (similar to target) images.

AIA techniques address some critical problems in image retrieval, including:

1. The need of manual image annotation for text-based image retrieval which is impractical due to the number of images produced every day, the time complexity of the manual annotation process, and the subjectivity of the annotators
2. the lack of surrounding textual information. Not all images can be related with text, for instance images in large database collections, medical images, etc.
3. the semantic gap between high-level concepts (keywords) and low-level features (e.g. image colour features).

AIA provides users with more qualitative and quantitative retrieval results. More images can be annotated automatically, so the number of relevant images is increased, increasing also the possibility to satisfy user needs. In addition, with the AIA, the user has the possibility to express his/her question with words, a method which users are familiar with, and the system can better understand than the query images in the content-based image retrieval.

The corresponding literature [13,44,45] suggests the classification of AIA methods into five categories:

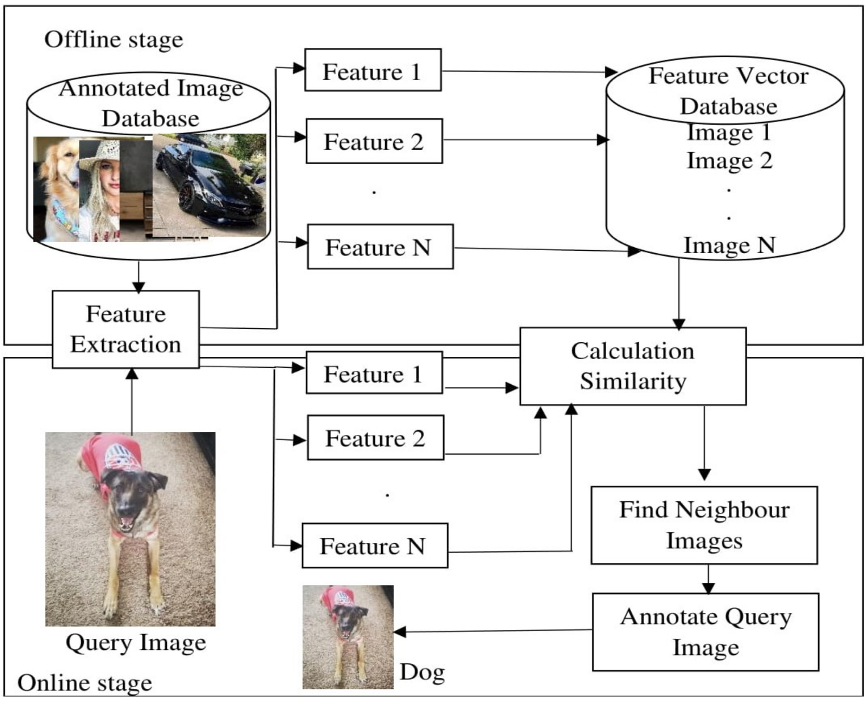


Figure 2.6: An example of Automatic Image Annotation [3].

1. *Generative / model-based*. The aim of this category of techniques is to maximize the common likelihood of image features and labels (tags). The generative model computes the joint probability of low-level features, non-tagged images, and the available tags in the training database to identify the tags that maximize that probability. Thresholding of the joint probability can also be applied so as to avoid annotation with irrelevant tags.

2. *Nearest neighbour methods*. The basic principle behind these techniques is that images that are visually "close" to each other (see Section 4.6.1 for more details) can be assigned the same tags.

3. *Discriminative models / detection models*. In this type of techniques, AIA is approached as a multi-label classification problem. In the first case (discriminative models), each untagged image is classified into one of a few categories with the aid of a properly trained multi-label classifier. Since scaling up (extending the number of categories to an arbitrary high number) is a severe problem of these methods, detection models were developed instead. In these methods, the learning algorithm creates a separate classifier for each keyword, and that classifier is used to predict ("detect") whether the target image belongs to the associated class (contains the modeled keyword) or not. All detected keywords are assigned as tags to the target untagged image.

4. *Tag completion methods*. In this type of method, we can achieve not only predict labels from images that are unlabeled but also correct noisy tags for given images. Moreover, an advance in tag completion methods is that missing tags can be filled automatically without training processes [46].In tag completion, the relationship between tags and images is represented as a matrix where the row and column represent images and tags. Then the algorithm learns the tag and visual similarities to assign new tags to images and correct noisy ones. [47,48].

5. *Deep learning methods*. Deep learning methods are also based on the learning by example paradigm;

in this respect, they are similar to both generative and detection models. The key difference is that in deep learning, selecting appropriate image features that will be used for training is not required. Convolution neural network structures are used to generate the visual features (see also Section 2.7) that feed the classifier. It is assumed that those visual features, with the help of deep learning techniques, effectively extract text information related to them based on the training examples. Deep learning methods seem very promising in the area of AIA, but they are not fully explored yet. Furthermore, they operate fully as black-boxes; it is practically impossible to justify why a tag is assigned to a target image. Nevertheless, deep learning techniques are state of the art and will be used in this thesis to explore the sixth research question (i.e., whether the image-hashtag pairs extracted from Instagram are appropriate for creating training sets for AIA).

2.4 Social media

Social media have a significant influence on modern society. People use them on a daily basis to post text and upload photos or other multimedia and, generally, share content that reflects their thoughts at the time of post / upload [49]. Social media platforms are also used for socialization, public debate, and information exchange. Social media first appeared in the mid-to-late 1990s when the users could create their websites through servers such as Geocities. At that time, blogging and social networks were launched. In 2002 Friendster was launched, and the resulting social media network became highly popular. Other social networking platforms such as MySpace, LinkedIn, iTunes, and the image-hosting website Flickr also appeared.

2.4.1 Facebook

The real revolution of social media networking took place with the advent of Facebook[3] in 2004. Facebook, founded by Mark Zuckerberg, and now has over 1.74 billion, active users. Zuckerberg, before Facebook, developed CourseMatch, a site that helped students register courses based on the selection of others, and Facemash, a site that allowed users to compare images of fellow students to socialize online. Inspired by the popularity of the sites mentioned above, Zuckerberg launched TheFaceBook.com on February 4, 2004. Facebook was first created at Harvard University and then managed to expand, reaching approximately 2 billion users [50].

In order to use Facebook, you have to register, and after that process, you can connect with other users, called “friends”. Facebook users can post text and upload images and multimedia and share them with their friends or publicly. The mission of Facebook is to let people have the power to build online communities and bring the world closer together. Facebook users can stay connected with friends and their family, despite the location of their residence, discover what’s happening in the world, and are able to share their thoughts[4].

[3] https://www.facebook.com/

[4] https://zephoria.com/top-15-valuable-facebook-statistics/

2.4.2 Twitter

Twitter[5], another popular social media network, appeared in 2006. In Twitter, users can construct a profile, follow other users, and post, limited in length, messages known as *tweets*. Initially, the length of a tweet was limited to 140 characters, but since September 2017, this limit was doubled to 280 characters [51]. The users also can upload images and videos both on Twitter or elsewhere (and share their links). Fig. 2.7 shows the anatomy of a tweet.

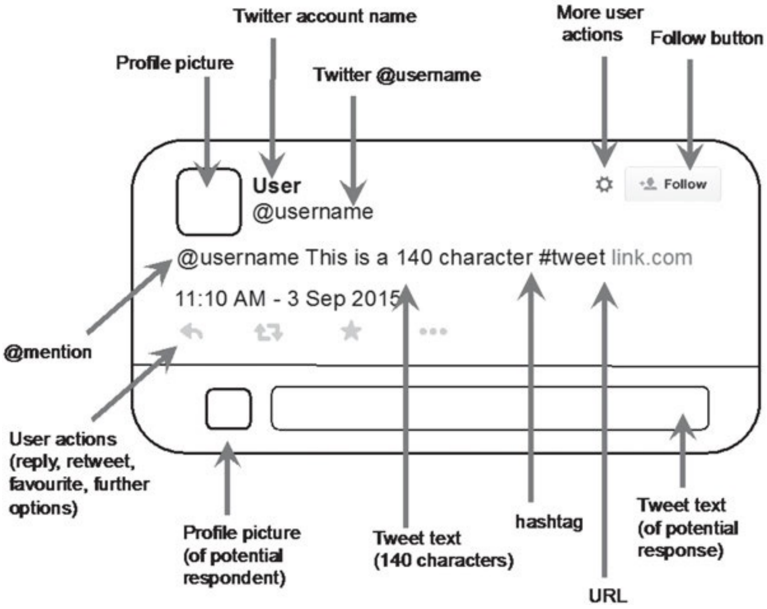

Twitter account name
More user actions
Follow button
Profile picture
Twitter @username
User
@username
Follow
@username This is a 140 character #tweet link.com
11:10 AM - 3 Sep 2015
@mention
User actions (reply, retweet, favourite, further options)
Profile picture (of potential respondent)
Tweet text (140 characters)
hashtag
Tweet text (of potential response)
URL

Figure 2.7: The anatomy of a tweet [4].

Twitter was founded by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams [52]. According to one of its founders, the definition of the word Twitter is “a short burst of inconsequential information” and “chirps from birds”[6]. The total number of active users in Twitter is 326 millions[7], but it plays an essential role in modern society because it is used as a major communication tool for politicians (including presidents and prime ministers of almost every country in the world), journalists and international organizations like EU and NATO. Twitter is also very popular among academics, and with the aid of Twitter API[8] many researchers have investigated Twitter from different scientific aspects such as sentiment analysis, event detection, and election forecasting.

Although collecting images from Twitter has the benefit of rich surrounding text through the body of the associated tweet or via possible comments and/or replies, the number of uploaded photos is quite low, compared to the number of tweets or with other social media platforms such as Facebook and Twitter. In addition, retrieving categories of Twitter images in one step is far more difficult than in Instagram.

[5] https://twitter.com/

[6] https://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html

[7] https://www.omnicoreagency.com/twitter-statistics/

[8] https://developer.twitter.com/

2.4.3 YouTube

YouTube[9] is a very popular video-sharing social media network. Users can upload, watch, review, share, create playlists, report, and comment on a video and have the possibility to subscribe to channels created by other users. The content of the videos varies largely: music, films, audio recordings, movie trailers, video from you-tubers, blogs, documentaries, educational presentations, etc. YouTubers are people who produce, upload, and share their videos on YouTube frequently.

YouTube was activated in February 2005 and was bought by Google in 2006 as an alternative medium to TV. Since then, it has become a part of the entertainment industry globally. The platform managed to gain 1.9 billion users per month, and the fact that more people prefer online video than TV is one of the YouTube effects. It is estimated that every minute users upload 400 hours of video [53, 54] on YouTube. YouTube content is also popular content for researchers working in the fields of image and video analysis and synthesis, computer vision, and many others corresponding to the entertainment industry. Nowadays, it is the primary research channel for multi-modal (combining video, audio/music, and text) video retrieval.

The aforementioned social media networks (Facebook, Twitter, Youtube) are based on the same idea of connecting people, i.e., they allow users to create a profile and build an online social network with their friends [55].

2.4.4 Flickr

Flickr[10] is one of the oldest social media that focuses on photo sharing. The purpose of Flick is to provide the user a place where they can upload their images and video to share them with friends and other users. Flickr users can make notes about the images they upload, subscribe to other users, and view their content. In addition, users can organize their images in albums of photos (formerly known as set) and categorize these albums into collections that can be part of higher-order collections [56].

Flickr was constructed by Stewart Butterfield and Caterina Fake. Flickr was a result of a game development named *Game Neverending*. A developer of the game added a tool that allowed players to share photos. That photo-sharing tool was developed independently and evolved into Flickr [57]. The innovation of Flickr was the interaction between users. Flickr went online on February 10, 2004, and by the end of that year, there were over 2 million photos on Flickr.

In Flickr, users can annotate their images with tags. Tags are keywords used to describe the photograph. These tags can help to organize and search images. Moreover, users can add geo-taggs in their photos that indicate where the photo was taken using location names.

2.4.5 Pinterest

Pinterest[11] is a free online mobile social-bookmarking website and social media service that focuses on images. Users can upload images they find online. In Pinterest, images are called pins and are linked from a website or uploaded [58]. The images are grouped in pinboards which are act as catalogs. Pinterest

[9] https://www.youtube.com/

[10] https://www.flickr.com/

[11] https://www.pinterest.com/

allows users to discover and save creative ideas expressed visually [59]. Users in Pinterest can add hashtags to their images since September 26, 2017.

Pinterest was developed by Ben Silbermann and Paul Sciarra with the help of Evan Sharp in 2009. Ben Silbermann, before Pinterest, developed a mobile shopping application with the name Tote. The purpose of Tote was to give users a personalized shopping experience presenting products from various retailers and locations. That project was not successful because it was too complicated, and Silbermann could not find someone to fund. However, Silbermann noticed that Tote users were collecting images of products. So, Silbermann and his team decided to create an image-based website. Pinterest in March 2010 was launched to a small group of people. After three months, the website managed to gain three thousand registered users. Pinterest growed exponentially the first few years of release and in September 2015 managed to have 100 million active monthly users, becoming the third most popular social networking site (at that time) behind Facebook and Twitter.

People use Pinterest to plan things they want to get or do in the future, to put in order things aspire for, such as dream car or dream vacation. Moreover, users can locate and organize solutions for their problems on Pinterest, such as making dinner and training their pets. In addition to these, users in Pinterest can find hobbies, discover something new to buy, or start and relax by browsing through the images [60].

2.4.6 Snapchat

Snapchat[12] is a mobile messaging application used to share photos, videos, drawings and text messages without any cost. In Snapchat the users are called Snapchatters and the picture and video a user sends and receives as a message to / from other user is called snap. The innovation of Snapchat is that the user can see the snaps only for a few seconds and after that they disappear. This is the reason Snapchat has a ghost symbol. The specific feature is the reason that app is so popular among young social users [61,62].

The founders of Snapchat were Evan Spiegel, Bobby Murphy, and Reggie Brown. The idea for Snapchat came up in the spring of 2011. The founders had the idea to create an application that someone send a photo and that photo disappear immediately after the receiver read the message[13]. The app went online on September 26, 2011. The initial name of the app was Picaboo. Within two years Snapchat became very popular and managed to gain 229 million daily active users, as of April 2020[14].

As mentioned earlier, the messages the users send in Spanchat are called snaps. In snaps the user can add emoticons, captions, filters. Moreover, in Snapchat users also have the opportunity to create stories, i.e., a collection of snaps that disappear in 24 hours [63]. Snapchat also provides the users a service called *Memories* that allows them to save snaps and story posts in their personal storage area.

2.4.7 Instagram

Instagram[15] is a social media network, initially targeted mobile devices, that allows users to share photos and videos. Its founders, Kevin Systrom and Mike Krieger, launched it on 6 October 2010, and rapidly gained popularity [64]. Their initial idea was to build a simple app that can inspire creativity when a

[12] https://www.snapchat.com/

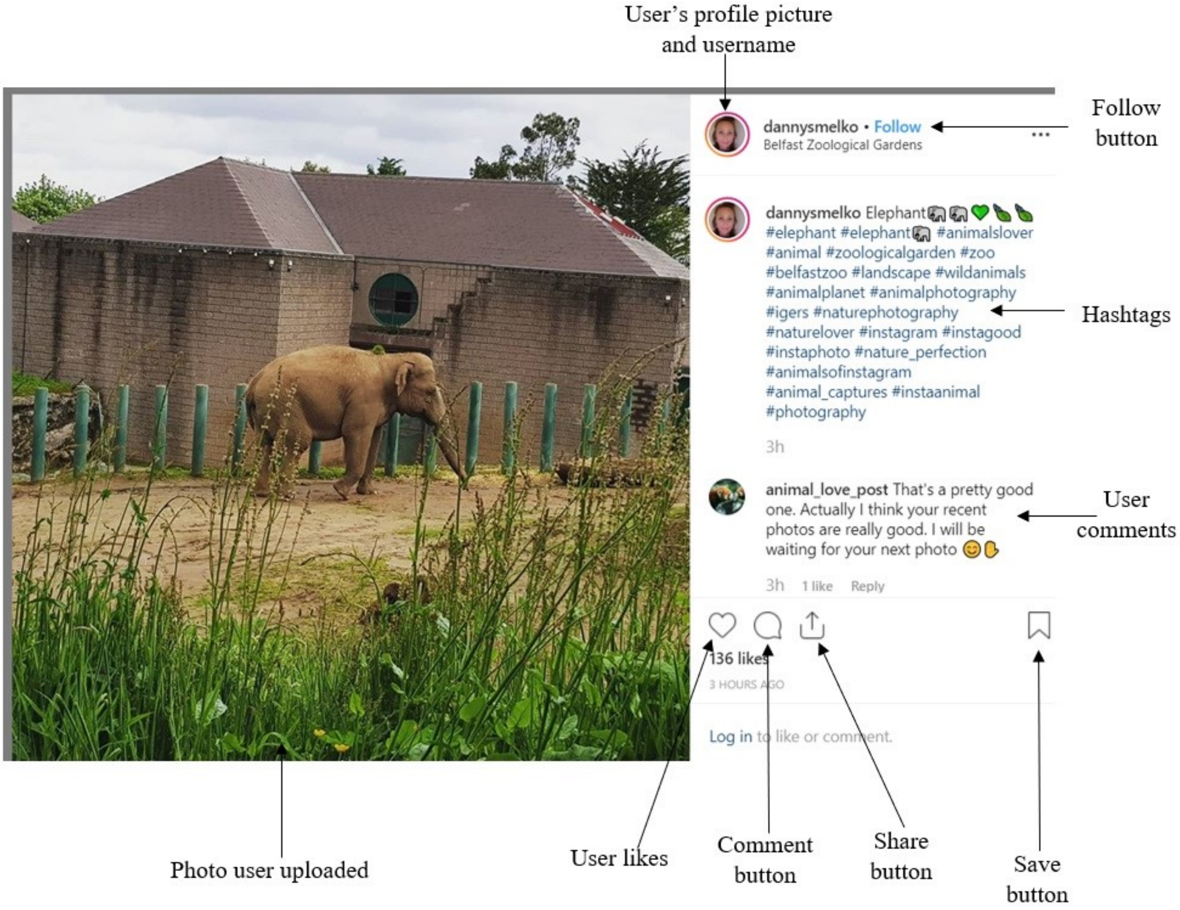
[13] https://www.businessinsider.com/snapchat-founders-lawsuit-internal-photos-texts-emails-2017-2?r=nordic

[14] https://investor.snap.com/news-releases/2020/04-21-2020-210949737

[15] https://www.instagram.com/

user captures everyday moments through the camera of his/her mobile phone. The Instagram developers chose to create a mobile application because Instagram has the purpose of allowing a user to produce photos on the go, in the real world and in real-time[16]. The term Instagram reflects their initial idea: it is a combination of the words “instant camera” and “telegram” [65]. The Instagram founders inspired it from a previous application they built which emphasized check-in at particular locations. The name of the application was *Burbn* and users could point locations and make plans for future places they were going to visit. Moreover, users of the Burbn app could win points for going out with friends and have the possibility to post pictures of the meets with those friends. That app was not so successful because it was very complicated, but its founders noticed that Burbn users used it to share photos. So after extended experimentation and development tests, they launched a simple photo-sharing application that they named Instagram [66]. The reason Kevin Systrom and Mike Krieger renamed it to Instagram is because they felt the purpose of the application was an instant telegram of sorts [67].

Instagram users have to register to the system, create their profile, and then they can upload pictures or short videos called stories. Instagram users also can follow other users and react to their posts (“like” them or comment on them). In Fig. 2.8 an example of an Instagram post is depicted.



dannysmelko • Follow
Belfast Zoological Gardens

dannysmelko Elephant🐘🐘💚🍃🍃 #elephant #elephant🐘 #animalslover #animal #zoologicalgarden #zoo #belfastzoo #landscape #wildanimals #animalplanet #animalphotography #igers #naturephotography #naturelover #instagram #instagood #instaphoto #nature_perfection #animalsofinstagram #animal_captures #instaanimal #photography

3h

animal_love_post That's a pretty good one. Actually I think your recent photos are really good. I will be waiting for your next photo 😊✋

3h 1 like Reply

Figure 2.8: The anatomy of an instagram post.

Instagram reached one billion active users with 40 billion photos, while 95 million posts are uploaded every day. It is the fastest growing social media network, especially among young people: 75% of Insta-

[16] https://instagram-press.com/blog/2013/02/05/introducing-your-instagram-feed-on-the-web/

gram users are aged between 18 and 24[17]. In January 2011 Instagram added hashtags [68] and from April 2015 users are able to use emoji as hashtags[18]. *Hashtags* are tags or words pre-pended with "#" 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.

Instagram managed to have so many active users due to the ability, offered by the application, to quickly and effectively process images, before uploading them. In Instagram, you have the possibility not only to capture and upload a photo but also to process (e.g. adjusting the contrast, brightness, and saturation [69]) it via a filter in order to achieve the desired result. Since August 2016, Instagram users could create and publish an Instagram story. Instagram stories allow users to take photos, add drawings, text, emojis, and swipe-able colour filters [70].

The maximum number of hashtags each Instagram image can contain is 30. Hashtags help other users to locate images they want; it is more likely to gain likes and comments on your posts if other users can easily locate your picture, so hashtags become popular and a lot of "advices" on the recommended use of hashtags to gain popularity appeared.

Hashtags are not new, neither on Twitter nor Instagram; users started to use them with the IRC (Internet Relay Chat) to categorize items into groups. The first who used hashtags in contemporary social media, especially on Twitter, was Chris Messina, a designer who asked his followers how they felt about using the pound sign to group conversations [71]. Thus, an essential role of hashtags was traditionally to organize knowledge, facilitate access and enable retrieval of information (see also the work of Small [72] on this).

2.5 Why Instagram

As we have seen in Section 2.4, there are many social media platforms focusing on images. So, it is essential to clarify why we chose Instagram instead of other social media. While this research question will be answered by carefully examining the literature devoted to AIA concerning significant media platforms, some initial hints and clues justifying the use of Instagram for AIA purposes are described below.

Facebook photos could be, indeed, used for AIA purposes since appropriate examples of photo-description pairs could be extracted. However, Facebook users are not very keen to use hashtags for image annotation. Instead, some text is usually associated with a photo. In this respect, retrieval of Facebook photos is similar to Web image retrieval. Thus, it is an application area of AIA and not a source for AIA training data. In addition to this, Facebook, and Twitter even less, is not a photo-oriented social medium since many users post text and/or video, participate in discussions, read news, etc.

While we could use Facebook API (Application Programming Interface) to develop applications that extract data from Facebook, it is not easy to search posts with a specific hashtag[19]. The result of those API applications is, usually, a graph in which nodes could be users, photos, pages, or comments, and edges

[17] https://www.brandwatch.com/blog/instagram-stats/

[18] Instagram: Our Story,https://instagram-press.com/our-story/

[19] http://www.socialmediainformer.com/api/facebook/hashtag/?open-article-id=8076187&article-title=what-the-facebook-and-instagram-api-changes-mean-for-you&blog-domain=agorapulse.com&blog-title=agora-pulse

could be photos on a page or comments on a photo. Thus, no explicit connection between photos and associated hashtags does exist; the closest combination being a photo-comment graph, which, as explained before, belongs to the general association photo-surrounding text, which is the case of web photos in general. The same argument also holds for Twitter API, which is far more flexible than Facebook API, but still does not support an obvious way for constructing *reliable* photo-hashtag association graphs.

Regarding Twitter and AIA training data, we should keep in mind that the portion of tweets containing text only is far more extensive than those consisting of images, videos, or gifs. As already explained, hashtags in Twitter aim to categorize tweets and facilitate tweet search and retrieval and not to tag photos[20]. While both Twitter and Instgram make extended use of hashtags, there is a fundamental difference between them: Instagram hashtags are the primary medium to search for images while in Twitter are used to annotate and help retrieve tweets; hashtags associated directly with a photo are less common.

YouTube is a platform whose primary aim is video (and music) sharing. The use of hashtags in YouTube is quite uncommon while the content of comments below every post is rarely relevant to the visual content of video; rather, in video clips, comments are usually related to the music/song in the video clip while in the other type of YouTube videos the comments are related to high-level video metadata (topic, creator, etc.).

Pinterest photos could also be considered for AIA purposes. However, Pinterest is a catalog of ideas that aim to motivate users to extent their creativity[21]. Adding hashtags to Pinterest images is a relatively new feature, and Pinterest users are not familiar with that functionality. Therefore, it is not easy, at the current stage, to locate representative pairs of photos-tags[22] for AIA training purposes. Pinterest photos are clustered into categories, which is, indeed, useful for AIA training data, and users can buy them[23]. That is, Pinterest image classification focuses on marketing purposes and for inspiring people's creativity. Therefore, we can conclude that neither Pinterest is currently an appropriate social medium for gathering AIA training. This situation, however, may change quite soon.

Snapchat also allows users to exchange pictures and video[24]. The role of hashtags in Snapchat is not the same as in other social media. Users can add hashtags, but Snapchat does not link images or videos that have the same hashtag[25]. Thus, in Snapchat, it is tough to mine pairs of images-tags, especially during the limited lifespan of a snap.

Flickr could be ideal for automatic image annotation because it contains images and tags. However, Flickr in the last years is not popular, almost unknown to social media users instead of Instagram, which is in the top ten popular social networks worldwide[26]. So, we can easily conclude that in Instagram, we can locate more images than Flickr necessary to create good training examples for automatic image annotation. In addition to this, we can locate research relating to the use of Flickr in the framework of automatic image annotation ([73–76]). On the contrary, we could not discover research focusing specifically on the use of Instagram for automatic image annotation purposes.

[20] https://www.lifewire.com/what-is-a-hashtag-on-twitter-3486592

[21] https://fortune.com/2015/07/13/pinterest-ceo-ben-silbermann/

[22] https://www.persuasion-nation.com/blog/pinterest-seo-best-practices-step-by-step-guide-2017-edition

[23] https://www.lifewire.com/top-social-networking-sites-people-are-using-3486554

[24] https://phys.org/news/2018-06-snapchat.html

[25] https://www.techjunkie.com/does-snapchat-use-hashtags/

[26] https://sproutsocial.com/insights/instagram-stats/

We conclude this section by referring to Table 2.1, where naive searches on the Scholar Google of the key terms of *AIA*, *photo(s)*, *hashtags* in conjunction with Facebook, Instagram and Twitter, the major social media platform, is presented. As already mentioned in the beginning of this section, the primary method to address the current research question (*Why Instagram?*) is via systematic and targeted literature review. We see in Table 2.1 that the percentage of Scholar indexed publications containing the pair of terms *Instagram* - *hashtags* is at least three times higher than the pair *Facebook* - *hashtags*, or the pair *Twitter* - *hashtags*. The difference is even higher if we consider the triples *Instagram* - *photo* - *hashtags*, *Facebook* - *photo* - *hashtags*, and *Twitter* - *photo* - *hashtags*. It is also interesting to note that the combination *Automatic Image Annotation* - *photo* - *hashtags* appears in 1.41% of the total publications in the field of Automatic Image Annotation (AIA) while among the publications containing the terms *AIA* and *hashtags* more than 80% contain also the term *photo*. It is even more interesting to mention that among the publications containing the terms *AIA* and *photos* 15.7% contain also the term *hashtags*.

	Total	+hashtags		+photos		+photo +hashtags	
		#	%	#	%	#	%
Facebook	6200K	41.8K	0.67	292K	4.71	16.2K	0.26
Instagram	1160K	24.9K	**2.15**	55.3K	**4.77**	11.0K	**0.95**
Twitter	7250K	52.1K	0.72	286K	3.94	16.6K	0.23
AIA	431K	7.66K	1.78	38.8K	9.00	6.1K	1.41

Table 2.1: Basic search in Scholar Google with key terms of the current thesis in accordance with three major social media platforms and the AIA term

2.6 Crowdourcing

The methods proposed in this thesis for gathering tags, filtering Instagram hashtags and identifying the relevance of Instagram hashtags to the visual content of the image they accompany make extended use of the ideas of crowdtagging. Thus, it is essential to explain the overall theory of crowdsourcing, emphasizing on its practical application and tools.

The term crowdsourcing is relatively new; it was first appeared in the June 2006 issue of *Wired magazine*, and by 2013 it was related to conducting business on the Internet [77]. Crowdsourcing can be defined as *the act of assigning a job, traditionally performed by an employee, to a large group of people in the form of an open call* [78]. We have to mention here, however, that in contemporary in crowdsourcing platforms like *Appen*[27] or *Amazon Mechanical Turk*[28], the call is usually done through the World Wide Web.

The platforms mentioned above take advantage of the "Wisdom of Crowd" theory, which suggests that groups can be more intelligent than the more thoughtful person in them under several circumstances. The theory was motivated by the early experiment of Francis Galton, who wanted to prove that the capabilities of the average voter were minimal. To test his assumption, Galton examined the eight hundred votes of a weight-judging competition about the weight of an ox in an International Exhibition. Galton was surprised

[27] https://appen.com/

[28] https://www.mturk.com/

when he discovered that the average crowd guess was very close to the weight of the ox [79].

Collective intelligence is a core element in Web 2.0 applications that collect reviews, comments, or tag content or boost participation in online communities to contribute to new content. Contemporary digital giants such as *Amazon, Netflix, Airbnb, IMDb* widely use collective intelligence in their everyday activities or to expand their market. Google, for instance, records and utilizes what users choose among the results presented to them by the search engine through a process known as "relevance feedback". The information gained through this process is used to adapt the retrieval results either in general or through personalization.

Collective intelligence is achieved for tasks in which a group of people expresses their free opinion or suggests a solution. It is not the right approach for tasks requiring heavy expertise, such as medical diagnosis tasks. In any case, three preconditions need to behold for a crowd to be wise and not mob [79, 80]:

1. *Diversity of opinion* each person needs to have private information even if it is just an eccentric interpretation of the known facts.
2. *Independence* It is vital to ensure that others' opinions do not influence people's opinions.
3. *Decentralization* People can specialize and draw on local knowledge
4. *Aggregation* Some mechanisms need to be applied to turn private judgments into a collective decision

Most of the contemporary online crowdsourcing platforms effectively facilitate the previously stated criteria for collective intelligence, which is why they are popular. The collective term *tagging*, usually expresses the overall process adopted to annotate and/or describe a digital item type such as images. With crowdtagging, a crowd is asked to tag digital objects [81] on the basis of some known motivation (payment, gaming, etc.).

2.7 Deep learning

It was already stated that machine learning, precisely artificial neural networks, is a key tool for AIA. For humans, learning is the process of gaining knowledge, understanding abstract concepts, and performed either by studying relevant information and getting instructions or by experience. Machines learn every time changes in data structures, software programs, or both take place [82]. For instance, the performance of automatic image recognition improves after analyzing several image examples showing a variety of cases (changes in data) based on mathematics or heuristic rules encoded in learning algorithms (software programs).

Machine learning has proven very successful in tasks involving pattern recognition and/or data classification. It has been effectively used in a variety of applications such as for object recognition, for supporting the doctors to diagnose patients, for automatically classifying data, or automatically separating materials [83], etc..

Artificial neural networks, a popular category of machine learning methods, are inspired by how neurons in the human body, especially in the brain, operate. Thus, in neural networks, neurons are simple information processors that can receive a signal, process it, and forward the result (output) to other neurons to which are interconnected.

In Fig.2.9 we show the structure of a neuron. The rectangular block responds to input data by multiplying them with the connection weights (W_i) and then thresholding the sum of the products, denoted as z in the diagram, via the F function (usually corresponding to a linear or sigmoid transformation). Neural networks usually contain hundreds of interconnected neurons organized in two or more layers; the layers between the inputs and the output(s) are known as hidden layers. Through proper training and appropriate training examples (pairs of inputs and outputs), neural networks are very effective in capturing hidden relations between the inputs and the output(s), creating models, i.e., encoded "rules", that predict the output(s) for never input data that have never seen before [5].

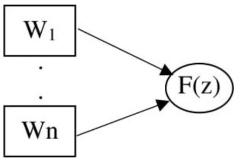


Figure 2.9: Structure of a neuron [5].

Deep learning (DL) belongs to machine learning methods and processes raw data in multiple layers to discover and distinguish high-level, abstract meanings [84]. One of the most common techniques in DL is Convolutional Neural Networks (CNN) [85]. The architecture of a CNN resembles a biological neuron, as we can see in Fig. 2.11. In the biological neuron, dendrites receive the signal; they process it and send it to other neurons. In the CNN shown in Fig.2.10 we have an input layer, where data are entered, and then the hidden layers in which input data are processed Fig.2.12. Finally, the output layer provides the confidence scores that high-level meanings (concepts such as 'dog', 'car', etc.) were detected in the raw inputs. The interconnection weights among the neurons within each layer and across layers are computed during the training phase.

In order to develop a machine learning-based AIA system, it is necessary to collect a large data set of images corresponding to the concepts we wish to model ('dog', 'car', etc.). In the training phase, the input to the DL architecture is images, and the output is a vector of scores, one for each concept modeled.

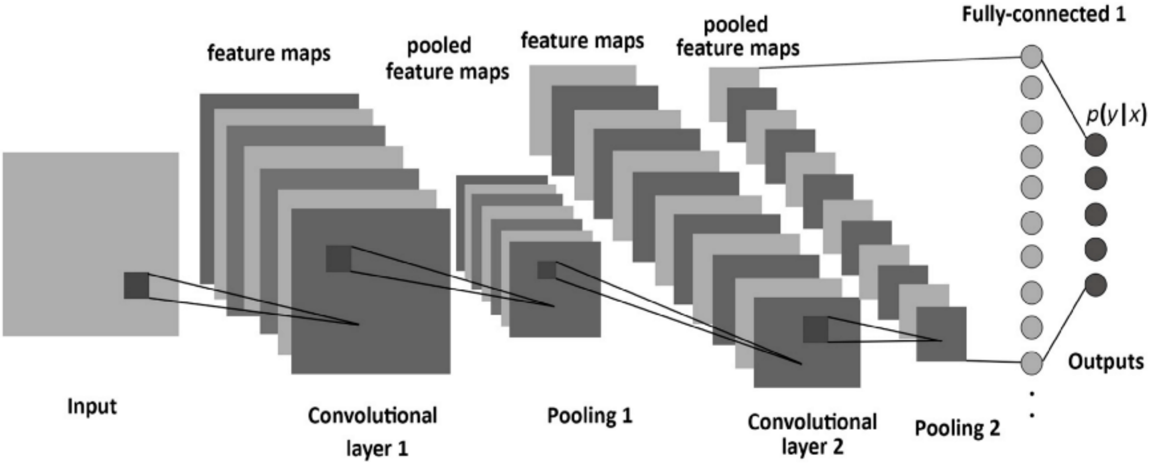


Figure 2.10: A diagram of Convolutional Neural Networks [6]

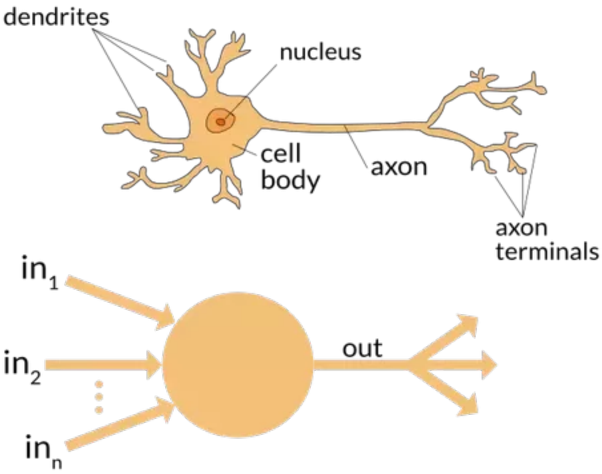


Figure 2.11: Artificial neurons are modeled to simulate the functionality of biological neurons [6]

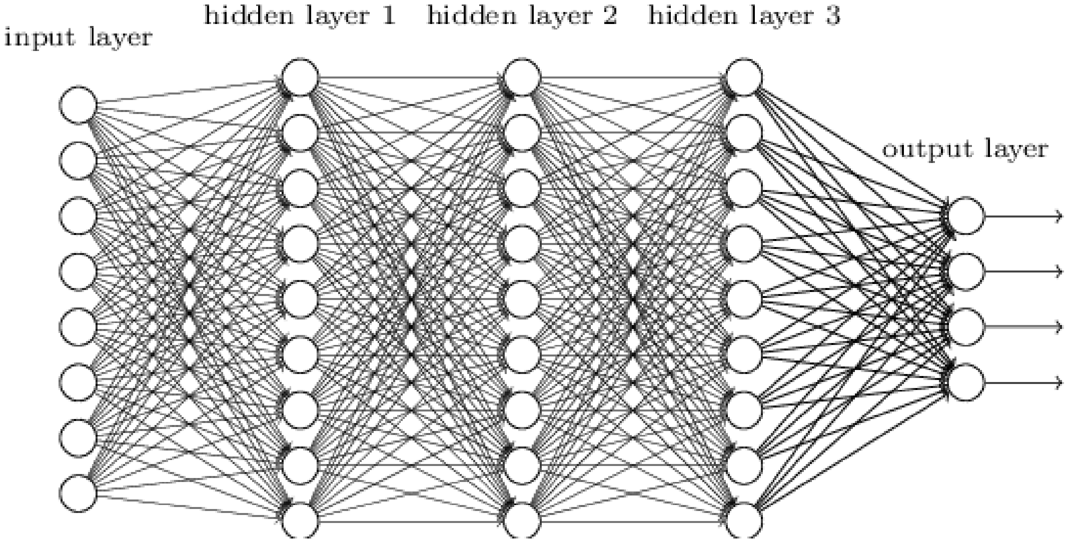


Figure 2.12: The hidden layers in Artificial Network [6]

During training, the goal is the output, say for concept ‘dog’, corresponding to the input image (i.e., an image showing a dog) to be higher in score than the other outputs, that is the outputs corresponding to other concepts. An objective function is used to calculate the difference (prediction error) between the desired output and the actual output to achieve that goal. The aim is the sum of the absolute values of prediction errors across all training data pairs (input-output) to be minimized. In order to do so, the adaptable parameters (i.e., the synaptic weights between neurons and the weights of input and output connections) of the DL architecture are changed. A deep-learning architecture can contain hundreds of millions of these adjustable weights and requires an equivalent amount of labeled examples (input-output pairs) to be properly trained. In order to overcome the necessity of so many training examples, stochastic gradient descent (SGD) is used. The SGD can perform with few examples because it uses the training data in several iterations [86] which in the terminology of machine learning are called *epochs*.

Liu *et al.* [87] tried to locate a method that could give food information from food images effectively and efficiently. The solution they suggest is based on Convolutional Neural Network, which automatically recognizes food images and gives information. To conduct their experiment, they used the UEC dataset that was developed DeepFoodCam project. This dataset contains a large volume of food categories with textual annotation. So, from the above, we can easily conclude that we can also implement Convolutional Neural Network in the case of image-tags from Instagram because we also have photos and text annotation.

Cheng *et al.* [13] presented a state-of-the-art machine learning method for AIA. They implemented AIA models in five databases (Corel 5K, ESP Game, IAPR TC-12, NUS-WIDE, MS-COCO) that are frequently used for assessment of AIA methods and concluded that deep learning-based AIA methods outperformed the other proposed techniques. The reason deep learning methods proved better than the other machine learning methods relies mainly on their ability to identify and acquire robust features from the input data, especially in complex and high-dimensional data types such as images and videos.

2.8 Transfer Learning

In the previous section we analyzed the importance of deep learning in AIA. As explained, traditional deep learning algorithms are using training and test data drawn from the same application domain, and, consequently feature space. Transfer learning methods are contemporary techniques that focus on storing knowledge gained while solving one problem and applying it to a different but related problem [88], [89]. The study of transfer learning is based on the human function of applying knowledge and experience learned previously to solve new problems faster or with better solutions. The researchers have focused more on transfer learning since 1995. In 2005 a new mission of transfer learning was set focusing on the ability of a system to understand and implement knowledge and skills learned in previous tasks to novel tasks [90].

Voulodimos *et al.* [91] in their review about deep learning for computer vision, point out that many pre-trained models used in transfer learning are constructed from large convolutional neural networks (CNN). In deep learning methods, especially in AIA, it is vital to collect images and train the models to classify images. That process is time-consuming and needs much effort. Transfer learning can help to reduce that effort. As Rawat and Wang [92] highlight, transfer learning is a standard method in computer vision because, with that techniques, we can build accurate models in a time-saving way. In Figure 2.13 we show an example of transfer learning for image annotation; in training from scratch method with the help of deep learning models, we enter the data, we train the CNN neural network, and as a result, we have the annotation of the image as a car. That process is time-consuming because it takes time for the machine to learn the model and produce the result. In the transfer learning process, we use pre-trained models, so we skip the training process, we enter the new data, and the model produces the results.

Ek [93], tried to categorize and annotate images from Finland's digital archives. To achieve that goal, Ek used pre-trained models on the ImageNet data set and managed to achieve 92.4% accuracy. Kieffer *et al.* [94] in their research they compare two methodologies in medical image classification based on Convolution Neural Networks. They conclude that pre-trained networks are quite competitive against training from scratch. Uricchio *et al.* [95] they propose a framework on automatic image annotation based on Kernel Canonical Correlation Analysis, which is used to build a latent semantic space where combine visual and textual features to annotate new images. The researchers in their architecture use pre-trained on ImageNet models. Singla*et al.* [96] they focus on food image classification and recognition.

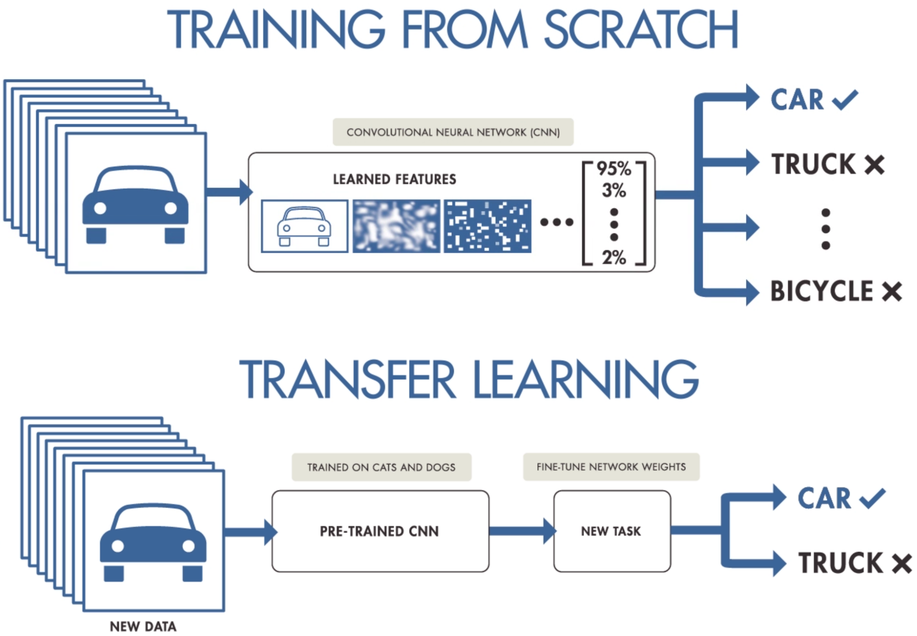


Figure 2.13: A diagram of Transfer Learning [7]

Using GoogLeNet pre-trained models, they managed to have high accuracy of 99.2% on food/non-food image classification and 83.6% on food categorization. Ma *et al.* [97] focusing on suggesting a hybrid method for thyroid nodule diagnosis they combine two pre-trained networks from ImageNet database. They managed to achieve high accuracy of 83.02% classification performance. Ashqar and Abu-Naser [98] in their effort to monitor invasive species, species that are not native to a specific location and with their tendency to spread can cause damage, they use the convolutional network and transfer learning. In their methodology to automatically locate invasive species from images, the researchers used pre-trained ImageNet, and they reached 99.71% accuracy.

From the above discussion, we can easily conclude that the use of transfer learning, especially using pre-trained models, can produce accurate results. So, we can explore transfer learning in our research to create training sets for AIA.

2.9 Summary

In Chapter 2, we have analyzed image retrieval methodologies, focusing on Automatic Image Annotation which we propose as a framework for our research. In addition, we analyzed the reasons Instagram was chosen as the data source of our empirical research, addressing also the first research question of the current thesis. In Chapter 3, we present a literature review spanning the research questions 2-7 (see Section 1.2).

Chapter 3

Literature Review

Literature review contents

3.1 Introduction

The current thesis focuses on effectiveness of creating learning-based AIA models using pairs of Instagram images and hashtags as training sets. In this chapter we conduct a literature review regarding the research questions 2-7 as set in Section 1.2). The chapter is divided into six sections: The first section provides a literature overview concerning the role of hashtags and the creation of training sets for AIA. The second section discusses the problem of identifying meaningless hashtags that are very popular on Instagram. The third section focuses on the possibility of using the HITS algorithm in a crowdtagging scenario to filter out irrelevant hashtags. The fourth section provides an overview of topic modeling approaches for locating relevant Instagram hashtags for developing AIA training sets. In this section we also address a literature review regarding the evaluation of topic models with the help of word clouds. The fifth section explores image similarity and word embeddings. In the sixth section, we conclude the current chapter with a literature review of transfer learning use in image classification.

3.2 Instagram hashtags and training set creation

On average on Instagram every day users share 86 million images[1] and while 350 million images per day are shared on Facebook[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).

3.2.1 Instagram Hashtags

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. As already mentioned (see Section 2.1), in text-based image retrieval images must be somehow related with specific keywords or textual description. 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) [99] emerged. AIA refers to the process of extracting low-level features from an image and assigning one or more semantic concepts (tags) to it [100].

A large portion of AIA methods involve machine learning techniques following the learning by example paradigm [2]. 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 [101]. 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 [102]. 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 (developing models for new concepts) and adaptability (modification of already learned models) issues.

It has been already mentioned that Instagram added hashtags in January 2011. 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 in Twitter, was Chris Messina, a designer, who asked from his followers how they felt about using the pound sign to group conversations [71]. 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 [72] on this). Tapastreet is 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 by harvesting location, time and hashtags [103]. It makes the assumption that hashtags are related to visual content of multimedia information. However, we know that users extend the function of hashtagging beyond findability and give hashtags a metacom-

[1] https://www.omnicoreagency.com/instagram-statistics/
[2] https://www.omnicoreagency.com/facebook-statistics/

municative use. According to Daer *et al.* [104] 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. Instagram hashtags, in addition, are also used for marketing purposes. Businesses use hashtags to raise user attention to their products and easily-track user generated content (UGC). Such examples include the hashtags *#loveloft* and *#worldsstrongestcoffee*. These hashtags are called branded hashtags and users, related to the corresponding marketing campaign through Instagram, use these hashtags on a variety of different photos [105].

Several researchers also suggest that hashtags carry emotional information [106] which is not directly related with the context they appear [107] in. 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 [108]. A study on gender difference in hashtag usage in Instagram for the hashtag 'Malaysianfood', revealed that women tend to use more emotional hashtags while men hashtags are more informative [109]. Ferrara *et al.* [110] 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 discussion 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 them carry significant descriptive information of their visual content. Thus, if we assume that it is the owner who can better expresses the real visual content or meaning of an image then choosing among the Instagram hashtags for assigning tags to images is much more safe than traditional text-based indexing approaches [111–113]. This is extremely important in training sets where pairs of images and tags have to be carefully selected because they affect the effectiveness of tag predicting models to be learned. However, Instagram 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. Thus, applying hashtag filtering approaches is necessary.

3.2.2 Creating training sets for AIA

Several approaches were proposed for creating training sets for AIA, such as (i) developing training datasets minded from the Web [114], (ii) use of the Flickr, a social network similar to Instagram, to construct image - tag pairs [115], (iii) getting advantage of clickthrough data and search logs in search engines to form image-tag pairs [116], (iv) combining linguistic description with visual data in order to achieve automatic image annotation [117] and (v) investigating the quality of manual image annotation [118]. In the following we examine the literature in these areas in more detail.

3.2.2.1 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 [119–121]. 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.* [114] 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.* [122] 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[3] a crowdsourcing Web service.

In an attempt to automate the image annotation process, NEIL (Never Ending Image Learner) [123], 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 [124] 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 identify shots to be used as training data for automatic detection of action concepts.

3.2.2.2 Image tagging with the aid of Flickr

According to the study of Sigurbjörnsson & Zwol [115] regarding users annotation on Flickr, users use only a few tags to annotate their photos and tend to annotate images according to their content. Ulges *et al.* [73] confirmed the results of Sigurbjörnsson & Zwol and proved also that users share, in the Web, images with specific structure and metadata.

Ntalianis *et al.* [125] 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.* [126] 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.* [127]. 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 stronger relationship they could identify and remove irrelevant tags.

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

3.2.2.3 Clickthrough approaches

Joachims *et al.* [116] 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 [128, 129]. 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 [130, 131], 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.* [132] examined the quality of clickthrough data for training concept detectors in images. They showed that clickthrough data, if properly filtered, could 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. Despite that, the use of clickthrough data for developing AIA models is an attractive approach and Microsoft Research announced, for three years in a row, a challenge based on data obtained from the Bing search engine[4].

Sarafis *et al.* [133], 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 [134] 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.

3.2.2.4 Visual and language data assignment techniques

Recently, several researchers started working on the alignment of visual and language data for image annotation. Karpathy *et al.* [117], in a notable work, investigated the relation between images and sentence description in order to produce novel sentence description of image regions. A dataset of images and sentence descriptions was used as input to a Multimodal Recurrent Neural Network aiming to learn generating description of image regions. Kiros *et al.* [135] built a log-bilinear model that generates phrase description from images. In their approach their model learn together word representation and image features with a help of Convolutional Neural Network. Their model relies on word representation for the image description based on high-level image features learned from deep neural networks. Johnson *et al.* [136] they propose a Fully Convolutional Localization Network architecture that can localize regions in an image and generate descriptions for those regions. In their model they managed generating meaningful description for regions of interest. Socher *et al.* [137] used supervised recursive neural networks in order to merge image segments or natural language words based on semantic transformations of their original features. WISE techniques [12] were also used to acquire rendered text from web pages, based on the idea that text which is close to an image in the HTML code is not necessarily close when a Web page is rendered.

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

3.2.2.5 Quality of manual image annotation

Several approaches deal with the quality of manual image annotation, especially under a crowdsourcing setting. Nowak & Ruger [138] 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 [118]. In order to examine the image retrieval from social media and especially the diversification of image retrieval results, Ionescu *et al.* [139] compared experts and crowdsourcing annotation. The results showed that in the crowdsourcing annotation the inter-rater agreement were a slightly lower than expert annotators. Veloso *et al.* [140] 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 more 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.* [141] 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.* [142], 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 [139]. The importance of crowdsourcing annotation lead 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 [143]. In a research about crowdsourcing annotators' consistency Theodosiou *et. al.* [144] used both vocabulary keywords and free keywords to check whether guided annotation (as assumed by the use of structured vocabulary) would increase annotation consistency. They concluded that, indeed, by combing free keywords and vocabulary keywords annotation consistency increases compared to the use of free keywords alone. Baba & Kashima [145] 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.* [146] developed a framework, called Requallo, in order to keep 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.* [147] 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.

3.2.2.6 The quality of crowdtagging

Image annotation by the crowd is a very popular trend nowadays. The validity of crowdsourced image annotation was examined and verified by several researchers. Mitry *et al.* [148] compared the accuracy of crowdsourced image classification with that of experts. They used 100 retinal fundus photography images selected by two experts. Each annotator was asked to classify 84 retinal images while the ability of annotators to correctly classify those images was first evaluated on 16 practice - training images. The study concluded that the performance of naive individuals to retinal image classifications was comparable to that of experts. Giuffrida *et al.* [149] measured the inconsistency among experienced and non-experienced users in that task of leaf counts in images of Arabidopsis Thaliana. According to their results everyday people can provide accurate leaf counts. Maier-Hein *et al.* [150] investigated the effectiveness of large-scale crowdsourcing on labelling endoscopic images and concluded that non-trained workers perform comparably to medical experts. Cabrall *et al.* [151] in their survey for drive scene categorization they used the crowd to annotate driving scene features such as presence of other road users and bicycles, pedestrians etc. They used the Crowdflower platform (now *Figure-eight*) in the categorization of large amounts of videos with diverse driving scene contents. As usual the Gold Test Questions in Crowdflower were used to verify that the annotators perform well in their job. The results indicated that crowdsourcing through the Crowdflower was effective in categorizing naturalistic driving scene contents.

3.3 Common non-descriptive hashtags

We have seen in previous section that some hashtags accompanying Instagram images are not descriptive. In Instagram there are hashtags that are very popular and users use them in their photos (see [152], [153]) just to draw attention, get likes and to become part of a group. In other words, users use these hashtags to annotate their photos regardless of what the picture show. The study of such hashtags received the attention of several researchers. From the AIA perspective, hashtags that are irrelevant to the visual content of the image they accompany should be filtered out.

According to Zhang *et al.* [154] the hashtag *#like4like* is appended to more than 290 million photos. This is clear case where a hashtag that appears in millions of Instagram photos is non descriptive. Unfortunately, this is the case for many other hashstags. For instance the *#photooftheday* hashtag appears in more than 500 million pictures while the *#instagood* hashtag is among the hashtags of more than 710 million photos (see https://top-hashtags.com/instagram/ for more recent statistics). Armano *et al.* [155], in their study aiming to locate stopwords in different document categories, propose specific metrics that capture the informative content of each term and measure their discrimination and characterisation capability. A rule of thumb is that "*a discriminating term has to distinguish a category against the others while a stopword has to be common all over the categories*".

Chua [156] *et al.* discovered that, approximately, 50% of the tags the Flickr users use to comment their images correspond to 'noise' (spam) hashtags while half of the 'true' labels are missing. Fan *et al.* [157] also dealt with the problem of spam hashtags in Flickr and developed an algorithm to clean spam tags through cross-modal tag cleansing and junk image filtering. Drewe [158], with the aid of Instagram API and a list of popular hashtags, created a list of unsearchable hashtags. This list, however, is unofficial, incomplete and needs regular update since it is created using ad-hoc processes and not a scientific methodology. Sedhai and Sun [159] in their effort to locate spam tweets concluded that 40% of spam tweets have three or more hashtags and it is more likely to use the word 'follow' as part of the tweet hashtags.

Yang and Lee [160] extract descriptive keywords from web pages and they measure the relatedness between web pages and tags in order to detect spam tags in social bookmarking. Tang *et al.* [161] in their effort to eliminate noise tags in a folksonomy system they propose a two-stage semantic-based method. First, they remove non-descriptive tags and then the semantic similarity between tags is examined in order to remove noise tags. Zhu *et al.* [162] in their approach for tag refinement they propose a form of convex optimization which considers, tag characteristics, error sparsity, content consistency and tag correlation.

3.4 Graph-based methods for filtering Instagram hashtags

In the previous section we have analyzed the common hashtags that are popular among Instagram users. Filtering these hashtags could be based with a lexicon based approach: Once common, non-descriptive hashtags are identified, a corresponding list will be created and mined hashtags contained in that list will be removed from the training data. The hashtag filtering problem can be also approached from a different perspective: We consider that image - hashtags (or in general image - tags) networks can be derived, either through crawling or via crowdtagging procedures, and that centrality measures, borrowed from the graph theory, can be applied so as to identify the best hashtags (tags) for each image. The underlying principle here is that a hashtag (tag) that many people argue that is relevant with an image is likely to be indeed relevant. After all, this is what the 'Wisdom of Crowds' theory tells us!

In order to address the current research question (see Section 1.2 research question 4) we envision the use of the HITS algorithm on graphs composed from image - hashtag associations. The Hyperlink-Induced Topic Search (HITS) algorithm provides the means to address centrality in bipartite networks. In the case of Instagram images we have users that tag pictures with hashtags and through this process they form an image-hashtag relation (an 'edge' in the terminology of graph theory). Thus, HITS is a ranking algorithm than we could use to filter Instagram hashtags and locate the most relevant one.

The purpose of HITS algorithm, developed by Jon Kleinberg, is to rate Web pages. The basic idea is that web page can provide information about a topic and also relevant links for a topic. Thus, web pages belong into 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 web page both a hub and an authoritative value [163]. In social network analysis the HITS algorithm, and specifically the hub and authority values it computes, is used for estimating the centrality of nodes especially in networks composed of two types of nodes, known as two-mode networks. A typical example of such networks are the bipartite networks which are usually modelled through bipartite graphs. A bipartite graph is a graph whose nodes can be divided into two distinctive groups (partitions) while its edges connect nodes among partitions but not within each partition [164, 165].

Two-mode (bipartite) networks are frequently used to model recommender systems [166], since consumers and products correspond to two different type of entities and usually the consumers choose or rate products. Mao *et al.* [167] applied HITS (and the PageRank as well) to improve user profiling in a social tagging system. The purpose of user profiling is to understand and code the personal interests of users so as to provide them advanced and personalized services. They modelled the social tagging system as a user-tag network and applied PageRank and HITS to refine the weights of tags. A diffusion process on the tag-item bipartite graph of the collection was then applied by using the estimated tag weights. The experiments, conducted on three different datasets, showed superiority of the proposed method over the traditional tag-based collaborative filtering approach that is usually adopted in recommender systems.

Zhang *et al.* [168] tried to extract people's opinions on features (characteristics) of electronic products such as mobile phones, tablets etc. In order to rank the importance of those characteristics they constructed a two-mode network where features were modelled as authorities and feature relevance indicators as hubs. With the aid of the HITS algorithm they were able to identify highly-relevant features and good feature indicators by thresholding the corresponding authority and hub values respectively. Nguyen and Jung [169] used a variation of the HITS algorithm, called GeoHITS, to rank locations with respect to specific tags such as those related with food types. Both tags and locations were collected from geotagged resources on social network services. The authors used a subset of tags that shared across several locations to act as hubs while the locations were considered as the authorities.

Cui *et al.* [170] proposed a healthcare fraud detection approach which is based on the trustworthiness of doctors to distinguish fraud cases from normal records. They created a doctor-patient two-mode network which was represented as a weighted bipartite graph. The prescription behavior in patients' healthcare records was used to compute the edge weights. According to the authors the hub scores of the HITS algorithm provide a good estimation of the trustworthiness of doctors. London and Csendes [171] applied a modified version of the HITS algorithm called Co-HITS to evaluate the professional skills of wine tasters. In order to achieve this goal, they constructed a weighted bipartite graph composed of wine tasters, modeled as hubs, and wines, modeled as authorities. The weights correspond to the scores given by the wine-tasters to wines. According to the authors, the computed hub values can be used to filter out incompetent tasters while they are highly correlated with the competence of wine tasters.

Tseng *et al.* [172] tried to distinguish fraudulent remote phone calls from normal ones by considering that the trust value of remote phone numbers is related with the hub score of the HITS algorithm. For that purpose they used telecommunication records to create directed bipartite graphs with incoming and outgoing calls between contact book entries of the users, assumed as authorities, and remote phone numbers (phone numbers not in contact books), assumed as hubs. The edge weights for each pair of user and remote phone number were computed based on duration and frequency relatedness between a user and a remote phone number. With the application of HITS the trust value for each remote phone number was computed and used to classify remote calls into fraudulent and normal.

There are also a few works in which the HITS algorithm was used in a crowdsourced environment, as we do in the current work for the specific case of image tagging. However, in the majority of cases the emphasis is put on the evaluation - enhancement of the quality of the crowdsourced data rather than to information mining. Sunahase *et al.* [173] applied the so called Pairwise HITS algorithm, a modification of the HITS algorithm which is applicable to pairwise comparisons, to three different tasks: image description, logo designing and article language translation. The aim was to estimate the quality of produced data and the ability of evaluators to assess those data through pairwise comparisons of image descriptions, logo designs and article translations created by two different creators - data producers. Schall *et al.* [174] tried to evaluate crowdsourcing participants (coordinators, supervisors and workers) used for business process. They created a two-mode social graph for each coordinator that processes a task from a customer. Supervisors, that separate the task into sub-tasks, and workers that perform the task, correspond to the two types of entities that compose the bipartite graph. The authority score is used to rank the performance of workers while the hub score is used to rank the effectiveness of supervisors to assign the right task to the right workers. Aydin *et al.* [175] tried to find the right answers to multiple-choice questions that had been aggregated from the crowd for the game “Who wants to be a millionaire?”. They created a big bipartite graph composed by multiple choice answers, assumed as authorities, and users, assumed as hubs. The computed hub scores, through the HITS algorithm, of the users were used as weights in a weighted voting scheme that predicts the right answer of a multiple choice question. The authors claimed a significantly increased accuracy of right prediction on the harder questions that are posed at the end of the game while the overall accuracy of prediction reaches 95%.

The structure of tuples {user, item, tags} in tagging systems has been termed folksonomy, being composed of folk, i.e., the users of the tagging system, and a taxonomy, i.e., a hierarchy is built from an “is-a” relationship. Traditional ranking algorithms such as the PageRank and HITS were proposed for ranking folksonomies [176]. However, the fact that folksonomies are composed from three different types of entities, and, therefore, can be only modelled as tripartite graphs, makes the direct application of those algorithms for ranking folksonomies problematic. As a result several modifications of the original PageRank and HITS algorithms were proposed. The FolkRank [177] is one of the algorithms that are based on the PageRank algorithm while a modification, called *differential FolkRank*, appropriate for ranking folksonomies that are modeled as uni-directed tripartite graphs was also proposed by the same authors [177].

There are different approaches in tag filtering including that of Xia *et al.* [127] who proposed a bi-layer clustering framework to locate relevant tags to social network images. In the first layers they try to locate relevant tags and images. In the second layer the image groups are divided into smaller using Affinity Propagation. Then they calculate the frequency and relevance of tags to keep only the relevant ones. Wang [178] *et al.* inspired by topic modelling and deep learning, proposed a method called regularized latent Dirichlet allocation to filter tags. In the deep learning model they use four layers combining tags and image features.

3.5 Topic modelling for Instagram hashtags filtering

In previous section we reviewed graph-based methods as an approach for filtering out non-descriptive Instagram hashtags. The graph-based method is based on the crowd and as a result can not be automated. In the current section we examine topic modelling as method for identifying relevant Instagram hashtags. Topic modeling is based on word probabilities. Words with higher probabilities in a corpus can give a good idea of what topics are discussed in that corpus [179]. Assuming that a corpus can be derived by compiling all hashtags appended to Instagram images retrieved via single hashtag query, we can use topic modelling find relevant hashtags to the query hashtag. This approach can be easily automated, because we can collect hashtags from relevant images, imply topic modelling and locate relevant hashtags.

Topic modelling algorithms use statistical analysis to discover the themes that best describe a collection of documents [180]. Thus, with topic model analysis, large archives of documents can be automatically tagged with thematic information. Topic modelling was applied on a variety of data sources. Below we concentrate on studies focusing on topic modelling applied on data crawled from social media platforms.

Rohani *et al.* [181] used topic models to extract topic facets from a dataset consisting of 90527 records related with the domain of aviation and airport management. The data were crawled within a period of 30 days from social media (the authors did not refer to the platform from which the data were collected). They developed an LDA topic modeling method while the data were pre-processed by removing punctuation and stop words. They identified five main topics and then they examined which one of the topics was the dominant in each date. The performance of topic modelling was qualitatively evaluated by domain experts who were asked to investigate the detected topics along with the discovered keywords and compare the results with their own interpretation about the top topics of the studied datasets.

Liu and Jansson [182] tried to identify city events from Instagram data. They created a dataset with posts, comments, and hashtags during the summer of 2016 from publicly accessible Instagram accounts in the Helsinki metropolitan region. Then, they applied the LDA topic modelling method to the set of relevant posts in order to discover clusters of targeted events. They thoroughly investigated the best number of topics that had been pursued and they concluded that as the numbers of topics increases the topics become heavily overlapped; thus, they decided to set an upper bound of topics (50) to their analysis. On the other hand, pursuing a small number of topics created non-coherent themes composed mainly from the most frequent words in the dataset. Instagram hashtags were kept during their analysis but only as a part of the container post / message. The authors do not provide any information on the pre-processing steps they applied nor on the way they evaluated the results of their analysis. In a newer analysis, the same authors (Liu and Jansson) [183] concluded that it is necessary to remove frequent non-topical terms, such as compliments, excitements or other positive tone and sentiments in order to bring up more novel topics. They examined, also, the importance of hashtags' presence in the Instagram posts and drew the conclusion that keeping hashtags in the analysis brings value into the mined topics.

In their effort to detect relevant content topics of pictures associated to a particular hashtag, Fiallos *et al.* [184] collected 7382 pictures associated with the hashtag #allyouneedisecuador. The aforementioned hashtag was created by a campaign entitled *"All you need is Ecuador"* as an effort to strengthen tourism in Ecuador. They calculated the similarity of topics mined from users description (hashtags and post text) and topics mined from visual analysis of the photos, called visual description. Visual descriptions were extracted with the aid of Microsoft Cognitive Services[5]. The visual descriptions produced 962 terms which after pre-processing were reduced to 838 while the users' descriptions initially produced 21972 terms and reduced to 18810 terms after the pre-processing stage. Topic modelling was applied to both description sets separately by combining TF-IDF with either the Non-Negative Matrix Factorization algorithm or the K-Means clustering algorithm. The authors discovered low similarity between the topics mined from the users description and the visual description and attributed this deviation to the fact users usually refer to situations or opinions regarding the photos while visual analysis produces tags more related with the actual content of the images.

Manikonda *et al.* [185] concluded that on Twitter you can locate informational content while on Instagram the content is more personal and social in nature. To reach this conclusion the researchers performed textual and visual analysis on the media content posted on these two platforms from the same set of users. For their textual analysis they used Latent Topic Models to extract topics users post on Instagram and Twitter with the aid of the Twitter-LDA API[6] which was developed for topic modeling of short text corpora to mine the latent topics [186]. The visual analysis targeted on the clustering of images using low level features (SURF features[7]) in an effort to investigate differences between the clusters created from the Instagram and Twitter photos.

Alkhodair *et al.* [187] tried to improve the performance of Twitter-LDA by combining it with the WordNet[8] and by including also hashtags in their analysis. They emphasized on the importance of different keywords to different topics based on the semantic relationships and the co-occurrences of keywords in hashtags. They also proposed a method to find the best number of topics to represent the text document collection. In order to evaluate the obtained results they used perplexity, topics coherence and users qualitative investigation of the mined topics.

The previous discussion shows that no other work so far dealt with topic modelling on Instagram hashtags, neither for extracting image tags nor for any other reason. In most cases Instagram (and Twitter) hashtags were used only as a part of the container post / message while in other were totally ignored. In addition two other important facts revealed from the literature review. The first is that LDA is the method of preference for most researchers regarding topic modelling. For short posts / messages such as tweets a the Twitter-LDA variation is usually used. The second, is that there is no common approach for evaluating topic modelling. Most researchers involve qualitative evaluation through user inspection while some others use quantitative metrics such as topic coherence.

3.5.1 Word Clouds

In Section 3.5 we have seen that the evaluation of topic modelling is mainly based on topic coherence. Nevertheless, evaluating the results of topic modelling is not an easy task. Several researchers approach (see [188], [189]), evaluated mined topics against humans performance. As Uglanova and Gius [189] mention human evaluation is still the gold standard in the evaluation of topic models. Topic models can be also seen as words clouds and interpreted by humans on the basis of this tool. In the following a short literature review on the use of word clouds is presented.

[5] https://azure.microsoft.com/en-us/services/cognitive-services/directory/vision/

[6] https://github.com/minghui/Twitter-LDA

[7] https://en.wikipedia.org/wiki/Speeded_up_robust_features

[8] https://wordnet.princeton.edu/

Word clouds is an informative data visualisation tool [190] primarily used to summarize textual information but it has been also applied for the analysis of social media data. Word clouds are used to depict word frequencies derived from a text or a set of text documents. The size of each depicted word in the cloud depends on its frequency: words that occur often are shown larger than words with rare appearance while stopwords are removed. Thus, a Word cloud can be seen as a synopsis of the main themes contained in textual information [191, 192]. Word clouds became popular in practical situations and are commonly used for summarizing a set of reviews presented as free texts (i.e., “open questions”).

In order to construct a classic word cloud it is necessary to calculate the word frequencies in a text or set of texts. However, word frequencies can be replaced by any other measure that reflects the importance of a word in a text document. In that respect word clouds can be used for the visualisation of topics derived from a collection of texts. Topic models infer probability distributions from frequency statistics, which can reflect co-occurrence relationships of words [193]. Through topic modeling we can reveal the subject of a document or a set of documents and present in a summarized fashion what the document(a) is / are about. This is why topic modeling is, nowadays, a state-of-the-art technique to organize, understand and summarize large collections of textual information [194]. Since with topic modelling we can measure the most relevant terms of a topic we can assume that by applying topic modelling on the hashtags sets [195] we can derive a set of terms best describing the set of Instagram photos grouped together within a subject.

Jin [196] used Twitter data about Hurricane Maria to identify and understand the main communication patterns of the related thread. She approached that problem in quantitative manner by topic modeling and word clouds to capture topics related to Hurricane Maria, and then, to qualitatively explain the results. Nogra analysed Instagram comments in order to locate words that are mentioned more frequently according to the media photo and visualised the results with word clouds [197]. The overall aim was to identify appropriate words to be associated with online product advertisements to better target possible customers.

In a study on how the Instagram is used to depict and portray breastfeeding, and how users share perspectives and information about that topic. Marcon *et al.* analysed 4089 images and 8331 corresponding comments posted with popular breastfeeding-related hashtags such as *#breastfeeding*, *#breastmilk*, *#breastisbest*, and *#normalizebreastfeeding*. They used word clouds to visualize the comment discussions in order to quickly identify the main discussion trends [198]. Vitale *et al.* [199] investigated how Igers (‘instagrammers’ which allow people who do not follow them to find their photos) represent themselves and their experience at museums in a textualised fashion. They analyzed the captions and hashtags of Igers’ Instagram photos and presented the most frequent words used in word clouds for quick interpretation.

Mittal *et al.* [200] study some user interaction properties, such as hashtags and post time, along with photo properties such as photo features or applied image filters to understand users’ engagements with Instagram posts. As a part of their analysis, they apply the Latent Dirichlet Allocation (LDA) algorithm in order to locate the most commonly used hashtags at a specific location. The most common hashtags per location are depicted as word clouds.

Kamil *et al.* [201] collected 1017 Instagram posts, tracked with the hashtag *#prayfornepal*, related to the Nepal earthquake in April 2015 to investigate how the people respond and express themselves emotionally for a disaster of such massive scale. By using posts’ date, time, geolocation, image, post ID, username and ID, caption, and associated hashtags they categorized the posts into seven categories and they created the word clouds for each one of those categories using the captions and the hashtags to visually illustrate the main topic facets related with the disaster.

In order to study the reactions of Instagram users on an Indonesian action entitled GERMAS, aiming to promote healthy living community movement, Habibi *et al.* [202] collected posts related to hashtag *#germas*. They applied topic modeling on the captions of those posts and used word clouds to illustrate the resulting topics. For topic modelling the authors used the Latent Dirichlet Allocation (LDA) algorithm.

3.6 Color histograms and hashtags sets

In contemporary Content-based Image Retrieval systems color-based features are quite common. Among them color-histograms are probably the most popular ones. Can we assume that the there is a correlation between the color histogram of an Instagram image and the associated hashtag set of that image? In this section we review the literature on color histograms and word embeddings which are the means we use in this thesis to investigate the previous question.

3.6.1 Color histograms and Bhattacharyya distance

In content-based retrieval, images are indexed by their visual content, usually based on low-level characteristics such as color, texture, shape, and spatial layout [203]. In practice, as in the case of search engines, color-based features are commonly adopted [204]. Among them, color histograms [205] are quite popular. Google Image[9], for instance, provides the users the ability to search images based on image examples (query images). Although the exact features that Google uses for the content-based retrieval are not known, if we compare the retrieved results with query images we can easily conclude that image similarity and ranking is based on color histogram comparisons. This conclusion is also evidenced by previous work: Takeishi *et al.* [206] compare the results of their content-based image retrieval system with that of Google image search. In this context, we assume that accurate estimation of color-based similarity is highly desirable for a variety of purposes such as image pre-indexing, easy creation of training sets in the Automatic Image Annotation (AIA) paradigm [12], and for establishing hybrid image retrieval methods.

Color-based low-level features are used for image classification and matching because of their effectiveness and ease of computation [14]. Color histogram is one of the methods that is used to extract low-level features. Color histograms are invariant to orientation and scale, and this property makes them more powerful for image classification [207]. Theodosiou [2] in his survey focuses on image retrieval using the AIA approach. In the proposed framework Theodosiou to extract low-level features uses spatial and color histograms to classify images. Zhang *et al.* [208] they propose an image retrieval algorithm that is based on color histograms. Mufarroha *et al.* [209] they build a system for content-based image retrieval that is based on color histogram and distance calculation between histograms to retrieve similar images. Alwan *et al.* [210] proposed a novel approach for identifying influential users on Instagram. They used color histograms in the RGB space to distinguish between influential and non-influential posts.

[9] https://www.google.com/imghp?hl=en

The approaches that calculate the similarity between images can be broadly divided into two categories on the basis of the metrics that are used. Intensity-based approaches are based on features (indices) derived from pixel color intensities while geometry-based approaches use geometric transformations between corresponding pixels [211, 212] in the compared pair of images. In intensity-based similarity computation, the metrics that are usually used are correlation, Chi-square, intersection, and Bhattacharrya [213, 214] distance. The geometry-based similarity metrics include Pixel Correspondence Metric, Closest Distance Metric, Figure of Merit, and Partial Hausdorff Distance Metric [212, 215]. The main drawback of geometry-based similarity metrics is their high computational cost. Thus, intensity-based metrics, usually involving histogram and histogram matching, are frequently adopted. While a variety of metrics is used for histogram matching, the most common metric is Bhattacharrya [213, 214] distance.

Bhattacharyya distance was widely used for computing the low-level content similarity of two images, video frames or image regions, in a variety of purposes. Chacon-Quesada and Siles-Canales [216] adopted Bhattacharyya distance as a metric for shot classification of soccer videos. They evaluated eight different histogram distance metrics and they concluded that Bhattacharyya distance is the best among them. Ong *et al.* [217] in their effort to develop a moving target tracking algorithm, used color histograms of frame regions to locate the target object in each frame. Abidi *et al.* [218], in their vision-based robot control system, use histogram of oriented gradients (HoGs) and minimize the Bhattacharyya distance between two sets of gradient orientations expressing the desired and current camera poses. Doulah and Sazonov [219] cluster food-related images using Bhattacharyya similarity. The images are extracted from meal video captured with a wearable camera and they are indexed using histograms in the HSV color space.

3.6.2 Word Embeddings

In the previous section, we discussed the distance between two color histograms and related research work. In this section, we discuss the second key technology dealing with the question we posed in the beginnig of Section 3.6, i.e., the word embeddings, focusing on their use in the context of Instagram hashtags.

Word embeddings, i.e., techniques that convert words to numerical vectors retaining semantic and syntactic information, is a state-of-the-art approach in natural language processing, especially in document classification, sentiment analysis [220] and topic modelling [221]. Word embedding techniques learn the relation between words via training on context examples of each word [222] using deep learning methods. The most common used word embeddings are GloVe [223], Word2vec [224], and WordRank [225]. Pre-trained word embeddings for every word in a variety of languages are available online and this is boosted their application on an impressive number of different fields and purposes.

Weston *et al.* [226] used a convolutional neural network to create specific word embeddings for hashtags. The overall aim was to predict hashtags from the text of an Instagram post. Liu and Jansson [182] tried to identify city events from Instagram posts and hashtags. They used Word2vec embeddings for query expansion, i.e., to identify terms related to the seed posts they used. Hammar *et al.* [227] classified Intastagam text posts into clothing categories using word embeddings. They used similarity matching via word embeddings to map text to their predefined ontology terms.

Prabowo and Purwarianti developed a system that helps online shop owners to response to Instagram comments [228]. The system classifies the comments to those that are necessary to answer, those that the online shop owner needs to read, and those to ignore. By comparing Support Vector Machines (SVMs) and Convolutional Neural Netowks (CNNs) they concluded that the combination of word embeddings with CNN learning provides the best combination.

Akbar Septiandri and Wibisono [229] used Word2vec to detect spam comments on Indonesian Instagram posts. They used the *fastText* library which allows easy expansion of word matching to short-text (paragraph) matching. Serafimov *et al.* [230] proposed hashtag recommendation for online posts using word to paragraph matching with the aid Word2vec vectors. Gomez *et al.* [223], based on Instagram posts related to the city of Barcelona, combine images and caption to learn the relations between images, words and neighborhoods. To achieve their goal they used pretrained Gensim Word2Vec models to discover the words the users relate with Barcelona's neighborhoods. Xu *et al.* [231] used word embeddings to locate relevant documents in an information filtering system. In their model create topic models based on documents of user interests. Then a topic model is estimated for incoming documents and the relevance of the document is estimated to determine if the document is relevant for the user oe not.

3.7 Transfer learning and image classification

Color histograms effectively capture the visual content of an image especially in the contest of AIA. However, the fact that color histograms do not retain spatial information leads to cases where images showing very different concepts (such as sky and sea) but still have very similar color histograms. In recent years, several researchers have focused their efforts for AIA on deep learning [232] which became nowadays the state of the art approach in the field. Cheng *et al.* [13] in their survey of different AIA methods concluded that deep learning-based approaches show the best performance. Below we review some recent studies that successfully applied transfer learning for image classification.

Abdullah and Hasan [233] in their study to classify 200 images in five categories, namely: Binoculars, Planes, Faces of people, Watches, and Motorbikes, used the AlexNet Model, a pre-trained CNN, and they concluded that with the help of that pre-trained model achieved improved accuracy of classification. Chaib *et al.* [234] used two pre-trained Convolutional Neural Networks (VGG-Net and CaffeNet) for the classification of Very High-Resolution (VHR) satellite images. They compare their method with other state-of-the-art methods, and they concluded that their transfer-learning based methodology outperforms the other state-of-the-art methods. Shima [235] also applied transfer learning with the aid of Alexa-Net pre-trained models. They investigated object classification, using the STL-10 database, for ten classes, and they achieved an average of 84.38% test-set accuracy. Nasiri *et al.* [236] examined automatic identification of grapevine cultivar by leaf image using pre-trained ImageNet models, in transfer learning context. They reported an overall accuracy of 99.11% on six grapevine cultivars. Taheri-Garavand *et al.* [237] proposed a transfer-learning based methodology for chickpea variety identification and discrimination. Four commercial chickpea varieties (Adel, Arman, Azad, and Saral) were used in their experiments. They used pre-trained ImageNet models to fine-tune their models and they reached an average classification accuracy of over 94%.

Tsapatsoulis and Diakoumopoulou [238] applied transfer learning for face verification on the specific case of old Greek actor Leonidas Arniotis using the ResNet-50 model, trained on VGGFace2 dataset and fine-tuned using confirmed photos and video frames of the actor. They concluded that transfer learning can be applied for face verification but special attention must be given when comparing images of highly different resolutions. In that case, the learned embeddings seem to mainly capture global image characteristics rather than specific facial characteristics that can discriminate different people.

3.8 Summary

In Chapter 3, we have presented an extended literature review spanning the six of the seven research questions of the current thesis. Several studies referring to the use of Instagram hashtags, and manual and crowdsourcing annotation for the formation of training datasets for AIA purposes were presented. We have also seen filtering methods that can be applied on Instagram hashtags emphasizing on graph-based methods. Topic modelling methods applied on social media data and verified through word clouds were also reviewed. Application of color histograms and word embeddings for AIA were presented while the literature review was concluded by reporting some recent methods of transfer-learning applied for image and object classification.

In Chapter 4, we formulate the mathematical background of the methods applied in the context of the current thesis while we explain the methodology we followed.

Chapter 4

Methodology and Mathematical Formulation

Methodology contents

4.1 Introduction

The purpose of this chapter is to explain the research design and methods adopted in the current study to answer the research questions set (see Section 1.2). We first describe the methodology that we followed to investigate whether Instagram hashtags accompanying each image describe the visual content of the image (and, thus, can be used for AIA purposes). Then, we deal with the problem of identifying and remove stophashtags, i.e., common non descriptive hashtags. Graph theory and relevant methodologies for hashtag filtering are presented next. The topic modelling approach for hashtag filtering and image retrieval is, then, covered and discussed. The methodology we followed to identify possible correlation between visual content and (filtered) hashtag sets is then presented, including a discussion for color histograms creation and matching. The chapter concludes with a discussion of the method we used for transfer learning.

4.2 Examining if Instagram hashtags describe the visual content of images

The purpose of this section is to analyze the methodology we followed to investigate if hashtags are appropriate for AIA. We further analyze that research question into two specific 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.

The basic assumption underlying this methodology is 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. 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). An advantage of using Instagram data for training AIA methods is that 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.

In order to examine the aforementioned goals we adopt a quantitative methodology based on an online questionnaire. Through the online questionnaire we examine if participants would choose the owner hashtags to annotate the image rather than random hashtags. With the choice of online questionnaire we increase the number of participants, reduce the time required to fill in the questionnaire and avoid fatigue effects. For the analysis of results we follow 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 organize 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 among students as we can conclude from the survey of Pew Research Internet Project [239]. 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. In addition the, AIA tries to address the problem of automatically assigning tags to images to be used by the contemporary search engines. So, to evaluate the descriptive power of hashtags we choose the aforementioned users that search in search engines daily.

4.2.1 Mathematical formulation

The current research question will be answered via an empirical study. Thus, a solid mathematical framework for evaluating the results, in terms of retrieval effectiveness measures such as Recall, Precision and F-score [240], is necessary. As it can be seen below those measures are adapted to the specific study.

Let us denote with P^i the *i-th* participant (*i=1,...,*N_P) of a total of N_P study participants. We also denote with I^j the*j-th* image (*j=1,...,*N_I) in the image dataset where N_I is the total number of images annotated at least from one user. By set $H = \{h_1, h_2, ..., h_{N_H}\}$ 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.

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:

$$R_{ij} = \frac{||\mathbf{T}_{jc} \cap \mathbf{T}_{ji}||}{||\mathbf{T}_{jc}||} \tag{4.1}$$

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. 4.1 across all images participant P^i annotated we get the overall per participant recall value:

$$R_i = \frac{\sum_{j=1}^{N_I} \left\| \mathbf{T}_{jc} \cap \mathbf{T}_{ji} \right\|}{\sum_{j=1, T_{ji} \neq \emptyset}^{N_I} \left\| \mathbf{T}_{jc} \right\|} \tag{4.2}$$

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. 4.3:

$$R_j = \frac{\sum_{i=1}^{N_P} \left\| \mathbf{T}_{jc} \cap \mathbf{T}_{ji} \right\|}{N_P^j \cdot \left\| \mathbf{T}_{jc} \right\|} \tag{4.3}$$

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

In a similar manner we define per image (see eq. 4.5) and per participant precision (see eq. 4.6), 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:

$$P_{ij} = \frac{||\mathbf{T}_{jc} \cap \mathbf{T}_{ji}||}{||\mathbf{T}_{ji}||} \tag{4.4}$$

(precision of participant's P^i choices for j-th image)

$$P_j = \frac{\sum_{i=1}^{N_P} \left\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\right\|}{\sum_{i=1}^{N_P} \left\|\mathbf{T}_{ji}\right\|} \tag{4.5}$$

$$P_i = \frac{\sum_{j=1}^{N_I} \left\|\mathbf{T}_{jc} \cap \mathbf{T}_{ji}\right\|}{\sum_{j=1}^{N_I} \left\|\mathbf{T}_{ji}\right\|} \tag{4.6}$$

$$F_j = \frac{2 \cdot R_j \cdot P_j}{R_j + P_j} \tag{4.7}$$

$$F_i = \frac{2 \cdot R_i \cdot P_i}{R_i + P_i} \tag{4.8}$$

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^ican 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 choice of Hamming distance [241] is evident. Hamming distance compares binary strings of equal length and outputs the number of positions at which they differ [241].

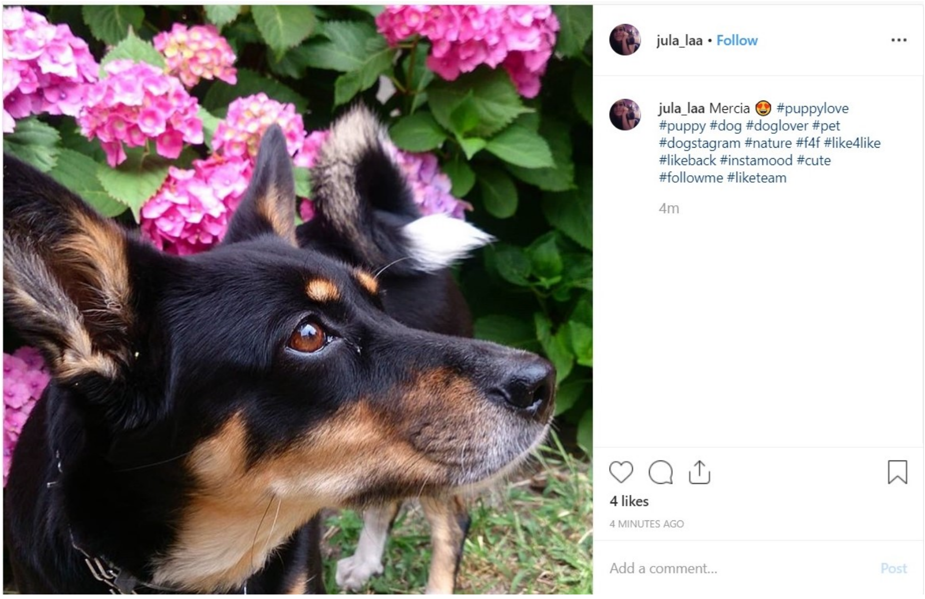
Thus, the similarity $S(C, P^i)$ between the choices a participant P^i made in order to characterize the images in the questionnaire with the actual hashtags the owners used, is given by:

$$S(C, P^i) = 1 - \frac{h(\vec{C}, \vec{P^i})}{L} \tag{4.9}$$

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 $\vec{V}$).

4.3 Locating Stophashtags

AIA techniques calculate the correlation between image features and textual words from example datasets in order to predict keywords for unseen images. The first step of AIA, is, therefore, the creation of good training examples, i.e., pairs of images and relevant tags. So hashtag filtering is essential for AIA purposes. Common hashtags among irrelevant images or non-descriptive words add noise to the machine learning process and could lead to non-representative concept models. In an analogy to document indexing, where stopwords are removed [242], so we can suggest the term 'stophashtags' to those hashtags. Thus, 'stophashtags' are usually meaningless hashtags that appear quite frequently in entirely different image categories [155].



jula_laa Mercia 😍 #puppylove #puppy #dog #doglover #pet #dogstagram #nature #f4f #like4like #likeback #instamood #cute #followme #liketeam

Figure 4.1: An example of Instagram image along with its hashtags. We can locate meaningless hashtags such as #f4f, #like4like, #likeback, #instamood, #cute, #followme, and #liketeam

Figure 4.1 shows an example of an image in Instagram depicting hashtags, among which some are related to the visual content of the image and some are irrelevant. The post includes the picture of a dog (to the left) along with the hashtags the creator/owner used to annotate it (the rigt part). Hashtags like #puppylove, #puppy, #dog, #doglover, #pet, #dogstagram, and #nature can be considered as descriptive of the visual content, but hashtags such as #f4f, #like4like, #likeback, #instamood, #followme, and #liketeam are used exclusively in a metacommunicative manner and clearly lack any descriptive value. These hashtags can be considered as stophashtags since they appear in many, visually irrelevant, images in a similar manner as common words (e.g. conjunction words or articles) appear in irrelevant documents. In document classification these words are considered stopwords and are discarded.

4.3.1 Methodology for locating stophashtags

As in the previous research question, it is important to define a mathematical framework which will allow us to proceed with 'stophashtags' identification and removal, as well as for effectiveness evaluation. The aims is to propose an algorithm for calculating a hashtag score based on which we can locate stophashtags. From the definition we have given (theoretically) to stophashtags, i.e., meaningless hashtags that appear in irrelevant image categories, we can devise the basic principles of our algorithm: We could select N subjects / hashtags that are thematic irrelevant (e.g.'dog' and 'tower') and collect all images related to that subjects. We can easily assume that common hashtags appear in different categories as mentioned before, they are not related with images' visual content, so these hashtags are not descriptive. Our proposal is an algorithm encompassing six steps:

1. Create a list of N independent 'subject' hashtags. Independence can be assessed on the basis of Word2Vec embeddings[1] or similar tools.
2. For each 'subject' hashtag H retrieve relevant images. We consider, therefore, here that images retrieved under the same 'subject' hashtag form an image category.
3. For each image I collect all hashtags appended to it.
4. For each 'subject' hahstag H compute a stophashtag score S_H
5. Compute a threshold T of S_H scores with the aid of Otsu method
6. If $S_H > T$ add to H Stophashtag List.

To the best of our knowledge this is the first research that attempts to locate meaningless hashtags in Instagram for automating the process of hashtag selection for AIA.

4.3.2 The stophashtag score

The mathematical formulation of stophashtag score is the key element in the above algorithm. We denote with N_H the number of subjects/hashtags that contain a specific 'subject' hashtag H, while N denotes the number of randomly selected 'subject' hashtags. It is essential to mention that the selection of 'subject' hashtags must be done with great caution because the subjects have to be independent (non-correlated) with each other. We denote, also, with I_H the number of distinct photos we locate through the 'subject' hashtag H and with I the total number of photos retrieved through all N 'subject' hashtags. In order to formally estimate the likelihood of a hashtag to be a stophashtag we define a score, S_H, that combines the normalized subject frequency of a hashtag, S_F, and the normalized image frequency of a hashtag, I_F, as follows:

$$S_F = \frac{N_H - 1}{N - 1} \tag{4.10}$$

where $N > 2$. We use $N_H - 1$ in the nominator of eq.4.10 to indicate that the 'subject' hashtag used to retrieve the photos within a subject (let's call it retrieval keyword) is, by definition, relevant to the content of photos retrieved within this particular subject and cannot be considered as stophashtag. For instance, assume that one of the N subjects/hashtags used to collect photos is the hashtag *#car*. To estimate the normalized subject score for the hashtag *#car* itself, we should count its frequency of appearance in the remaining subjects excluding the one it was used as a retrieval keyword. In this respect, the number of independent subjects (used in the denominator of eq.4.10) is $N - 1$.

[1] https://radimrehurek.com/gensim/models/word2vec.html

Algorithm 1 Locating Stophashtags

```
procedure Hashtags($H_{List}$)
/* $H_{List} = \{H_1, H_2, ..., H_N\}$: $N$ independent hashtags given by the user */
    $H \leftarrow []$                                                    ▷ $H$: an empty hash table
    for $H_i$ in $H_{List}$ do:
        $H[H_i] \leftarrow$ List of URLs of Instagram images tagged with $H_i$
        for $I_j$ in $H[H_i]$ do:
            $H[H_i][I_j] \leftarrow$ list of hashtags of image $I_j$
    return $H$

procedure HashtagLists($H, H_{List}$)
/* $H$: hash table obtained by procedure Hashtags */
/* $H_{List}$: List of $N$ independent hashtags given by the user */
/* $h[s] \subseteq H_{List}$: List of hashtag topics in whose images hashtag $s$ appears in */
/* $p[s]$: List of images, obtained by $H_{List}$ topics, that have $s$ as a hashtag in */
    $S \leftarrow$ set of distinct hashtags in hash table $H$
    $P \leftarrow$ set of distinct image URLs in hash table $H$
    for $s$ in $S$ do:
        $h[s] \leftarrow \{\}, p[s] \leftarrow \{\}$
        for $H_i$ in $H_{List}$ do:
            for $I_j$ in $H[H_i]$ do:
                if $s \in H[H_i][I_j]$ then
                    if $I_j \notin p[s]$ then
                        $p[s] \leftarrow p[s] + \{I_j\}$
                    if $H_i \notin h[s]$ then
                        $h[s] \leftarrow h[s] + \{H_i\}$
    return $h, p, S, P$

procedure HashtagScore($h, p, S, P, H_{List}$)
/* $H_{List}$: List of $N$ independent hashtags given by the user */
/*$S$: set of distinct hashtags in hash table $H$ */
/* $h$: Hash table showing for each hashtag the list of hashtag topics it appears in*/
/* $p$: Hash table showing for each hashtag the list of images it appears in*/
    $N \leftarrow$ length of $H_{List}$                              ▷ Total number of topics / subjects
    $I \leftarrow$ length of $P$                          ▷ Total number of distinct image URLs retrieved
    $S_H \leftarrow []; S_F \leftarrow []; I_F \leftarrow []; a_N \leftarrow 0.8$          ▷ $a_N$: a weighting factor indicated in eq.1
    for $s$ in $S$ do:
        $N_H \leftarrow length(h[s]); I_H \leftarrow length(p[s])$
        $S_F[s] \leftarrow (N_H - 1)/(N - 1)$
        if $N_H == 1$ then
            $I_F[s] \leftarrow 0$
        else
            $I_F[s] \leftarrow I_H/I$
        $S_H[s] \leftarrow a_N \cdot S_F[s] + (1 - a_N) \cdot I_F[s]$
    return $S_H, S_F, I_F$

procedure StopHashtagList($S_H, S$)
/*$S_H$: list of hashtag scores */
/*$S$: set of distinct hashtags in hash table $H$ */
    $SHL \leftarrow \{\}$                                              ▷ Empty stophashtag list
    $T \leftarrow$ Otsu threshold of $S_H$ list values
    for $s$ in $S$ do:
        if $S_H > T$ then
            $SHL \leftarrow SHL + \{s\}$
    return $SHL$
```

$$I_F = \begin{cases} 0, N_H = 1 \\ \\ \frac{I_H}{I}, N_H > 1 \end{cases} \tag{4.11}$$

$$S_H = a_N \cdot S_F + (1 - a_N) \cdot I_F \tag{4.12}$$

where $0 < a_N < 1$ is a weighting factor indicating the relative importance of normalized subject and image frequencies. Recommended values, found through experimentation, for a_N range in the interval [0.75 1).

4.4 Graph-based methodology for filtering Instagram hashtags

As we have seen in Section 3.4 the HITS algorithm has been successfully applied in real-world problems that can be modeled through bipartite graphs. At the same time crowdsourced image annotation is gaining popularity through the wide use of dedicated crowdsourcing platforms. The conceptual framework of the proposed method is presented in Fig. 4.2. So, we apply HITS algorithm in order to identify image tags in a crowdsourcing environment. To the best of our knowledge, this is the first work that suggests the HITS algorithm in order to identify image tags in a crowdsourcing environment. Moreover, we can locate the appropriate hashtags in each image with the proposed methodology. Furthermore, we use the crowd's wisdom to annotate images considered close to an expert.

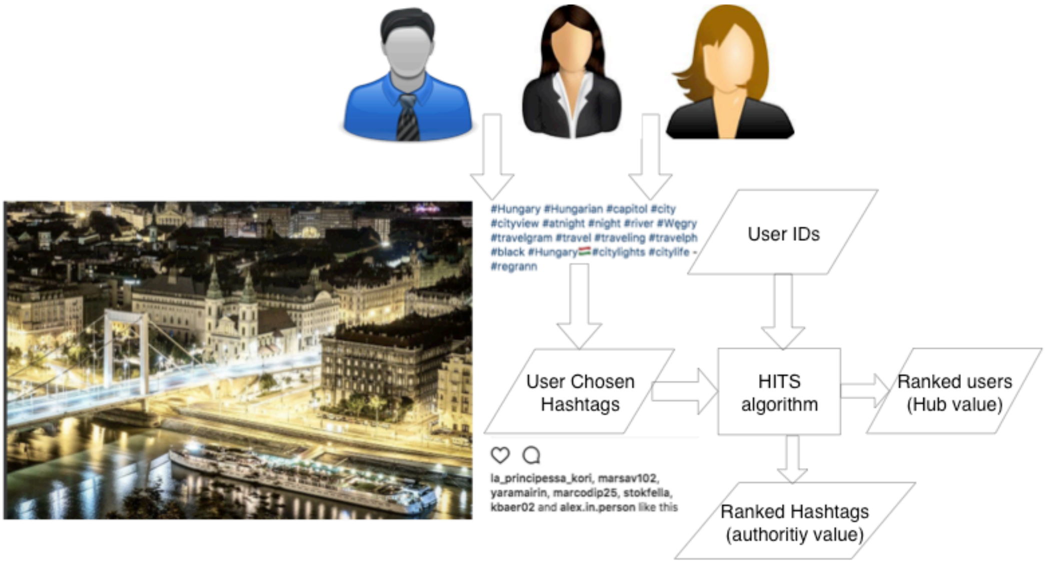

#Hungary #Hungarian #capitol #city #cityview #atnight #night #river #Węgry #travelgram #travel #traveling #travelph #black #Hungary #citylights #citylife - #regrann
User IDs
User Chosen Hashtags
HITS algorithm
Ranked users (Hub value)
la_principessa_kori, marsav102, yaramairin, marcodip25, stokfella, kbaer02 and alex.in.person like this
Ranked Hashtags (authoritiy value)

Figure 4.2: Conceptual framework for tag selection through HITS in a crowdsourcing scenario

A two-stage search was conducted. The first stage is based on data from the online questionnaire we used in the previous search (see Section 4.2). In the first stage we apply the HITS algorithm per image since we created bipartite network for a single image. So, as we see in Fig. 4.2 the reliability and performance of annotators is reflected on the hub value while the suitability of each hashtag to be used as image tag is reflected on the authority value are calculated per image. In the second stage we applied the HITS algorithm in data collected from a crowdsourcing platform. The problem of crowdsourced image tagging has never been modeled as a two-mode network probably because it involves three different types of entities: annotators, images and tags. We overcome the three entities problem by applying the HITS algorithm in two consecutive steps and on two different bipartite graphs. We first estimate the reliability of annotators by utilizing the hub value of the full bipartite graph consisting of the annotators and the tags they selected-used across all images. Then the annotator hub values are used as tie-weights on bipartite graphs constructed per Instagram image. The authority values of the tags, computed through the HITS algorithm, give us a ranking in terms of relevance between the hashtags and the image they accompany and is used to filter out the relevant from the irrelevant hashtags. Moreover, in the second stage FolkRank is used as baseline to evaluate the performance of the proposed method.

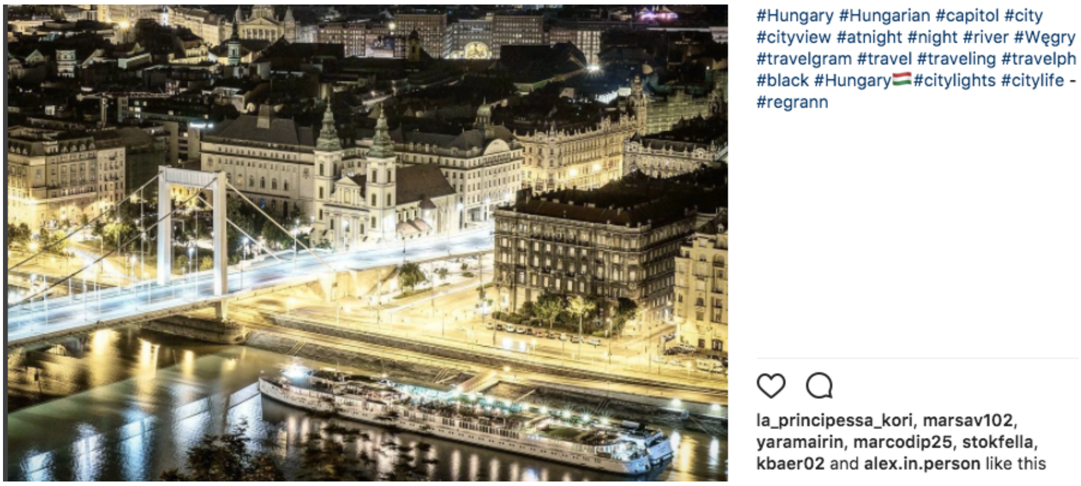
4.4.1 Mathematical formulation

Let us assume an Instagram image I_j and the set $\mathcal{T}^j$=$\{t_1^j, t_2^j, \ldots, t_k^j \ldots, t_{K_j}^j\}$ of K_j hashtags that accompany it (see Figure 4.3 for an example). We denote by r_k^j the relevance of hashtag t_k^j with the visual content of image I_j. We assume that the relevance scores $R[t_k^j]$, $k = 1, 2, \ldots, K_j$, $j = 1, 2, \ldots, M$ are computed with the aid of a crowd of N annotators (crowdtaggers) as explained in Section 5.4.4.

The aim of this study is to create a ranked set of tags for each one of the Instagram images I_j in terms of their relevance with its visual content, such as:

$$\mathcal{T}_r^j = \{t_{r,1}^j, t_{r,2}^j, \ldots, t_{r,k}^j \ldots, t_{r,k+1}^j \ldots, t_{r,K_j}^j\} \tag{4.13}$$

where $R[t_{r,k}^j] > R[t_{r,k+1}^j]$



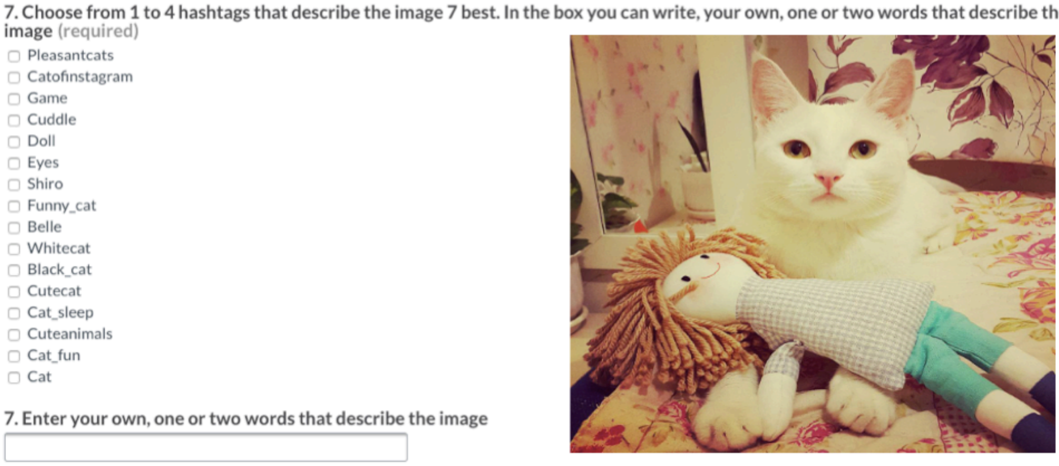
#Hungary #Hungarian #capitol #city #cityview #atnight #night #river #Węgry #travelgram #travel #traveling #travelph #black #Hungary🇭🇺#citylights #citylife - #regrann

la_principessa_kori, marsav102, yaramairin, marcodip25, stokfella, kbaer02 and alex.in.person like this

Figure 4.3: An example of an instagram image: At the top right we see the associated hashtags appended to it.

4.4.2 The proposed algorithm

We assume that a set $\mathcal{I} = \{I_1, I_2, \ldots, I_M\}$ of M Instagram images along with their associated hashtags $\mathcal{T} = \{\mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^j, \ldots, \mathcal{T}^M\}$ crawled according to the procedure described in Section 5.4.4. The methodology we follow to solve the problem mentioned in the previous section consists of the following steps. For the convenience of the readers who are interested to re-run the process detailed Python code is given in Section 5.4.6.



7. Choose from 1 to 4 hashtags that describe the image 7 best. In the box you can write, your own, one or two words that describe the image (required)

□ Pleasantcats
□ Catofinstagram
□ Game
□ Cuddle
□ Doll
□ Eyes
□ Shiro
□ Funny_cat
□ Belle
□ Whitecat
□ Black_cat
□ Cutecat
□ Cat_sleep
□ Cuteanimals
□ Cat_fun
□ Cat

7. Enter your own, one or two words that describe the image

Figure 4.4: An example of hashtag selection process that took place via *Figure-eight*

- *Step 1*: The relevance $R[t_k^j]$, $k = 1, 2, \dots, K_j$ of each hashtag with respect to the visual content of the associated image I_j is assessed by a set $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ of N users (annotators) with the aid of a crowdsourcing platform as it can be seen in Figure 4.4.

- *Step 2*: Given that all users assessed all image hashtags we can rank their effectiveness by considering the HITS algorithm. For that purpose we construct a bipartite graph:

$$\begin{aligned} \mathcal{B} &= \{\mathcal{V}, \mathcal{E}\} \\ \mathcal{V} &= \mathcal{V}_U \bigcup \mathcal{V}_T \\ \mathcal{V}_U \bigcap \mathcal{V}_T &= \emptyset \end{aligned} \tag{4.14}$$

where $\mathcal{V}_U$ and $\mathcal{V}_T$ are the sets of vertices corresponding to the annotators and hashtags, respectively, while $\mathcal{E} = \{e_{ik}^j\}$ is the set of edges denoting that the i-th user selected (considered as visually relevant) the tag t_k^j of image I_j.

- *Step 3*: The effectiveness (reliability) of annotators is approximated with the set of hub values $\mathcal{H} = \{h[v_1], h[v_2], \dots, h[v_i], \dots, h[v_N]\}$, where $h[v_i]$ is the hub value of vertex $v_i \in \mathcal{V}_U$, computed with the aid of the HITS algorithm (see also Section 4.4.4).

- *Step 4*: For each image I_j we construct a weighted bipartite graph as follows:

$$\begin{aligned} \mathcal{B}^j &= \{\mathcal{V}^j, \mathcal{E}^j\} \\ \mathcal{V}^j &= \mathcal{V}_U \bigcup \mathcal{V}_T^j \\ \mathcal{V}_U \bigcap \mathcal{V}_T^j &= \emptyset \\ \mathcal{E}^j = \{(v_i, v_k, h[v_i]) &| v_i \in \mathcal{V}_U, v_k \in \mathcal{V}_T^j, h[v_i] \in \mathcal{H}\} \end{aligned} \tag{4.15}$$

where $\mathcal{V}_U$ is the set of vertices corresponding to the annotators, $\mathcal{V}_T^j$ is the set of vertices corresponding to the hashtags of the j-th image and $\mathcal{E}^j$ is the set of weighted edges denoting that the i-th-user selected (considered as visually relevant) the tag t^j_k of image I_j.

- *Step 5*: A ranked set of tags, $\mathcal{T}^j_r$=$\{t^j_{r,1}, t^j_{r,2}, \ldots, t^j_{r,k}, t^j_{r,k+1} \ldots, t^j_{r,K_j}\}$, for each Instagram image I_j is achieved through the set of authority values $\mathcal{A}^j = \{a^j[v_1], a^j[v_2], \ldots, a^j[v_k], a^j[v_{k+1}], \ldots, a^j[v_{K_j}]\}$, where $a^j[v_k]$ is the authority value of vertex $v_k \in \mathcal{V}^j_T$, computed with the aid of the HITS algorithm when it is applied on the weighted bipartite graphs that were created in the previous step.

4.4.3 Bipartite networks

As we mentioned in Section 4.4 we conducted a two stage research. In the first stage we created a bipartite graph for each image and in the second stage we created a full bipartite graph. In order to fully explain the two stages we provide with one example of each stage.

4.4.3.1 Creation of a bipartite graph

In Fig. 4.6 is shown the bipartite graph of image with ID =2023 (see Fig. 4.5), i.e., j=2023. In this example the set of users is $\mathbf{u}^j$=$\{58, 77, 85, 122, 145\}$ while the sets $\mathbf{t}^j$, $\mathbf{e}^j$ are as follows:
$\mathbf{t}^j$=$\{2025, 2030, 2039, ..., 2226\}$
$\mathbf{e}^j$=$\{(58, 2025), (58, 2049), ..., (77, 2052), ..., (145, 2073)\}$

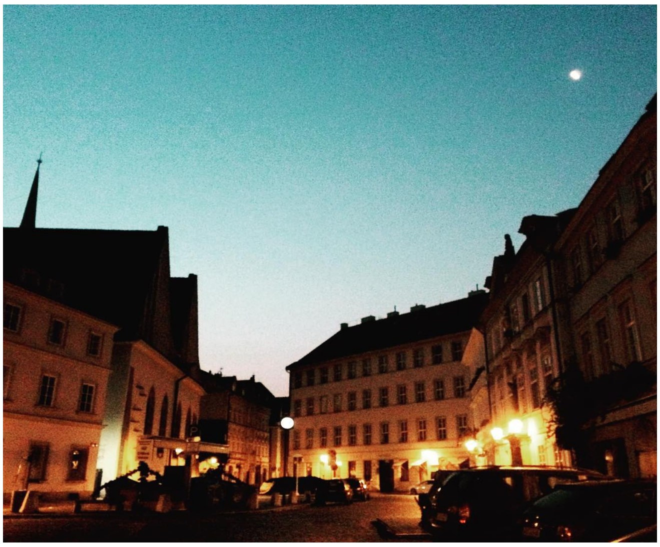


Figure 4.5: The image with ID 2023

In the HITS terminology the set $\mathbf{u}^j$ represents the hubs, i.e., the users that select hashtags, while the set $\mathbf{t}^j$ represents the authorities. By observing Table 4.1 in conjunction with Fig. 4.6 we see that the user 145 is highly unreliable. None of the hashtags she/he chose was selected by anyone of the other users. As a result her/his hub value is practically zero while the authority values of the hashtags she/he chose are also zero. In contrary, the user 58 selected five hashtags in total from which the three were also selected by other users. As a results she/he gets high hub value and the corresponding hashtags receive high authority values.

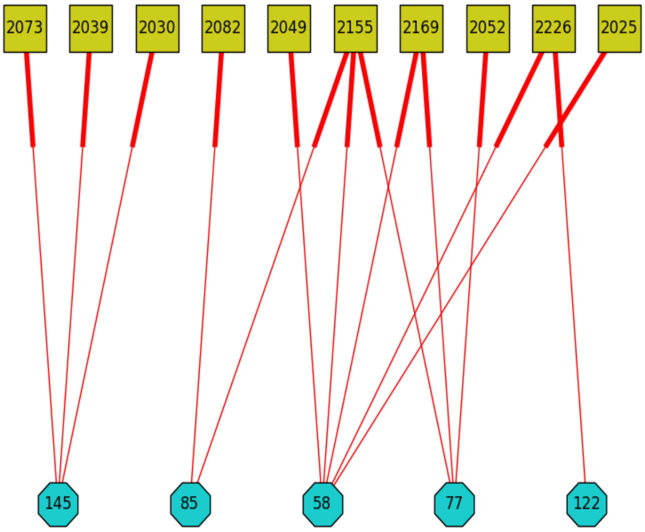


Figure 4.6: An example of the user-tag network for image 14. The squares show the tags while the circles show the users that selected those tags for the annotation of image 14.

Hashtag ID	Hashtag	Authority value	User ID	Hub value
2155	Lights	0.2542	58	0.4679
2169	Sky	0.2101	77	0.2913
2226	Moon	0.1520	85	0.1595
2049	Street	0.1295	122	0.0812
2025	Summer	0.1295	145	0.0000
2052	Houses	0.0806		
2082	Library	0.0441		
2030	Getinspired	0.0000		
2039	TakeOut	0.0000		
2073	Piranha	0.0000		

Table 4.1: Authority and Hub scores for the bipartite network of image with ID 2023 (see also Fig. 4.5)

4.4.3.2 The HITS algorithm: an example

In Figure 4.7 it is shown, for better visualization, the k-core[2] (k=6) of the bipartite graph corresponding to image 7 (the one shown in Figure 4.4). The radius of each tag is analogous to the weighted degree of the corresponding vertex. The whole bipartite graph for image 7 consists of 607 vertices: 499 annotators (users), the 16 hashtags of image 7 and another 92 tags suggested by the annotators.

Table 4.2 shows the authority values for the hashtags associated with image 7 along with the hub values of the 16 most reliable annotators (for this specific image) after the application of the proposed methodology.

4.4.4 The application of HITS algorithm on bipartite and weighted bipartite graphs

The HITS (Hyperlink-Induced Topic Search) algorithm was initially introduced by Kleinberg [243, 244] in order to analyze a collection of web-pages, relevant to a topic, and locate the most “authoritative”

[2] https://networkx.github.io/documentation/stable/reference/algorithms/core.html

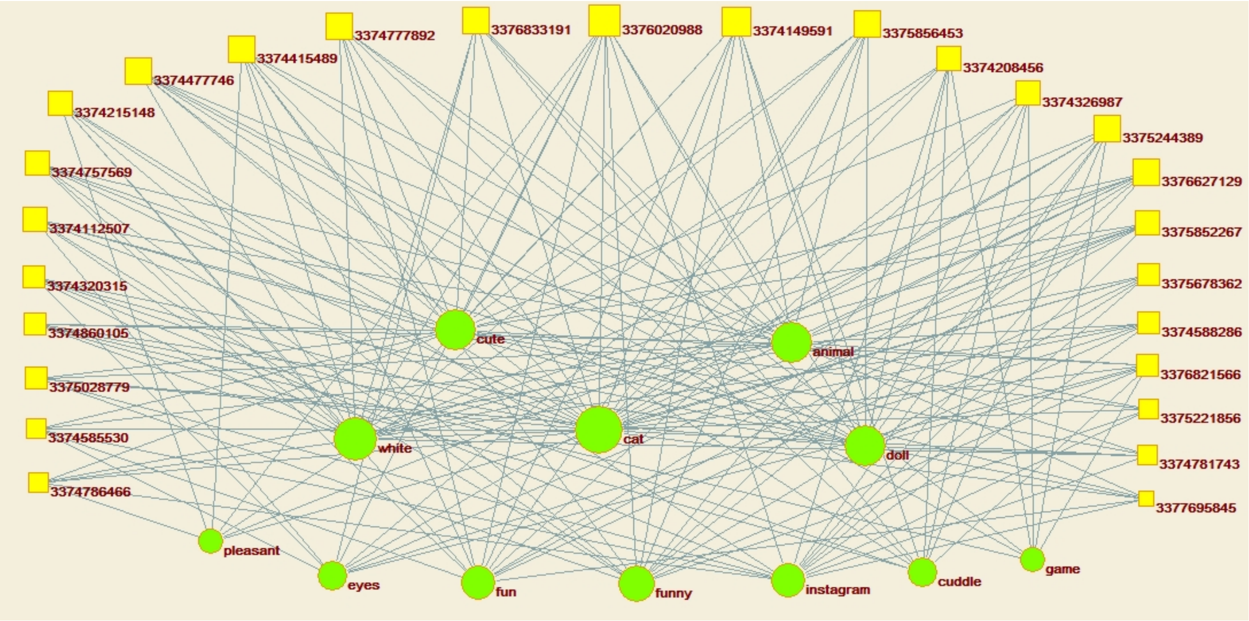


Figure 4.7: A subgraph of user-tag bipartite network for image #7. Circles show the tags while the boxes show the annotators that selected those tags.

Hashtag	Authority	Annotator ID	Hub ($x10^{-2}$)
cat	0.2027	3376020988	0.5582
doll	0.1314	3374149591	0.5163
white	0.1264	3374415489	0.4872
cute	0.1171	3374112507	0.4806
animal	0.0635	3374477746	0.4680
funny	0.0621	3376833191	0.4563
eyes	0.0471	3375771052	0.4556
instagram	0.0434	3375856453	0.4513
fun	0.0389	3374757569	0.4489
game	0.0279	3374777892	0.4256
pleasant	0.0267	3374647452	0.4037
cuddle	0.0256	3374505202	0.4029
belle	0.0092	3376453894	0.3996
shiro	0.0077	3374248101	0.3981
sleep	0.0060	3375852267	0.3976
black	0.0040	3374781743	0.3964

Table 4.2: Authority and Hub values for the bipartite network of image #7 (see also Fig. 4.7) - only the 16 most reliable annotators are shown

ones in that topic. It performs link analysis on those web pages in order to rank them in terms of two measures: hub value and authoritativeness. The authority score estimates the importance of the content of the page while the hub score estimates the quality of its links to other pages. Thus, a web-page that has many inlinks from other pages with high hub value is considered an authority while a page with many outlinks to high authority web-pages is a hub [245,246]. In simple words, the main principle of the HITS algorithm is that an informed hub points to many effective authorities and an effective authority is pointed out by many informed hubs. Thus, authorities and hubs have a mutual reinforcement relationship [247].

The HITS algorithm is commonly used for the analysis of two-mode networks represented as bipartite graphs. In that case both authority and hub values are used as measures of centrality[3], however, their interpretation differs significantly. A vertex with high authority score is considered as an expert while a vertex with high hub value is assumed as a good recommender. The authority $a[v]$ and hub value $h[v]$ of a vertex v in a bipartite graph are (iteratively) computed with the aid of the following equations:

$$\begin{aligned} a[v] &= \sum_{v_i \in \mathcal{N}_{v,U}} h[v_i] \\ \mathcal{N}_{v,U} &= \{v_i | v_i \in \mathcal{V}_U,\ (v_i, v) \in \mathcal{E}\} \end{aligned} \tag{4.16}$$

$$\begin{aligned} h[v] &= \sum_{v_i \in \mathcal{N}_{v,T}} a[v_i] \\ \mathcal{N}_{v,T} &= \{v_i | v_i \in \mathcal{V}_T,\ (v, v_i) \in \mathcal{E}\} \end{aligned} \tag{4.17}$$

where $\mathcal{N}_{v,U}$ is the set of vertices in $\mathcal{V}_U$ that point to vertex v and $\mathcal{N}_{v,T}$ is the set of vertices in $\mathcal{V}_T$ that vertex v points to (see also eq. 4.14).

It can be seen in eq 4.16 and 4.17 that a vertex's authority value is the sum of the hub score of all vertices pointing to it while its hub value is the sum of authority scores of all vertices that it points to. The final hub-authority values of a vertex are determined after infinite repetitions of the algorithm but in practice typical convergence tests, based on the number of iterations or the change of hub - authority scores between consecutive iterations, are applied. Given that directly and iteratively applying the above equations leads to diverging values, it is necessary to normalize hub and authority values after every iteration so as to sum to 1, i.e., $\sum_v h[v] = 1$, $\sum_v a[v] = 1$. By definition the initial values of $a[p]$ and $h[p]$ are set to 1.

For weighted undirected bipartite graphs $\mathcal{B}^j$, such as those corresponding to a user-tag bipartite network for a specific image I_j (see eq. 4.15), the equations of the HITS algorithm are modified as follows:

$$\begin{aligned} a^j[v] &= \sum_{v_i \in \mathcal{N}^j_{v,U}} h[v_i] \cdot h^j[v_i] \\ \mathcal{N}^j_{v,U} &= \{v_i | v_i \in \mathcal{V}_U,\ (v_i, v, h[v_i]) \in \mathcal{E}^j\} \end{aligned} \tag{4.18}$$

$$\begin{aligned} h^j[v] &= \sum_{v_i \in \mathcal{N}^j_{v,T}} h[v_i] \cdot a^j[v_i] \\ \mathcal{N}^j_{v,T} &= \{v_i | v_i \in \mathcal{V}^j_T,\ (v, v_i, h[v_i]) \in \mathcal{E}^j\} \end{aligned} \tag{4.19}$$

where $\mathcal{N}^j_{v,U}$ is the set of vertices in $\mathcal{V}_U$ that point to vertex v and $\mathcal{N}^j_{v,T}$ is the set of vertices in $\mathcal{V}^j_T$ that vertex v points to (see also eq. 4.15).

[3] https://en.wikipedia.org/wiki/Centrality

4.4.5 Folksonomies and the FolkRank algorithm

While our approach is a modification of the HITS algorithm to handle {user, images, hashtags} folksonomies, the FolkRank [177] is a known modification of the PageRank algorithm towards this direction. FolkRank makes use of the personalization component of the PageRank algorithm and applies single entity optimization. By doing so, Folk rank is capable of handling the inherent difficulty to adapt a single entity ranking algorithm (PageRank) to a three entity structure (folksonomy). An additional difficulty comes from the fact folksonomies are usually modelled as uni-directed graphs, i.e., humans select tags for an item. In order to handle this problem Hotho *et al.* [248] proposed a modified version of the FolkRank algoritm, called *differential FolkRank*. It is this algorithm that is used for comparison with the proposed method in the next section.

4.4.6 Evaluation metrics

Let us denote with $\mathcal{G}$ =$\{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^M\}$ the set of hashtags in the gold standard set, where $\mathcal{G}^j$ is the gold standard set for the j-th image. In our case, we define the gold standard as the set of descriptive annotations by the creator. Let us also denote with $\mathcal{T}^j_{r,\theta}$=$\{t^j_{r,1}, t^j_{r,2}, \ldots, t^j_{r,k}\}$ the ordered set of tags for image I_j such that $a^j[t^j_{r,1}] \geq a^j[t^j_{r,2}] \geq \ldots \geq a^j[t^j_{r,m}] \ldots \geq a^j[t^j_{r,k}]$ and $a^j[t^j_{r,k}] > \theta$, where $a^j[t^j_{r,m}]$ is the authority value of the vertex of bipartite graph $\mathcal{B}^j$ corresponding to the tag $t^j_{r,m}$.

The recall value $R_{j,\theta}$ for image I_j at the authority threshold value θ, i.e., the portion of tags in the gold standard set that were identified by the HITS algorithm when only the annotator tags with authority score higher than θ were kept, is given by:

$$R_{j,\theta} = \frac{||\mathcal{T}^j_{r,\theta} \cap \mathcal{G}^j||}{||\mathcal{G}^j||} \tag{4.20}$$

where $\cap$ denotes the set intersection operation and $||\Omega||$ refers to the cardinality of set Ω.

In a similar manner we define the precision value $P_{j,\theta}$ for image I_j at the authority threshold value θ, as the portion of the tags that were identified by the HITS algorithm that are included in the gold standard set of image I_j:

$$P_{j,\theta} = \frac{||\mathcal{T}^j_{r,\theta} \cap \mathcal{G}^j||}{||\mathcal{T}^j_{r,\theta}||} \tag{4.21}$$

With the aid of eq. 4.20 and 4.21 we can compute the Recall, Precision and F_1-measure, at the authority threshold value θ, for the whole image dataset as follows:

$$R_\theta = \frac{1}{M}\sum_{j=1}^{M} R_{j,\theta} \tag{4.22}$$

$$P_\theta = \frac{1}{M}\sum_{j=1}^{M} P_{j,\theta} \tag{4.23}$$

$$F_{1,\theta} = \frac{2 \cdot P_\theta \cdot R_\theta}{P_\theta + R_\theta} \tag{4.24}$$

The effectiveness of the proposed method is also evaluated with the aid of Mean Reciprocal Rank (MRR) [249]. The MRR of an image I_j is computed as follows:

$$MRR_j = \frac{1}{||\mathcal{T}_r^j \cap \mathcal{G}^j||} \sum_{i=1, t_{r,i}^j \in \mathcal{G}^j}^{K_j} \frac{1}{r_i^j} \tag{4.25}$$

where $\mathcal{T}_r^j$=$\{t_{r,1}^j, t_{r,2}^j, \ldots, t_{r,K_j}^j\}$ is the ordered set of tags for image I_j, $\mathcal{G}^j$ is the corresponding gold standard set, and r_i^j is the ranking of tag $t_{r,i}^j$.

The MRR is computed as the average of MRR_j across all images.

Another key performance metric in information retrieval is Mean Average Precision (MAP). The purpose of MAP is to calculate the average of the precision value of the top set of k results. It is defined as follows:

$$MAP_j = \frac{1}{||\mathcal{T}_r^j \cap \mathcal{G}^j||} \sum_{k=1}^{K_j} \frac{||\mathcal{T}_{r,k}^j \cap \mathcal{G}^j||}{||\mathcal{T}_{r,k}^j||} \tag{4.26}$$

where $\mathcal{T}_{r,k}^j$=$\{t_{r,1}^j, t_{r,2}^j, \ldots, t_{r,k}^j\}$ is the ordered set of the k first tags of image I_j.

A practical example on how the MAP and MRR scores are computed is shown in Table 4.3 for the particular case of Image #6.

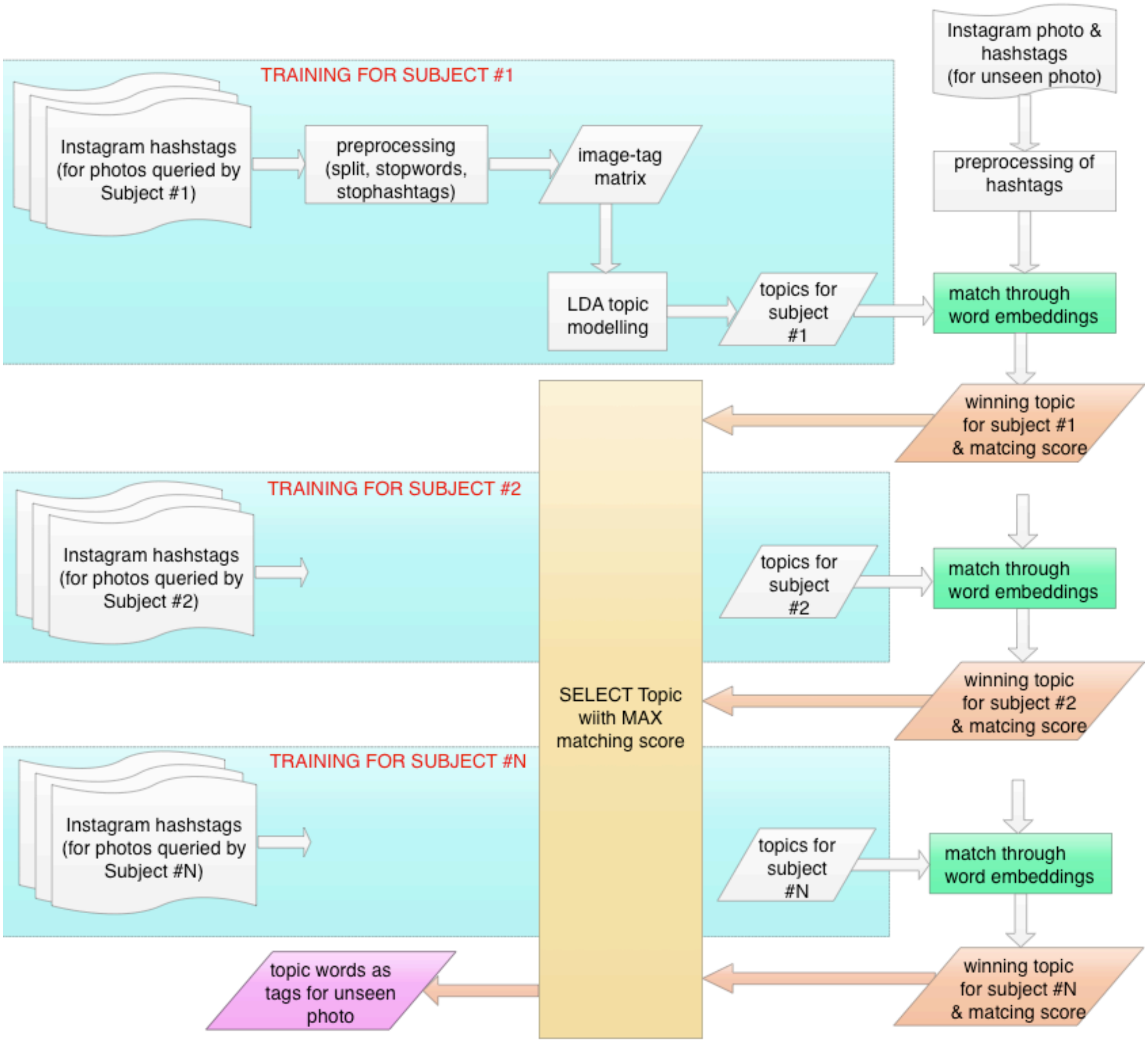
Hashtag	In Gold Standard	ASR	Precision	RR
vacation		1	0	
beach	x	2	1/2 (0.500)	1/2 (0.500)
sand	x	3	2/3 (0.667)	1/3 (0.333)
sun		4	0	0
bikini	x	5	3/5 (0.600)	1/5 (0.200)
sea	x			
sky	x			
woman	x			
hat	x			
Sum			1.767	1.033
Average			0.589	0.344

Table 4.3: Average Precision and Mean Reciprocal Rank for Image #6 hashtags according to authority score rank (ASR)

.

4.5 Image retrieval using topic modelling

The aim of that approach is to identify the relevance of an Instagram photo to an image category based on the hashtags appended to that image and the topics derived from the hashtags of all images belonging to that category. Here, an image category corresponds to all images collected under the same 'subject' hashtag. While the primary aim of this study is to serve as an alternative way of AIA for Instagram photos, it can be also used to filter out Instagram hashtags. This will be done on the basis of a simple rule. Assuming that an Instagram photo belongs to an image category, then its hashtags that do not coincide with the topics developed for that category will be removed. In addition, the training data of that specific image category will only include the hashtags belonging to the corresponding topic models. Moreover, in Section 3.5 we mentioned that we evaluated the topic modelling approach with topic coherence and human evaluation. In this section we analyze the methodologies we followed for the aforementioned evaluation methods. To the best of our knowledge, this is the first work that suggests the topic model analysis in order to identify image tags in the environment of Instagram. Moreover, the methodology is straightforward to implement since data collection is not complex, and there are just a few steps to follow.



TRAINING FOR SUBJECT #1
Instagram hashtags (for photos queried by Subject #1)
preprocessing (split, stopwords, stophashtags)
image-tag matrix
LDA topic modelling
topics for subject #1
Instagram photo & hashtags (for unseen photo)
preprocessing of hashtags
match through word embeddings
winning topic for subject #1 & matcing score
TRAINING FOR SUBJECT #2
Instagram hashtags (for photos queried by Subject #2)
topics for subject #2
match through word embeddings
winning topic for subject #2 & matcing score
SELECT Topic wiith MAX matching score
TRAINING FOR SUBJECT #N
Instagram hashtags (for photos queried by Subject #N)
topics for subject #N
match through word embeddings
winning topic for subject #N & matcing score
topic words as tags for unseen photo

Figure 4.8: The conceptualised architecture adopted for topic modelling of Instagram hashtags and their use for AIA

The architecture of the proposed technique to mine relevant tags for an image from its Instagram hashtags is shown in Figure 4.8. First, topics models are created from a collection of Instagram hashtags of photos belonging to the same subject (i.e., queried by the same hashtags, say #airplane). This results in a set of topics $\mathcal{S}^q = \{\mathcal{T}_1^q, \mathcal{T}_2^q, ...\mathcal{T}_k^q\}$ for the q-th subject.

Second, the matching score of the hashtags $\mathcal{H}_I$ of an unseen Instagram image I, preprocessed in the same way as the training instances, with each one of the topics $\mathcal{T}_t^q$ of the q-th subject is computed. Both $\mathcal{H}_I$ and $\mathcal{T}_t^q$ are sets of words while the latter includes also the importance of each word expressed through its relative frequency in the topic. The matching score $R(\mathcal{H}_I, \mathcal{T}_t^q)$ between these two sets can be computed as a weighted sum of the pair similarities of their word embeddings[4], pre-trained on external sources such Google News and Wikipedia, as shown in eq. 4.27:

$$R(\mathcal{H}_I, \mathcal{T}_t^q) = \frac{1}{|\mathcal{H}| \cdot |\mathcal{T}_t^q|} \sum_{h_I \in \mathcal{H}_I} \sum_{w_t^q \in \mathcal{T}_t^q} f(w_t^q) \cdot cc(\vec{h}_I, \vec{w}_t^q) \tag{4.27}$$

where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$, $\vec{h}_I$ and $\vec{w}_t^q$ are the word embeddings of hashtag h_I and topic word w_t^q respectively, $f(w_t^q)$ is the relative frequency of topic word w_t^q and $cc(.,.)$ is the correlation coefficient[5] operator.

In the third step, the best matching, to the set of hashtags $\mathcal{H}_I$, topic $\mathcal{T}_{opt}^q(\mathcal{H}_I)$ of the q-th subject is selected with aid of eq. 4.28:

$$\mathcal{T}_{opt}^q(\mathcal{H}_I) = argmax\{R(\mathcal{H}_I, \mathcal{T}_t^q)\} \tag{4.28}$$

Finally, the best matching topic $\mathcal{T}_{opt}(\mathcal{H}_I)$ for set $\mathcal{H}_I$, across all subjects, denotes the tags that will be assigned to Instagram image I and is given by eq. 4.29:

$$\mathcal{T}_{opt}(\mathcal{H}_I) = argmax\{\mathcal{T}_{opt}^q(\mathcal{H}_I)\} \tag{4.29}$$

4.5.1 Training architecture and corpus enrichment

In order to train our models we built on the approach suggested by Chen *et al.* [250] for short text classification. Figure 4.9 summarizes the steps that we followed while the tools we used include Jupyter Notebook[6], Python 3.5[7] and Prabhakaran's source code[8].

Along with Instagram hashtags for each subject, related material was also crawled from Wikipedia and used for the purpose of training. As Chen *et al.* [250] explain this approach produces more general, interpretable and less application / platform specific topic models.

4.5.2 Preprocessing

Data cleaning is very important for generating useful topic models. For normal texts data cleaning includes some standard steps such as: (a) *tokenization*, that is, splitting a document to its atomic elements (e.g. words) called tokens, (b) *stopping*, that is, removing words that are frequent to any text (e.g. 'and',

[4] https://en.wikipedia.org/wiki/Word_embedding

[5] https://en.wikipedia.org/wiki/Correlation_coefficient

[6] http://jupyter.org/

[7] https://www.python.org/downloads/release/python-350/

[8] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

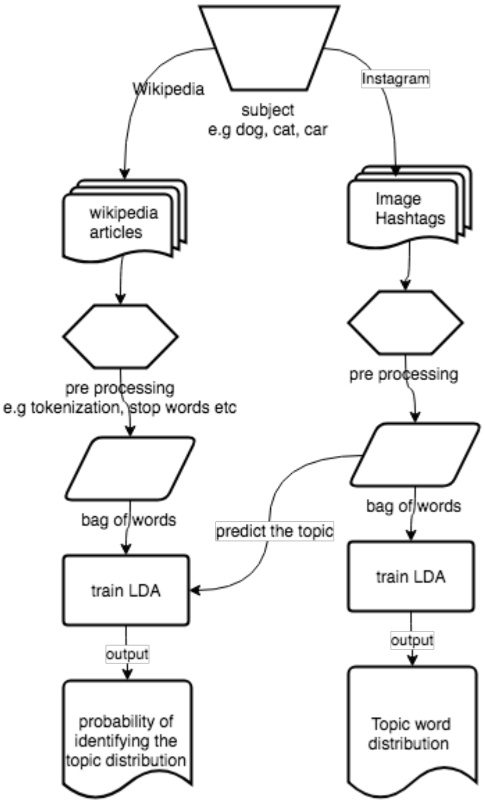


Figure 4.9: LDA with Wikipedia

'the', etc) and thus, meaningless for the description of a topic, and (c) *lemmatization*, that is, merging of words that share the same or similar meaning.

Instagram hashtags, on the other hand, are unstructured and ungrammatical, and it is important to use linguistic processing to (a) split a composite hashtag to its consisting words (e.g. the hashtag '#picoftheday' should be split to four words: 'pic', 'of', 'the', 'day'), (b) remove stopwords that are produced in the previous stage, (c) remove stop hashtags(see Section 4.3), that is hashtags that are used to fool the search results of the Instagram platform, (d) perform spelling checks to account for (usually intentionally) misspelled hashtags (e.g. '#airoplane', '#airoplanes' should be changed to '#airplane'), and (e) perform lemmatization as with the Wikipedia corpus.

All the previously mentioned preprocessing was done with the help of NLT[9], Wordnet[10] and personally developed code in Python.

4.5.3 Training with the LDA model

Using the prepared corpus mentioned above we trained our LDA models with the help of Python Gensim[11] and specifically the MALLET[12] implementation. According to various researchers, MALLET (MAchine Learning for LanguagE Toolkit) is an efficient implementation of the LDA that not only runs faster than the classic LDA implementation but also gives better topics segregation. The result of each LDA model

[9] https://www.nltk.org/

[10] https://wordnet.princeton.edu/

[11] https://radimrehurek.com/gensim/

[12] http://mallet.cs.umass.edu/

is a list of topics and the topic word distribution which denotes the probability distribution of words for each topic.

The trained models' output determines and categorizes topics by checking the distribution of words for each topic, quantifies the distance between the topic distribution vectors and calculates the topic coherence. The pyLDAVis tool[13] help us visualize and investigate all three previously mentioned parameters concurrently and qualitatively assess the performance of the proposed method. As shown in Figures 5.22 and 5.23 the topics are presented as discs. The disc area corresponds to topic coherence while the correlation between the two topics is reflected on the geodesic distance between the corresponding discs; neighboring or overlapping discs have high correlation. Finally, by selecting each disc you can see the consisting words of the corresponding topic along with their relative frequency.

4.5.4 Evaluation of topic models

Topic coherence [251] is a measure used to evaluate topic models for methods that automatically generate topics from a collection of documents, using latent variable models. It is defined as the average / median of the pairwise word-similarity scores of the words in the topic. A good model will generate coherent topics, i.e., topics with high topic coherence scores. Good topics are topics that can be described by a short label, therefore this is what the topic coherence measure should capture. A simple coherence measure is the UCI [252]. Assuming that a generated topic $\mathcal{T}_k$ consists of an ordered set of words $\mathcal{T}_k =< w_1^k, \cdots, w_n^k, \cdots >$, such as, $f(w_i^k) \geqslant f(w_j^k), \forall i < j$, where $f(w)$ is the frequency of word w, the UCI coherence $C_{UCI}[\mathcal{T}_k]$ of topic $\mathcal{T}_k$ is computed based on the top n words in terms of frequency of appearance and is given by:

$$C_{UCI}[\mathcal{T}_k] = \frac{2}{n \cdot (n-1)} \sum_{i<j\leqslant n} s(w_i^k, w_j^k) \tag{4.30}$$

where $s(w_i^k, w_j^k)$ is the similarity score between topic words i and j of the k-th topic. As similarity measure $s(.,.)$ is usually used the Pointwise Mutual Information (PMI[14])(see [253], [254]). The similarity score must be computed on a corpus different than the one used to extract the topics. Wikipedia is a corpus that is usually used to calculate topic coherence. Newman *et. al.* calculated the semantically similar words among the top 10 terms in a topic and measured the semantic similarity of the words using external resources, e.g. WordNet and Wikipedia. They concluded that evaluation metric based on the Pointwise Mutual Information estimate of the word pairs generated from Wikipedia was the closest to human judgments [255]. Niraula *et. al.* [256] proposed a methodology for semantic similarity in two texts based on probabilistic method Latent Dirichlet Allocation (LDA). Since LDA produces a specific number of topics the researchers used topic coherence to select the number of topics upfront. To calculate the topic coherence they used a Wikipedia-based corpus. Fang *et. al.* [257] in their study to calculate topic coherence they used Wikipedia.

[13] http://pyldavis.readthedocs.io/en/latest/readme.html

[14] https://en.wikipedia.org/wiki/Pointwise_mutual_information

4.5.5 Human interpretation with the aid of word clouds

In this section we investigate how humans interpret of topics with the aid of word clouds. Specifically we study how humans understands the topics derived from the hashtag sets of Instagram photos that were grouped together by a common query hashtag which we call subject. The topics are illustrated as word clouds with the queried hashtags (subjects) hidden and the humans are asked to guess the hidden hashtag providing their best guesses. The aim of the current research is to examine the performance accuracy interpretation in topic modeling we created from Instagram hashtags. In case humans' choice coincides with the subject of the word cloud we can conclude that word clouds are, indeed, related with the subject.

From the literature review (see Section 3.5.1) we can conclude that while presentation of Instagram related textual data, such as captions, comments and hashtags, via word clouds is quite common, no meta analysis of the word clouds themselves has been conducted in anyone of those works. Word clouds have been mainly used for visualisation purposes but the appropriateness of this visualisation format was never assessed. Thus, in addition to the application perspective of our work, which emphasizes on mining terms from Instagram hashtags for image tagging, the crowd-based meta analysis of word clouds provides also useful insights about their appropriateness for topic visualisation. Some of the reported works applied topic modelling to summarize textual information using the classic LDA approach. Our topic modeling algorithm [195] is quite different and tailored to the specific case of Instagram posts.

In order to investigate the appropriateness of word clouds for topic assessment we compare and discuss the crowd-based and student-based interpretation of word clouds created from Instagram hashtags. So we selected a number subjects / hashtags and for each subject/hashtag retrieve relevant images. Then for each image collect all hashtags appended to it. All collected hashtags were undergone preprocessing so as to derive meaningful tokens (words in English) as we described in 4.5.2. Preprocessing was conducted with the help of Natural Language ToolKit (NLTK - https://www.nltk.org/), Wordnet[15] and personally developed code in Python. Instagram photos and the associated hashtag sets belonging to a common subject were grouped together and modeled as a bipartite network. Then, topic models were created for each one of the subjects following the approach described in [195]. For each one of the topics a word cloud was created. The token corresponding to the associated subject (query hashtag) was excluded in order to examine whether the crowd and student would guess it correctly. Word clouds visualization was done with the help of WordCloud[16] Python library.

4.6 Assessment of relevance between visual content and hashtag sets in Instagram photos

In this section we describe the methodology we followed to assess the relevance between visual content and hashtag sets in Instagram photos. The proposed methodology is summarized in the flowchart shown in Figure 4.10. In order to numerically represent the visual content we use color histograms (a widely applied feature set in content-based image retrieval) while the hashtag set of an Instagram image is numerically represented with the aid of word embeddings.

[15] https://wordnet.princeton.edu/

[16] https://amueller.github.io/word_cloud/

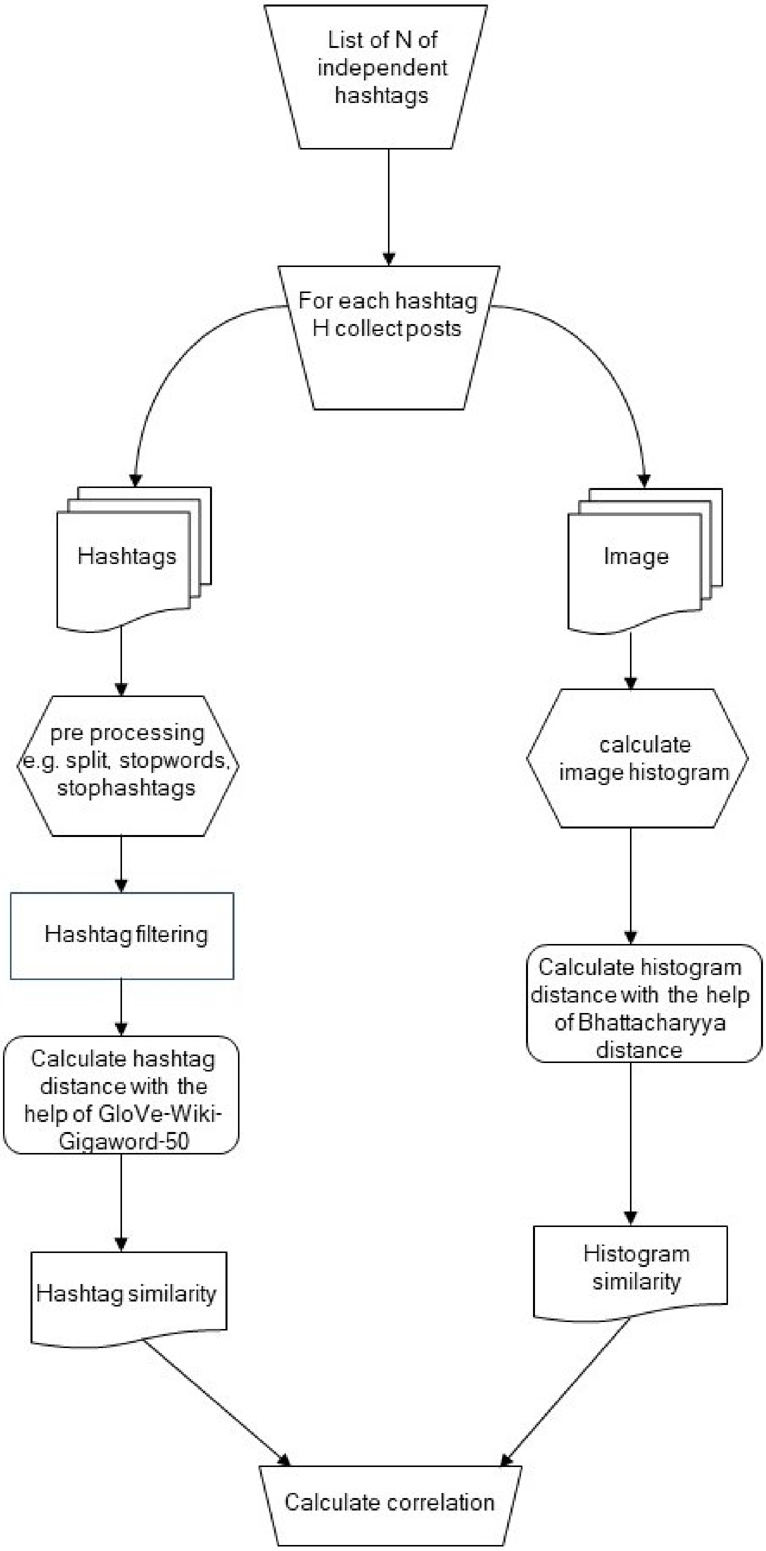

List of N of independent hashtags
For each hashtag H collect posts
Hashtags
Image
pre processing e.g. split, stopwords, stophashtags
calculate image histogram
Hashtag filtering
Calculate histogram distance with the help of Bhattacharyya distance
Calculate hashtag distance with the help of GloVe-Wiki-Gigaword-50
Hashtag similarity
Histogram similarity
Calculate correlation

Figure 4.10: Flowchart of the methodology for the assessment of relevance between visual content and hashtag sets in Instagram photos

We investigate if we can achieve AIA with the help of Instagram images and hashtags. So, we explore if we can bridge the semantic gap between image low-level features such as color histogram and high-level semantic content as hashtags. To the best of our knowledge, no research quantified the similarity color histogram and hashtag in Instagram.

4.6.1 Image similarity and Bhattacharyya distance

Assuming that an image I is indexed by, a usually fixed length, feature vector $\vec{h}$, then the similarity $S(I_1, I_2)$ between two images I_1 and I_2 is computed with the aid of the distance $d(\vec{c}_1, \vec{c}_2)$ between their feature vectors [258]. To calculate the distance between two probability distributions $p(x)$ and $q(x)$ which are approximated by the corresponding normalized (see eq. 4.33) histogram vectors $\vec{c}_1$ and $\vec{c}_2$. Thus, the Bhattacharyya distance between two images I_1 and I_2 whose histogram elements $c_1(x)$ and $c_2(x)$, were computed on a set of color hues $\mathcal{X}$, is given by:

$$d(I_1, I_2) = -ln(BC(\vec{c}_1, \vec{c}_2)) \tag{4.31}$$

$$BC(\vec{c}_1, \vec{c}_2) = \sum_{x \in \mathcal{X}} \sqrt{c_1(x) \cdot c_2(x)} \tag{4.32}$$

where:

$$\sum_{x \in \mathcal{X}} c_1(x) = 1 \qquad \sum_{x \in \mathcal{X}} c_2(x) = 1 \tag{4.33}$$

In order to use Bhattacharyya distance as a similarity metric reformulation is required as indicated in eq. 4.34. However, many different reformulations expressing a similar logic can be applied.

$$S(I_1, I_2) = \frac{1}{1 + d(I_1, I_2)} \tag{4.34}$$

It is clear from eq. 4.34 that the similarity among two images ranges in the interval $(0, 1]$ with values close to 0 indicating very low similarity while 1 denotes perfect match [259].

4.6.2 Word Embeddings

Word embeddings are computed using the pre-trained models of the Glove project [223]. Those word embeddings were learned on Google News articles and Wikipedia content using the following optimization criterion [260]:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w} + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{4.35}$$

where $f(X_{ij})$ tabulates the number of times word j occurs in the context of word i, $w \in \mathbb{R}^d$ are word vectors and $\tilde{w} \in \mathbb{R}^d$ are separate context word vectors, V is the vocabulary size and b_i is a bias for w_i.

4.6.3 Matching hashtag sets

Let $\mathcal{H}_i$ and $\mathcal{H}_j$ be the filtered hashtag sets of Instagram posts corresponding to the i-th and j-th Instagram images respectively. The matching score $R(\mathcal{H}_i, \mathcal{H}_j)$ between these two sets is computed as a weighted sum of the pair similarities between the word embeddings of their constituting hashtags, as shown in eq. 4.36.

$$R(\mathcal{H}_i, \mathcal{H}_j) = \frac{1}{|\mathcal{H}_i| \cdot |\mathcal{H}_j|} \sum_{h_{ik} \in \mathcal{H}_i} \sum_{h_{j\xi} \in \mathcal{H}_j} cc(\vec{h}_{ik}, \vec{h}_{jxi}) \tag{4.36}$$

where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$, $\vec{h}_{ik}$ and $\vec{h}_{j\xi}$ are the word embeddings of hashtags h_{ik} and $h_{j\xi}$ belonging to hashtags sets $\mathcal{H}_i$ and $\mathcal{H}_i$ respectively, and $cc(.,.)$ is the similarity measure used with the word embeddings of Gensim models[17].

4.6.4 Research Hypotheses

The current work is formulated as an experimental study expressed through two null hypotheses:

$H0_1$*: In relevant posts, there is no significant correlation between the color similarity of Instagram image pairs and the similarity of their (filtered) hashtags sets.*

$H0_2$*: There is no significant difference in the average correlation between color histograms and hashtag sets in relevant posts and irrelevant posts.*

In order to confirm or reject the null hypotheses, the process shown in Fig. 4.10 is followed. First, we select N independent hashtags which in the context of the current work are referred to as hashtag subjects. We decided to create a set of 26 different subjects. For each subject, we create two collections of Instagram posts, one corresponding to images visually relevant to the subject and one containing images which are not visually relevant to the subject. For each subject images and hashtags were automatically collected using the Beautiful Soup[18] library of Python. For instance, if we query with the hashtag subject #dog we randomly select Instagram posts that depict dog(s) and posts that, despite containing the hashtag #dog, do not show any dog. The selection process is random and only the confirmation regarding the visual relevance is done through human intervention.

For each pair of posts P_i and P_j we isolate the corresponding photos I_i and I_j, we compute their color histograms and express them as vectors $\vec{c_i}$ and $\vec{c_j}$ and we compute their Bhattacharyya similarity with the aid of eq.4.31. At the same time the similarity of the associated hashtag sets of the posts, say $\mathcal{H}_i$ and $\mathcal{H}_2$, is computed through the process described in Section 4.6.3. All collected hashtags were undergone preprocessing so as to derive meaningful tokens (words in English) as described in 4.5.2.

The final step is to compute the correlation between hashtag set (mean) similarities and color histogram (mean) similarities with the help of Pearson correlation coefficient for both the relevant and irrelevant posts. By rejecting the $H0_1$ null hypothesis we can conclude that color similarity of images can be predicted by the similarity of their associated hashtag sets. Failing to reject the $H0_1$ null hypothesis indicates that the information obtained from hashtag sets and color histograms, respectively, could be

[17] https://radimrehurek.com/gensim/models/word2vec.html

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

seen either as complementary or totally uncorrelated. This depends on the decision regarding the second ($H0_2$) null hypothesis. By rejecting the $H0_2$ null hypothesis, we can conclude that the information provided by color histograms and hashtag sets much more correlated in the relevant posts (as one would expect from the fact that both describe the same visual content) than in irrelevant posts.

The purpose is to study the correlation between Instagram image and filtered hashtags sets. The primary purpose of the data collected (Instagram images and hashtags), is exactly to achieve the purpose of our paper. Pearson correlation measures the relationship between objects. Moreover, Pearson correlation is among the most commonly used [261]. In addition to these, in Zhang*et al.* [262] in their research to measure the similarity between brands via posts of brands' followers on social network services they use Pearson correlation to calculate the similarity between image and tag vectors. In order to reject the $H0_2$ hypothesis, it is important to compare correlation coefficient of relevant and irrelevant posts. The methods for comparing coefficient are Zou's confidence interval and z-score [263], the most straightforward way, and the approach taken here is through z-score. The z-score is used in bibliography to compare correlation coefficient [264], [265], [266].

4.7 Transfer learning

The actual test whether training sets mined from the Instagram, by applying the methods mentioned in the previous sections, can be used for developing effective AIA models is to try them for this purpose. Transfer learning provides such a framework.

Deep learning for Automatic Image Annotation purposes has recently gained increased research interest. The main problem of deep learning is the need for a huge number of training examples. Transfer learning is an alternative method to overcome those problems. With transfer learning, we can use pre-trained models developed for a different purpose and use our training data to fine-tuned them in the context of a new application. As we have seen in Section 3.7 the results with transfer learning are quite impressive. In addition, pre-trained ConvNets with fine-tuning policies can exceed in efficiency deep networks trained from scratch. Furthermore, fine-tuning can lead to faster convergence than training from the scratch [268].

Training deep neural networks requires a lot of data and huge computational power whici is impossible to achieve with personal computers. Google has created Colaboratory (a.k.a. Colab), a cloud-based service that can assist scientists to execute the machine learning in the cloud. Collab use Jupyter, an open-source and browser-based tool that integrates interpreting languages, libraries, and tools for visualization. Collab allows to write and execute Python code and importing essential libraries for machine learning, such as TensorFlow, Matplotlib, and Keras [269, 270].

In the context of the current research question we use the Residual Neural Networks (ResNets) developed by He *et al.* [271]. ResNets deal with the vanishing gradient and the degradation problem that appear during the ultra-deep CNNs train. They include stacked residual units (building blocks) containing skip connections to link the input and output of each unit. CNNs with residual units were proved to perform better than plain counterparts. ResNet-152, the specific architecture adopted, consists of 152 layers, as shown in Figure 4.11, providing residual connection between them. The weights for this model were created after ILSVRC-2012-CLS dataset for image classification training[19].

[19] https://tfhub.dev/google/imagenet/resnet_v1_152/classification/5

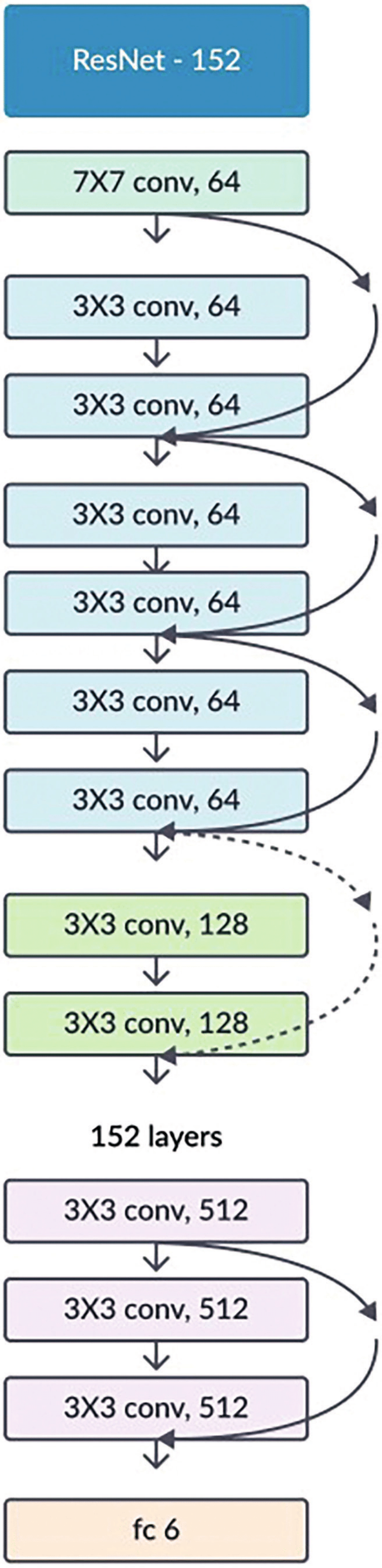


Figure 4.11: The ResNet-152 architecture
[267]

The Resnet-152 pre-trained model could be ideal for Instagram image classification in the context of transfer learning. In this context it has been used in a variety of studies dealing with image classification [268, 272, 273] with excellent results.

4.8 Summary

In Chapter 4 we have summarised the the methodologies adopted in the current thesis for answering the research questions set. We have described the technique to investigate whether we can locate hashtags related to the Instagram images they accompany, we have proposed a methodology to identify stophashtags and we have analyzed two approaches to filter out hashtags, the first based on HITS and the latter on topic modeling. Topic modelling was also applied in an image retrieval context. Chapter 4 ended with discussing the transfer learning approach for image classification. In Chapter 5, we explain the data collection and experimental assessment of the methodologies described in the current chapter.

Chapter 5

Data Collection and Empirical Assessment

Data collection contents

5.1 Introduction

In this chapter we describe data collection, presentation and analysis of results, and conclusions regarding the research questions 2-7 of the current thesis. Each of the following sections is devoted to one research question. References to scientific publications, produced during the corresponding study, are also given therein.

5.2 The descriptive power of Instagram hashtags

The purpose of this research question was to examine if we can use Instagram hashtags for AIA purposes and to assess whether the Instagram hashtags are semantically related with the visual content of Instagram photos. A rough estimation of the proportion of Instagram hashtags that are related with the visual content of the accompanied images was also pursued. Further details regarding this study as whole can be found at [15, 16]

5.2.1 Data collection

Data collection was conducted in two-stages. In the first stage 30 Instagram pictures were assigned to three separate online questionnaires containing 10 pictures each. Motivated by the results of the first stage research we decided to extend the first stage survey by expanding the number of images for annotation from 30 to 1000. A set of 1000 Instagram images which were selected from 100 different subjects / hashtags (10 relevant images per subject / hashtag) was used for that purpose. While in first stage images were collected randomly in the second stage we systematically collected 100 different query hashtags (as mentioned in the previous chapters we call the query hashtags *subjects*) and 10 relevant images from each query hashtag. While in the stage we used online questionnaire commercial service (SurveyMonkey[1]), in the second stage, for a higher flexibility, we constructed our own online questionnaire. We conducted the second stage research to confirm the results of the first stage in larger number of images systematically collected and with a significantly higher number of participants.

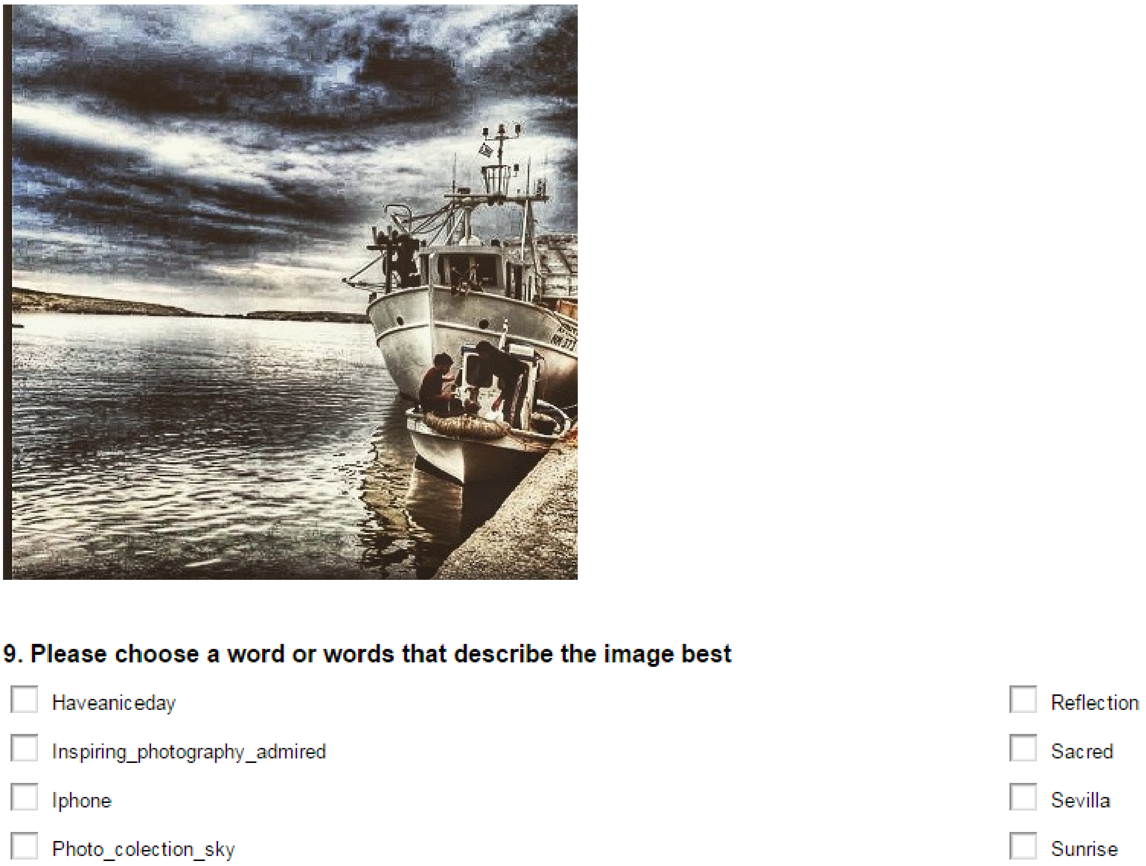
In both stage owners' hashtags surrounding images were automatically were crawled using the Beautiful Soup[2] library of Python. Then, one to four hashtags were manually chosen for each picture, which, according to our interpretation, better describe its visual content since, as mentioned, not all hashtags are intended to describe image. In both stages the participants were given four or eight choices 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 were four (including the hashtag of the owner); otherwise the participants were given eight options to select from. This rule was applied in order to keep a minimum chance level higher than or equal to 25%.

In Fig.5.1 is shown an example of a question presented to the participants of the first stage. The participants were asked to select the hashtags that better describe the visual content of the shown picture.

The three online questionnaires of the first stage were initially distributed, by electronic mail, to three experts in order to evaluate them. The results of the evaluation assisted the creation of a revised version

[1] https://www.surveymonkey.com/

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



9. Please choose a word or words that describe the image best

□ Haveaniceday

□ Inspiring_photography_admired

□ Iphone

□ Photo_colection_sky

□ Reflection

□ Sacred

□ Sevilla

□ Sunrise

Figure 5.1: An example of an image interpretation multiple choice question

which was, then, distributed by electronic mail to librarians of the library of Cyprus University of Technology and to undergraduate and postgraduate students of the department of Communication & Internet Studies. Each participant group, i.e., librarians, undergraduate and postgraduates students were randomly split so as to distribute the three questionnaires equally in each group. The survey was conducted between March, 19th and March 31st, 2015. A total of 39 questionnaires were collected and used for analysis.

As mentioned before, in the second stage we have collected of 1000 images from 100 different subjects (query hashtags). Those images were uploaded to Instagram by 970 different Instagram users. 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 Figure 5.3. 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 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. Because, a few of the participants in the previous stage were also participated in that stage 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.

In order to avoid fatigue effects, each participant was asked to 'annotate' only 20, randomly selected from the database, images in each session. However, users were allowed to repeat the process through another session as many times as they wished. The 'false' hashtags were randomly selected among the hashtags given to other images stored in the database in order to fully automate the process. In any case, participants were not aware that any of the given choices were related in any respect with the picture; thus, they were free to select as many of them as they wish according to their interpretation of the shown photo.

In order to reduce the probability of 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 were also asked to fill in information about their age, gender, years of internet experience and social media usage (see Tables 5.1 & 5.2).

Table 5.1: Users' demographics

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

Table 5.2: 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%

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. The survey was conducted between February, 15th and March 20th, 2016. 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.

5.2.2 Results

The data of the 39 filled in questionnaires of the first study were analyzed with aid of SPSS[3], MS Excel[4] and MATLAB[5] platform using the metrics defined in Section 4.2.1.

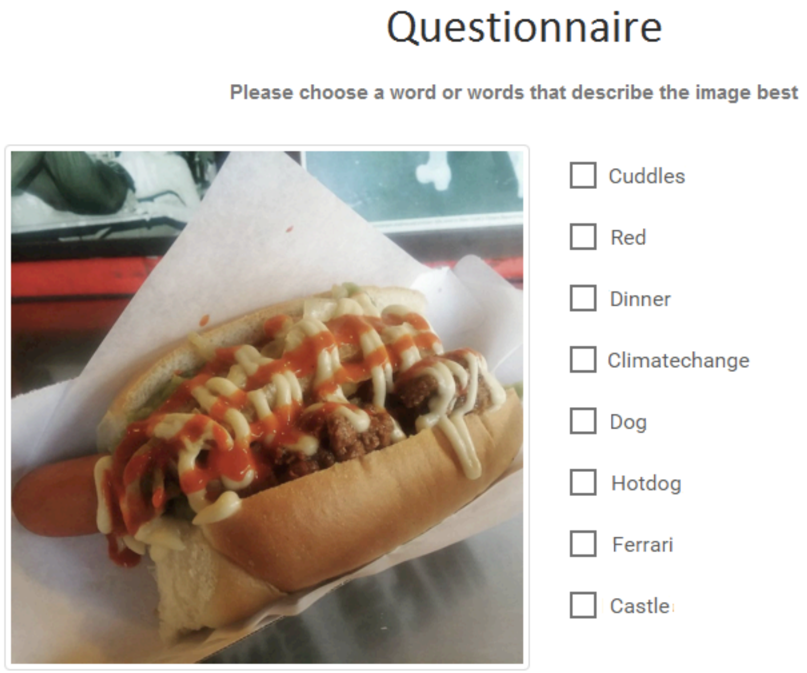
Fig.5.9 shows the per participants' Recall (eq. 4.2), Precision (eq. 4.6) and F-measure (eq. 4.8) of the 10 pictures each participant had to interpret in the experiment. As already explained not all participants evaluated all images; thus, the computations were done using the subsets of images shown to the participants according to the questionnaire they were given. In practice, this means that N_I in equations 4.2, 4.6, 4.8 was equal to ten (the number of images in each questionnaire).

Some basic statistics of the per participant Recall, Precision and F-measure are shown in Table 5.3.We see there that the recall performance per participant is 0.55 ± 0.15 with the extreme values being 0.23 (minimum) and 0.78 (maximum). Thus, the conclusion is that at least one out two hashtags used by the

[3] http://www-01.ibm.com/software/analytics/spss/

[4] https://products.office.com/en-us/excel

[5] http://www.mathworks.com/products/matlab/



Questionnaire

Please choose a word or words that describe the image best

□ Cuddles

□ Red

□ Dinner

□ Climatechange

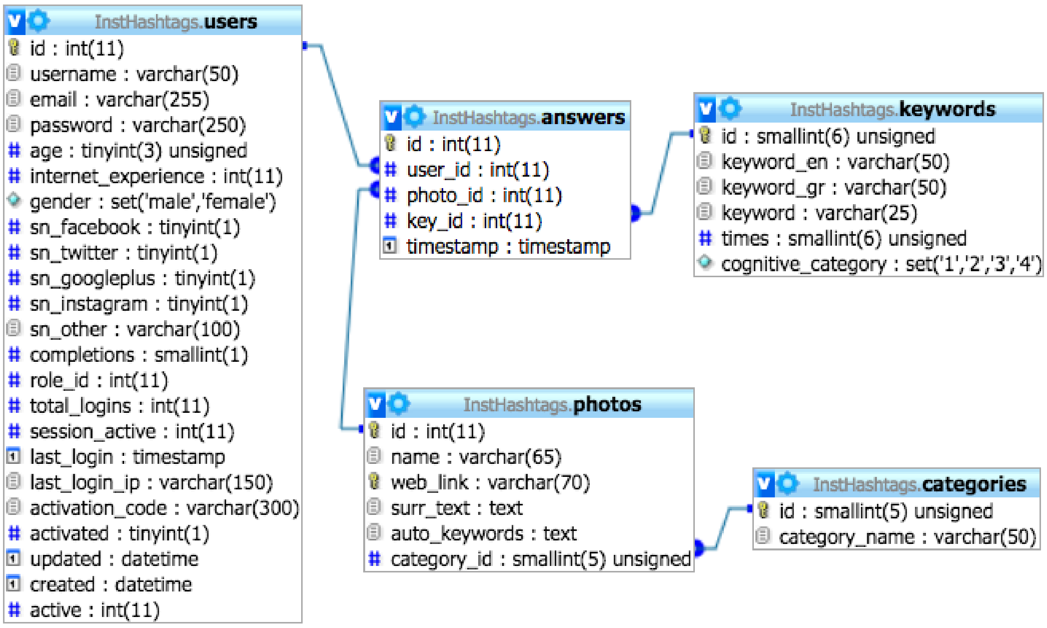
□ Dog

□ Hotdog

□ Ferrari

□ Castle

Figure 5.2: An example of an image interpretation multiple choice question



InstHashtags.users
id : int(11)
username : varchar(50)
email : varchar(255)
password : varchar(250)
age : tinyint(3) unsigned
internet_experience : int(11)
gender : set('male','female')
sn_facebook : tinyint(1)
sn_twitter : tinyint(1)
sn_googleplus : tinyint(1)
sn_instagram : tinyint(1)
sn_other : varchar(100)
completions : smallint(1)
role_id : int(11)
total_logins : int(11)
session_active : int(11)
last_login : timestamp
last_login_ip : varchar(150)
activation_code : varchar(300)
activated : tinyint(1)
updated : datetime
created : datetime
active : int(11)

InstHashtags.answers
id : int(11)
user_id : int(11)
photo_id : int(11)
key_id : int(11)
timestamp : timestamp

InstHashtags.keywords
id : smallint(6) unsigned
keyword_en : varchar(50)
keyword_gr : varchar(50)
keyword : varchar(25)
times : smallint(6) unsigned
cognitive_category : set('1','2','3','4')

InstHashtags.photos
id : int(11)
name : varchar(65)
web_link : varchar(70)
surr_text : text
auto_keywords : text
category_id : smallint(5) unsigned

InstHashtags.categories
id : smallint(5) unsigned
category_name : varchar(50)

Figure 5.3: The database schema used to store annotation data and questionnaire results

Table 5.3: Per participant Recall, Precision and F-measure value statistics

	Mean	St. Dev.	Minimum	Maximum
Recall	0.55	0.15	0.23	0.78
Precision	0.67	0.12	0.47	1.00
F-measure	0.56	0.12	0.33	0.77

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. The per participant precision is significantly higher (0.67 ± 0.12) 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. 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.

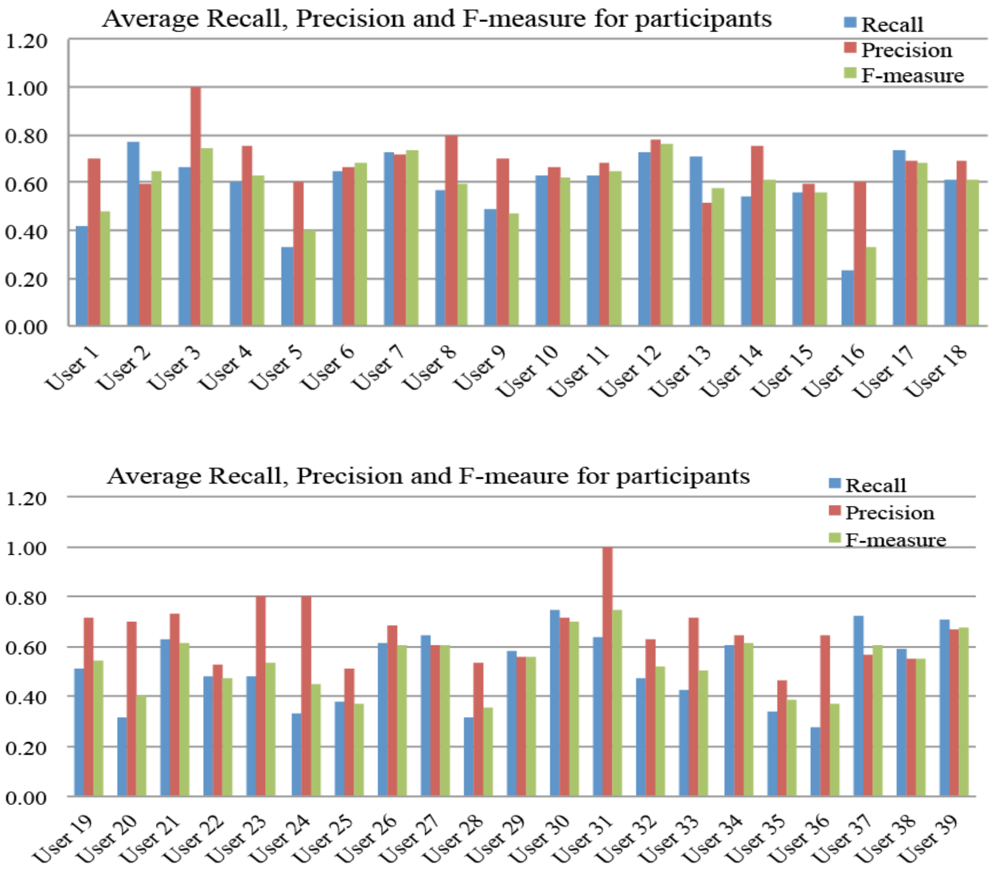


Figure 5.4: Average hashtags' recall, precision and F-measure per participant

Overall, with the aid of Fig. 5.4 and Table 5.3 we can conclude on both research questions set in this study. Given that the participants in our experiments can be seen as experts (librarians and students of Internet and Communication studies) we can claim that around 55% of the Instagram hashtags that accompany images are relevant to the actual content of the images and can be used for training purposes. By pointing out that on average only 40% of the (owner's) Instagram image hashtags are relevant to the images close to which they appear we can state that on average 22% $(0.55 \cdot 0.4)$ of Instagram hashtags are related with the visual content of images.

Figure 5.5 shows the dissimilarity of image interpretation between each one of the 39 participants, of the first stage study, and the photo owners with the aid of (normalized) Hamming distance and the mathe-

Table 5.4: Per image Recall, Precision and F-measure value statistics

	Mean	St. Dev.	Minimum	Maximum
Recall	0.55	0.18	0.15	0.91
Precision	0.67	0.22	0.22	1.00
F-measure	0.56	0.17	0.17	0.90

matical formulation presented in Section 4.2.1. 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 in the 10 questions of the questionnaire they took). As we see in Table 5.5 the average normalized Hamming distance between the photo owners and the participants is 0.245 ± 0.048. This means that there is less than 25% disagreement (only one out of four hashtag 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. 5.5 we see that only two users (those with ids 22 and 25) show somehow low performance (high dissimilarity with the interpretation of picture owners) but even in these cases no random or dishonest behavior can be justified since the average Hamming distance is well below 40%. On the other hand, the users with ids 3 and 31 present 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).

Table 5.5: Statistics of normalized Hamming distance between participants and photo owners in image interpretation

Mean	St. Dev.	Minimum	Maximum
0.245	0.048	0.138	0.375

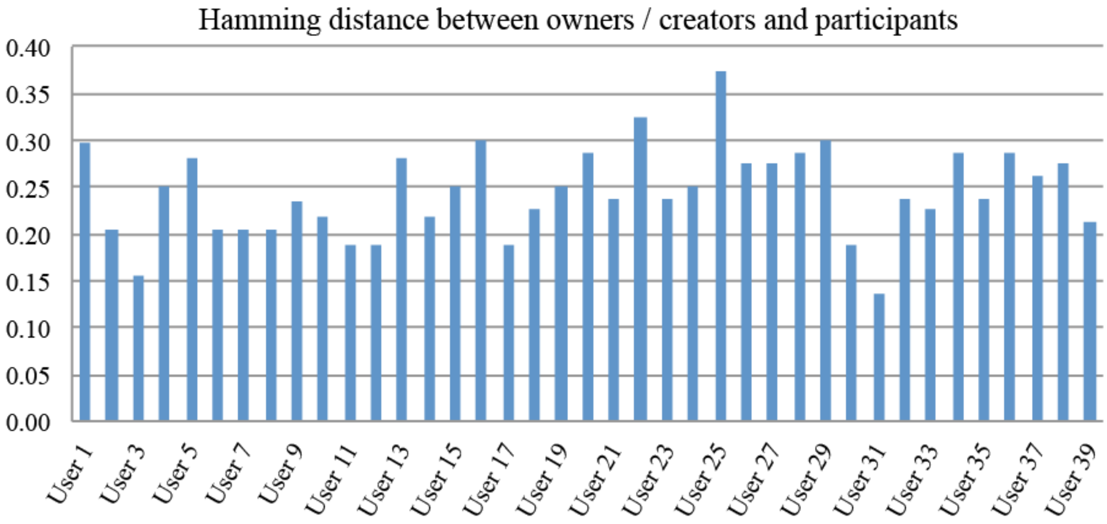


Figure 5.5: The hamming distance between participants and image owners / creators

In Fig. 5.7 we present the per image Recall (eq.4.3), Precision (eq.4.5), and F-measure (eq.4.7) values while in Table 5.4 are shown summary statistics for those values. The basic aim of this analysis is to check whether the difficulty of interpreting images depends on their visual content. Comparing Tables 5.3 and 5.4 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. On the other hand, in Fig. 5.6 we show the images with the lowest recall and precision scores (from left to right images with ids 2, 20 and 28). In a first glance it does not seem that these images present abstract concepts, which are, generally, difficult to interpret. Thus, probably the owners hashtags for these image might be irrelevant with their visual content causing a different interpretation by the experiment participants.

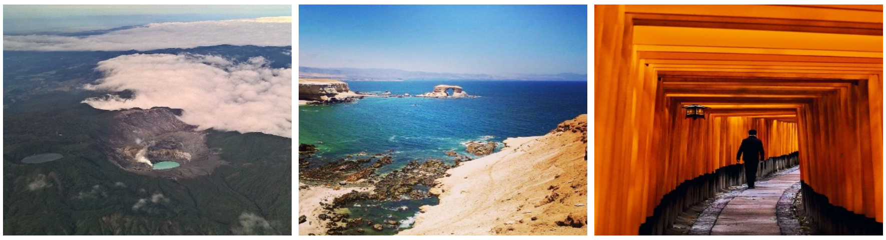


Figure 5.6: Difficult to interpret images

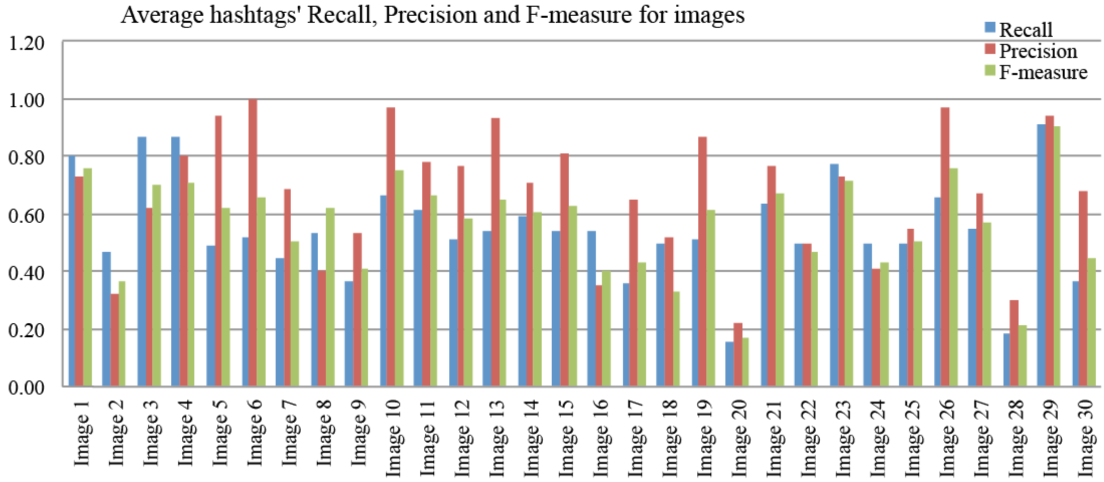


Figure 5.7: Average hashtags' recall, precision and F-measure per image

In the last part of our analysis, regarding the first stage, 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. Figure 5.8 presents the recall values for all owners' hashtags (the set H mentioned in Section 4.2.1). It is clear that abstract concepts tend to have low recall values, as expected, however, the three out of four concepts that have zero recall refer to places (Florida, Chile, Indochina). 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.

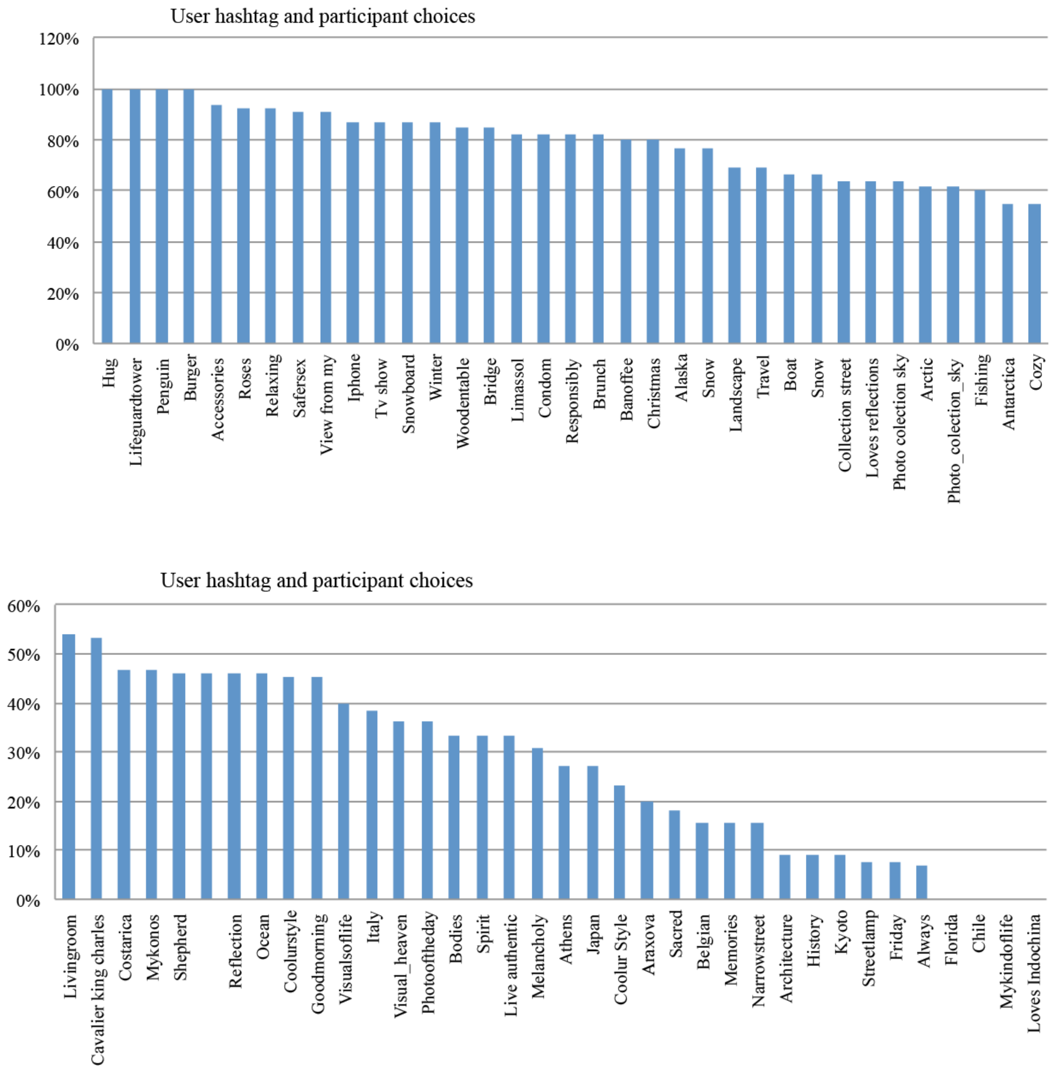


Figure 5.8: Percentage of the participants that chose each of the owner / creator hashtags

5.2.3 The extended study

As in the first stage, the data of the 295 filled in questionnaires of the second stage were analyzed with aid of SPSS[6], MS Excel[7] and MATLAB[8] platform using the metrics defined in the Section 4.2.1. Three users were identified as outliers, due to extremely low F-measure (users with ids 145 and 212) or unexpectedly high number of keywords per image (user with id 203), and their answers were ignored. Fig.5.9 shows the per participants' Recall (eq. 4.2), Precision (eq. 4.6) and F-measure (eq. 4.8) 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 computations were done using the subsets of images shown to the participants according to the questionnaire they were given. Figure 5.11 shows the average hashtags' recall, precision and F-measure for all participants. For ease interpretation the user ids were sorted based on F-measure

[6] http://www-01.ibm.com/software/analytics/spss/
[7] https://products.office.com/en-us/excel
[8] http://www.mathworks.com/products/matlab/

(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 5.6. 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 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 similar findings regarding the number of hashtags accompanying Instagram images [110].

Table 5.6: 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

Overall, with the aid of Fig.5.9 and Table 5.6 we can conclude on both research questions set in this study. 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 present results confirm the preliminary research, conducted in the first stage, that hashtags accompanying Instagram images, are relevant to the actual content of the images. The fact that in the second stage the average Recall value is higher, climbing form from 0.55 to 0.66 than in the previous one, can be explained by the fact that increased participation from 39 to 362 participants. 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% ($0.66 \cdot 0.3$) of Instagram hashtags are related with the visual content of Instagram images.

Figure 5.10 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 5.7 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. 5.12 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 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

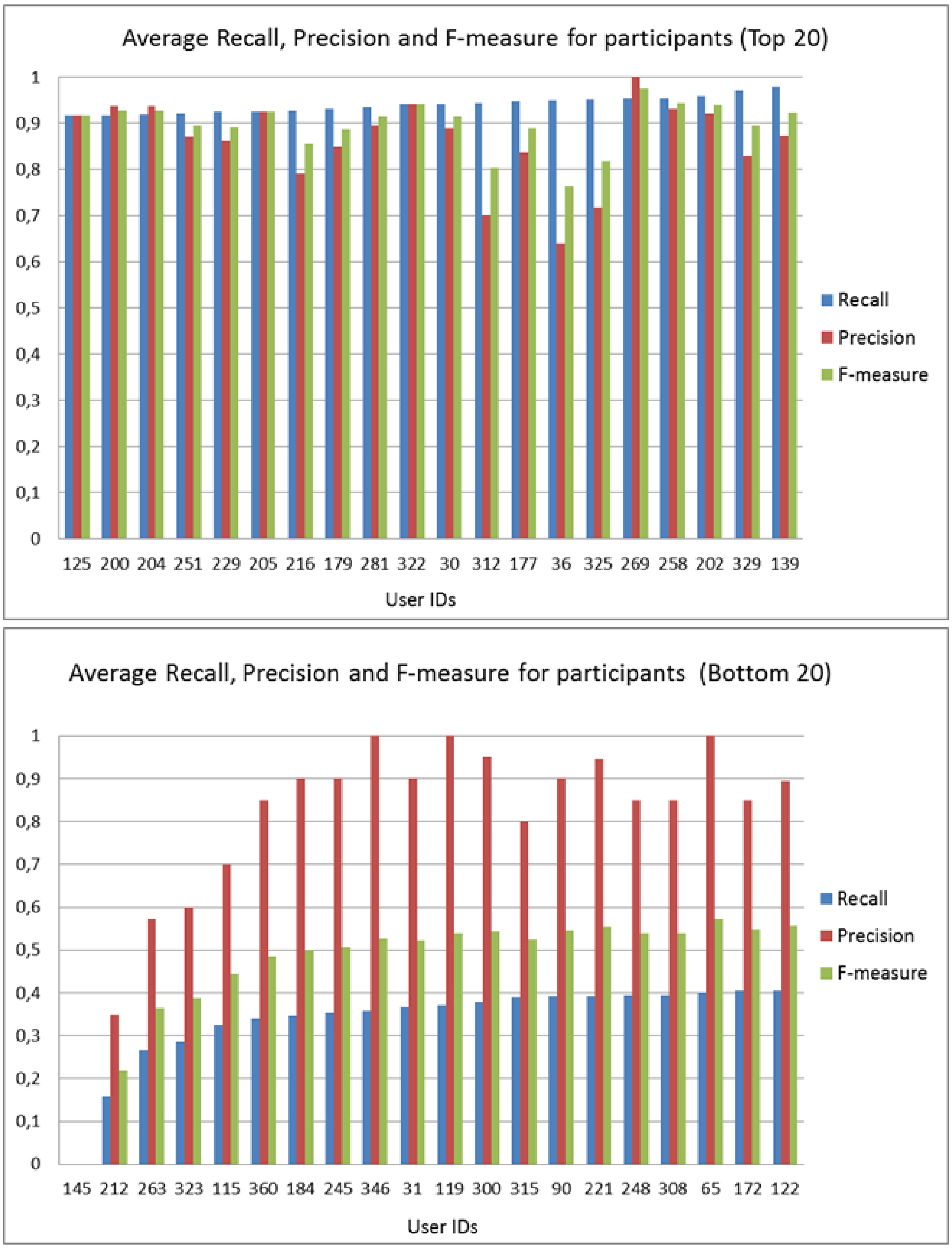


Figure 5.9: Average hashtags' recall, precision and F-measure per participant. The top 20 and bottom 20 performing participants are shown

information).

Table 5.7: 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

In Fig. 5.13 we present the per image Recall (eq.4.3), Precision (eq.4.5), and F-measure (eq.4.7) values while in Table 5.8 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'

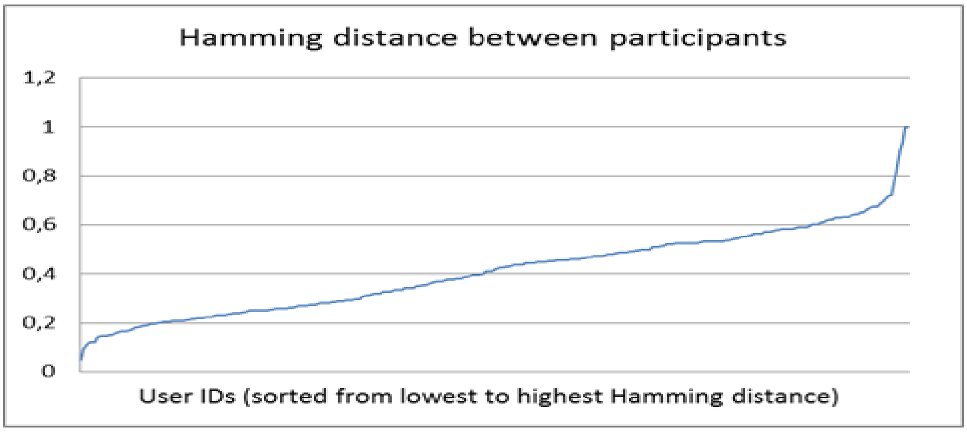


Figure 5.10: The hamming distance between participants and image owners / creators

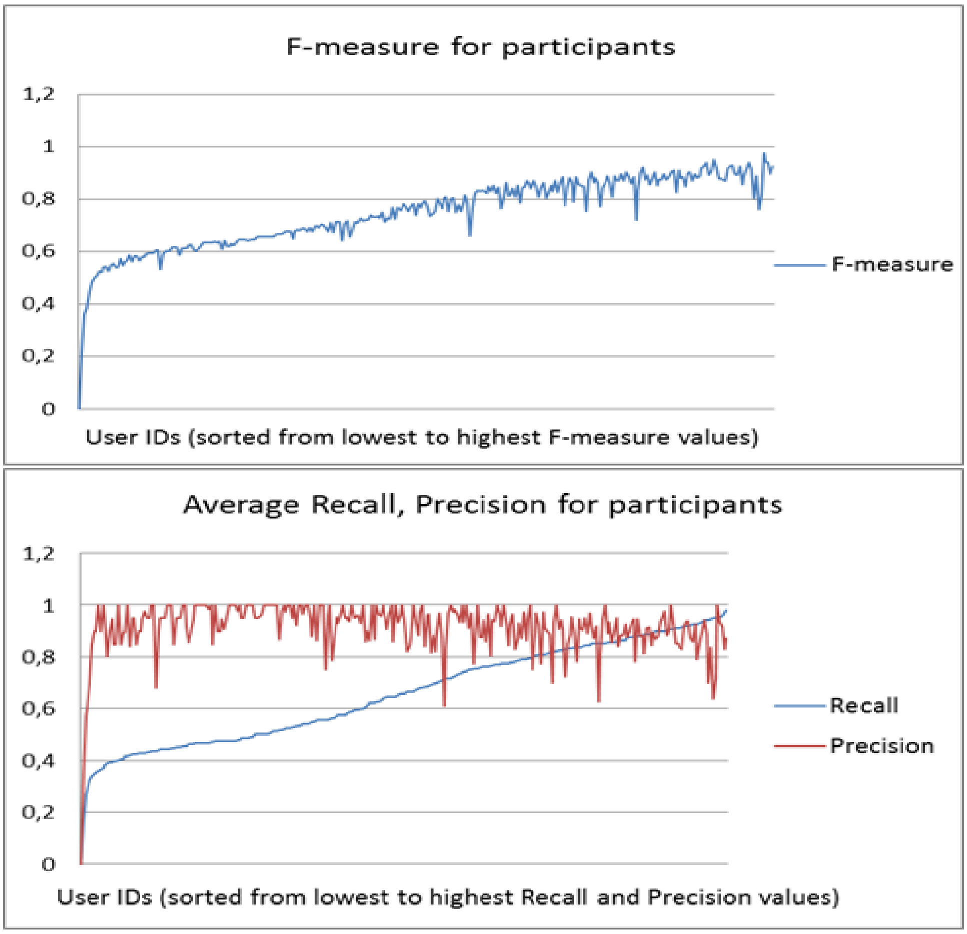


Figure 5.11: Average hashtags' recall, precision and F-measure per participant. For ease interpretation the user ids were sorted based on F-measure (top) and recall (bottom) from lowest to highest values

and were excluded from the calculations. The basic aim of this analysis is to check whether the difficulty of interpreting images depends on their visual content. Comparing Tables 5.6 and 5.8 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. 5.14 we show the images, annotated by at least 35 users each, with the lowest recall scores (from

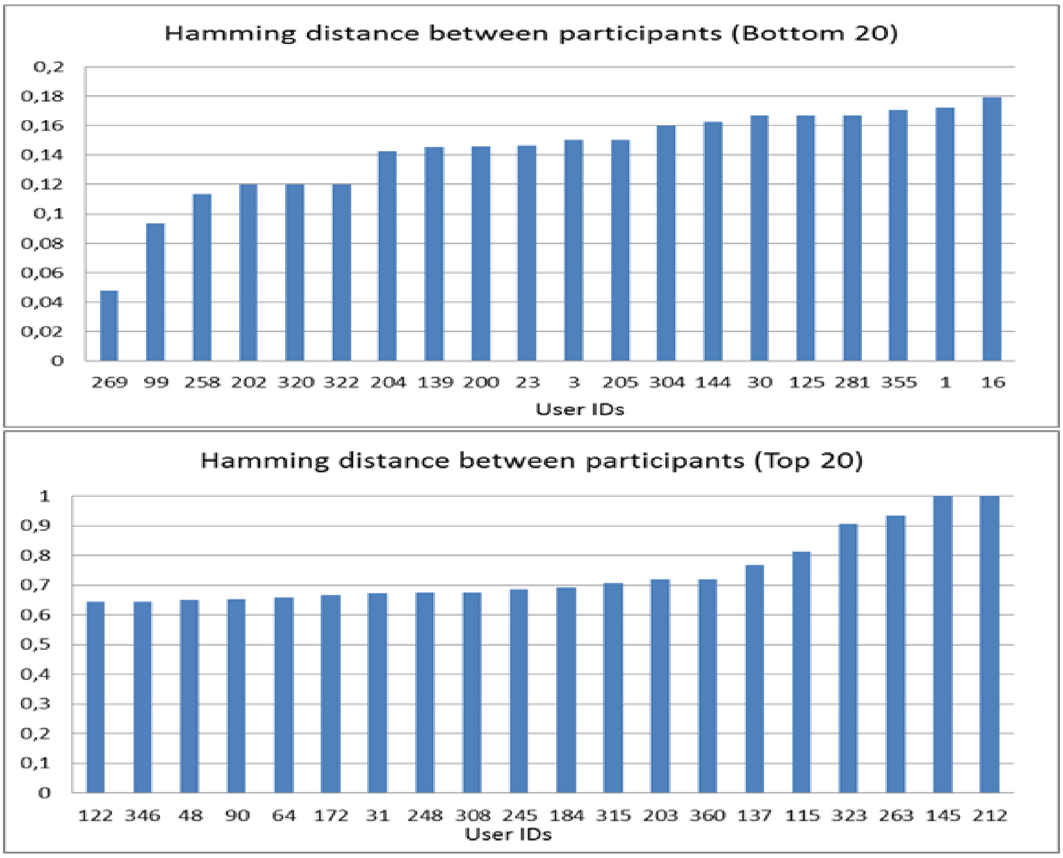


Figure 5.12: Hamming distance between participants and image owners / creators. In the top diagram we show the normalized distances for the 20 users with the best performance (lowest Hamming distance) while in the bottom diagram we show the distances of the 20 least performing users

Table 5.8: 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

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 spring; thus, half of the users made the association and selected spring as a

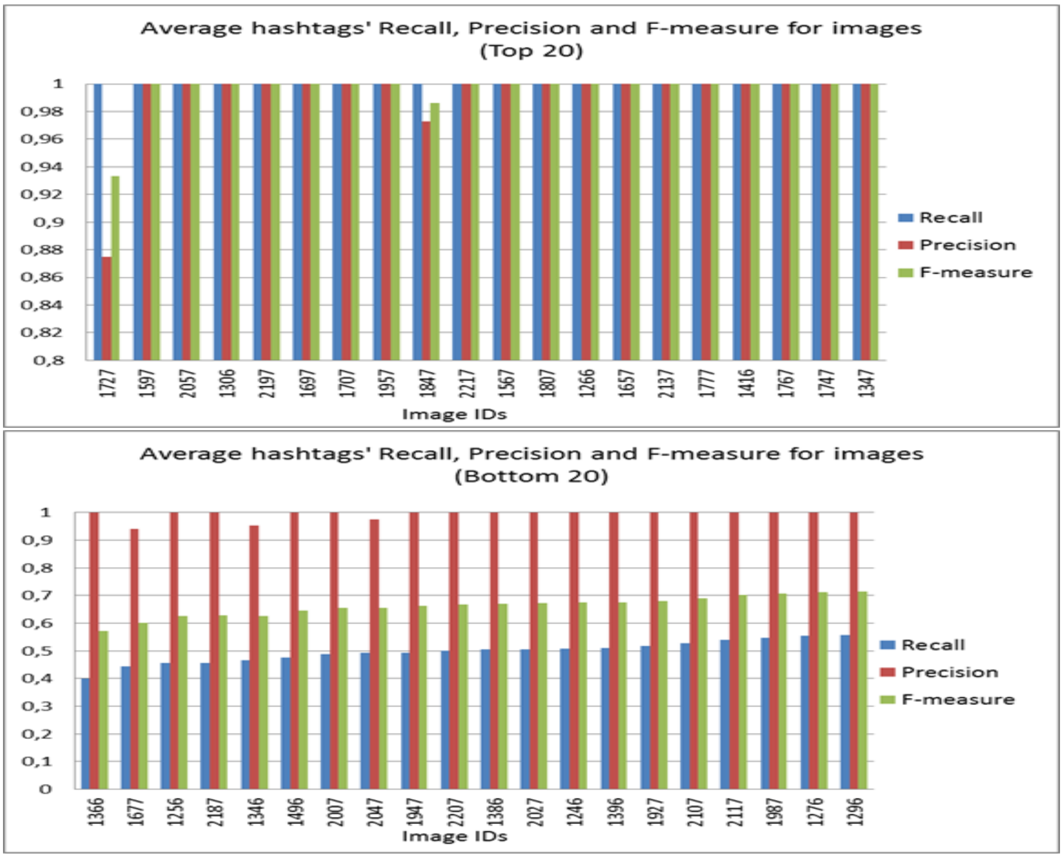


Figure 5.13: Average hashtags' recall, precision and F-measure per image. We show the scores for the 20 most easy and most difficult (in terms of recall) to interpret photos

keyword for this particular photo.

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. Figure 5.15 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').

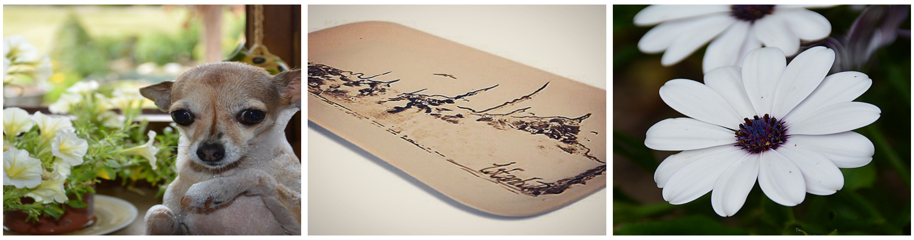


Figure 5.14: Difficult to interpret images

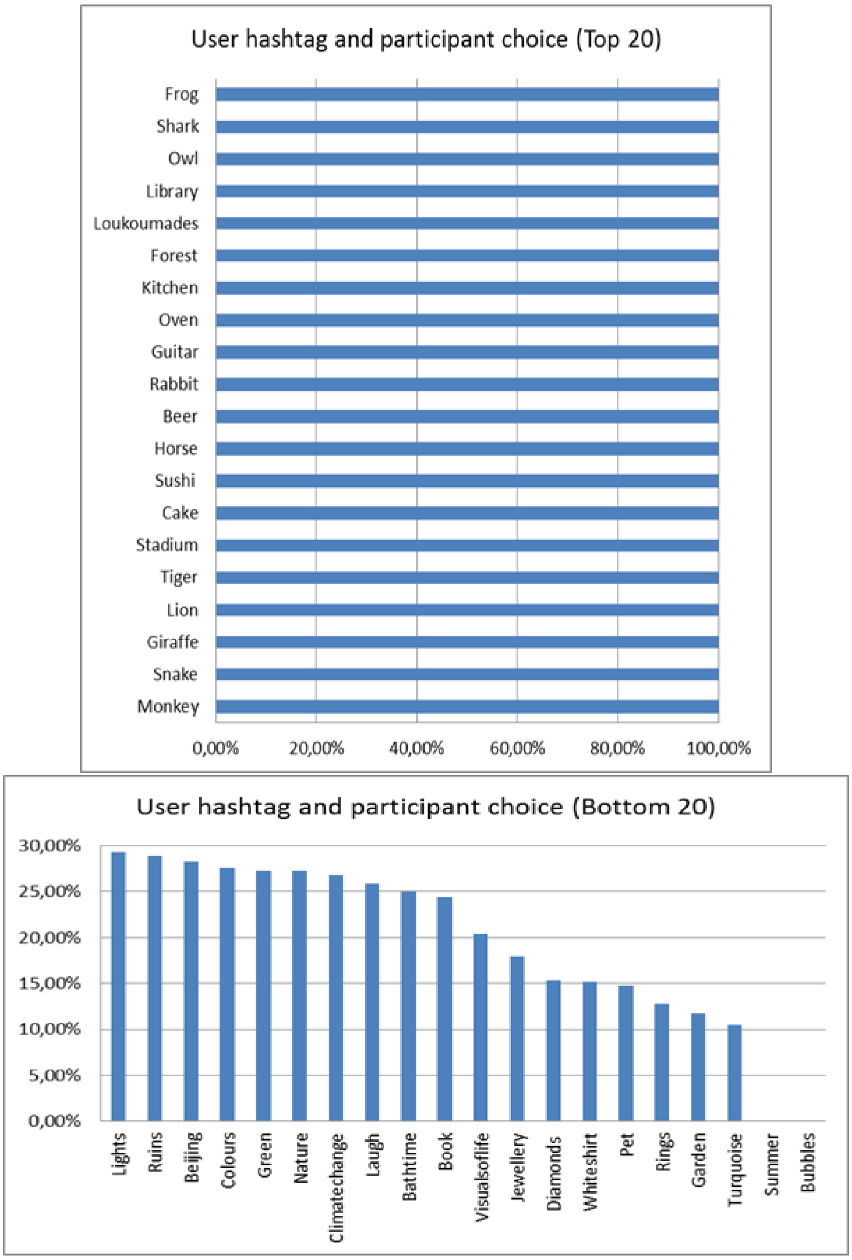


Figure 5.15: Percentage of the participants that chose each of the owner / creator hashtags. The 20 most easily (top) and hardest (bottom) to retrieve hashtags are shown

Table 5.9: Stophashtag list identified using the proposed methodology (with * are denoted stophashtags that confirmed through human evaluation - see Table 5.10)

Hashtag	Score (S_H)	Hashtag	Score (S_H)	Hashtag	Score (S_H)
love*	0.53	like4like*	0.49	likeforlike*	0.49
l4l*	0.49	likes4likes*	0.49	likesforlikes*	0.49
Like4Like*	0.49	L4L*	0.49	l4like*	0.49
instagood*	0.46	beautiful*	0.45	photooftheday	0.42
follow4follow*	0.37	f4f*	0.37	followforfollow*	0.37
picoftheday*	0.35	follow*	0.34	photo	0.34
instalike*	0.32	instalikes*	0.32	igers	0.32
instagramers*	0.32	instagram*	0.29	amazing	0.29
style	0.26	followme*	0.26	travel	0.25
vscocam*	0.24	vsco*	0.24	vscophile	0.24
vscogood	0.24	vscogram	0.24	vsco_hub	0.24
nature	0.24	sun	0.24	spring	0.24
instamood*	0.24	fun	0.24	cute	0.22
food	0.22	handmade	0.22	girl	0.21
photography	0.21	like*	0.21	likes*	0.21

5.3 Stophashtags

This section describes data collection, analysis of results and conclusions regarding the third research question, which deals with the definition and identification of stophashtags. The methodology we followed has been already described in Section 4.3. Further details regarding this study, including the exact methodology we followed, can be found at [17].

5.3.1 Data collection and initial results

In order to apply our methodology (see Section 4.3) first we had to define the N (in our case 30) independent subjects (query hashtags) we would examine to locate the stophashtags. Then for each subject image URLs appeared in the results page were automatically collected using the Beautiful Soup[9] library of Python. Frome each query hashtags we retrieved up to 35 images; from those we chose, manually, only those images that according to our interpretation, are more related to the query hashtag, i.e., retrieval keyword, we used for retrieving the photos. We then collected automatically all hashtags that accompany the selected photos for each subject. The photos and the related hashtags were gathered between 20-25 March 2016.

In order to define which hashtags are stophahstags we used the Otsu method [274] to calculate the threshold T that would be applied to stophashtag score S_H (see eq. 4.12). It appeared that a value of T around 0.20 provides the best classification of hashtags into descriptive ($S_H \leqslant 0.20$) and stophahstags ($S_H > 0.20$). We see in Table 5.9 that 45 hashtags were identified as stophashtags. However, human evaluation was required to confirm which of them are not related with the visual content of images to which they appear as hashtags. Thus, the study described in the next section was conducted for that purpose.

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

Table 5.10: Stophashtags identified through human judgement

Hashtag	Recall	Hashtag	Recall	Hashtag	Recall
cunard	0.00	detailersofinstagram	0.00	flex	0.00
instacool	0.00	iphoneonly	0.00	likealways	0.00
lotr	0.00	motorsport	0.00	police	0.00
phonephotography	0.00	recentforrecent	0.00	tagstagram	0.00
teamfollowback	0.00	tv_transport	0.00	tweegram	0.00
vsco	0.00	webstagram	0.00	igdaily	0.02
F4f (Follow4follow)	0.02	vscocam	0.02	follow	0.03
followme	0.03	artesanal	0.04	corn	0.04
likers	0.04	paint	0.04	string	0.04
insane	0.04	sick	0.04	followforfollow	0.05
instagood	0.05	instadaily	0.05	instafollow	0.06
australiagram	0.06	igersaustria	0.06	instagramers	0.06
scratchmap	0.06	tagsforfollow	0.06	tv_sea	0.06
L4l (Like4like(s))	0.08	instagram	0.07	bestoftheday	0.08
instalike(s)	0.08	crocodile	0.08	facebook	0.08
followers	0.08	tagsforlike(s)	0.08	instamood	0.11
corncobpipe	0.12	frenchieoftheday	0.12	gourmet	0.12
landscape	0.12	instaphotos	0.13	lionporn	0.13
pic(ture)oftheday	0.15	instapic(ture)	0.16	instacolourful	0.16
love	0.17				

5.3.2 Human judgment evaluation

To evaluate the proposed methodology for stophashtag identification, human judgment is needed to verify the results obtained automatically (see Table 5.9). So, we asked users to select from a list of tokens (here hashtags), composed from the hashtag list that appends to an image randomly selected stophashtags, the ones that are related to the visual content of a photo in question. The participants were not aware that among the given choices stophashtags were added.

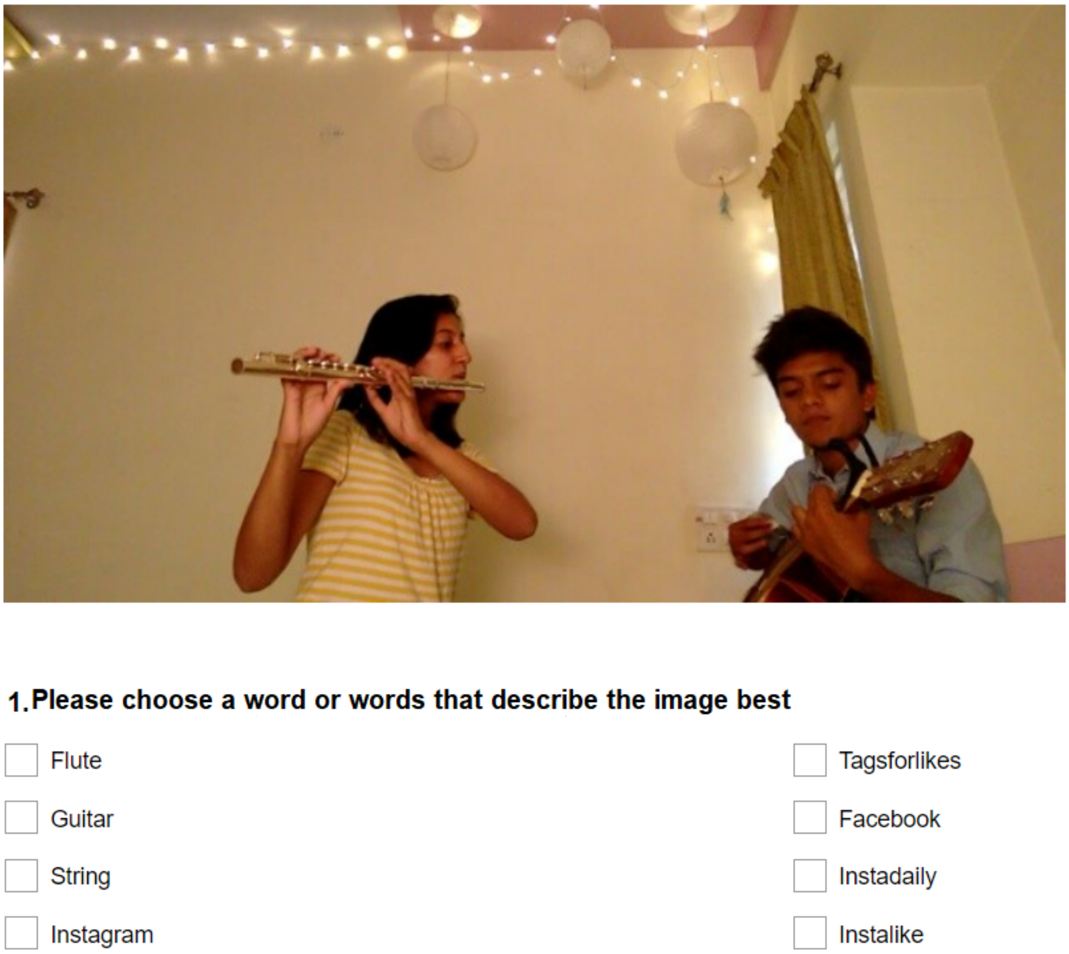
Human judgement was designed as follows: Users were asked to choose among hashtags associated with a photo the ones that describe the visual content of the photo in question. From the retrieved, through the N subjects, photos 30 were randomly selected and included in a online questionnaire which was delivered to users through SurveyMokey[10]. Participants had to select among eight hashtags, including both descriptive hashtags and stophashtags but all of them assigned to the photo by its owner / creator, for each one of the questionnaire photos.

The 30 photos were put into two separate questionnaires, containing 15 pictures each, in order to reduce the fill in time and avoid fatigue effects. Among the eight choices given for each photo of 2-3 correspond to descriptive hashtags and the rest to stophashtags. As we see in Figure 5.16 the words flute, guitar and string are descriptive and instagram, tagsforlikes, tagsforlike, facebook, instadaily, instalike are meaningless hashtags (regarding the content). In any case, participants were not aware that any of the given choices were considered either descriptive or stophashtags; thus, they were free to select as many of them as they wished according to their interpretation of the shown photo.

Choices not selected by participants are likely to correspond to stophashtags since participants consider them as not descriptive for the photo presented to them. In that context we can use effectiveness measures

[10] http://www.surveymonkey.com/

such as Precision and Recall to identify which of the identified by the algorithm stophashtags were also -indirectly- classified as such by humans and vice versa. For this purpose, however, we needed to create the list of stophashtags according to human judgement.



1.Please choose a word or words that describe the image best

□ Flute

□ Guitar

□ String

□ Instagram

□ Tagsforlikes

□ Facebook

□ Instadaily

□ Instalike

Figure 5.16: An example of an image interpretation multiple choice question

Given that the participants can select any token in the list, according to their interpretation, the frequency of selection of hashtags should be inversely proportional to the stophashtag score S_H obtained via eq. 4.12 and shown in Table 5.9. Thus, if H_T denotes the times that a hashtag is selected by the users and H_S denotes the times that users have the possibility to choose it as descriptive for a picture's visual content, the descriptive score D_H for each hashtag H is given by:

$$D_H = \frac{H_T}{H_S} \tag{5.1}$$

As already mentioned above, we argue that D_H and H_S would be inversely proportional. For instance consider that a hashtag was selected two times and that hashtag was given to the users as a choice only for one image. In case this photo was accessed 25 times, that is appeared in 25 filled questionnaires, then D_H equals $2/25$. Based on D_H and by using a threshold T_D we can classify the hashtags to descriptive ($D_H > T_D$) and stophashtags ($D_H \leqslant T_D$) according to human judgement. The threshold T_D was, again, computed using the Otsu method [274] and found to be 0.17

Table 5.10 shows the stophashtags identified through human judgement with the procedure explained above. By comparing Tables 5.9 and 5.10 we can see that 26 hashtags appear in both tables while the 17 hashtags with the highest stophashtag score S_H, shown in Table 5.9, were -indirectly- verified by human judgement. Thus, we have a clear indication about the effectiveness of the proposed method especially as fas as the definition of stophashtag score (see eq. 4.12) is concerned. On the other hand, there are some hashtags identified by the proposed algorithm, such as #picoftheday, #vscogood, #vsco_hub, etc., that were not confirmed by human judgement although it is clear that they lack descriptive power.

For a typical information retrieval based evaluation we consider the list of hashtags in Table 5.10 as the benchmark set and compute the recall R, precision P and F-scores. This gives us $R = 26/58 = 0.448$, $P = 26/45 = 0.578$ and $F = 0.505$. From the above discussion we conclude that the thresholds obtained through the Otsu method are not optimal in terms of precision and F-score. For instance a value of T equal to 0.30 would lead to $R = 20/58 = 0.345$, $P = 20/22 = 0.909$ and $F = 0.521$.

5.4 Graph-based data collection and results

In this section we describe data collection and present and analyze the results of the application of the methodologies described in Section 4.4 for Instagram hashtags filtering using graph-based methods and especially the HITS algorithm. The study was conducted in two stages and further details can be found at [18, 19].

5.4.1 Data Collection

A two-stage research was conducted. In the first stage (pilot study) research a subset of 100 photos from a set of 1000 Instagram images, we used in our previous experiment(see Section 5.2), was used. For each one of the 100 images we have manually selected 1-4 hashtags which, according to our interpretation, better describe its visual content. These hashtags consist the ground truth which we used to evaluate the proposed method.

In the second stage (extended study) we applied the proposed methodology on the data collected using a real crowdtagging environment facilitated by the *Figure-eight*, formerly known as *Crowdflower*, crowdsourcing platform (for details for *Figure-eight* see 5.4.2). 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, in the second stage we used also FolkRank as baseline to evaluate the performance of the proposed method.

5.4.2 The *Appen* interface

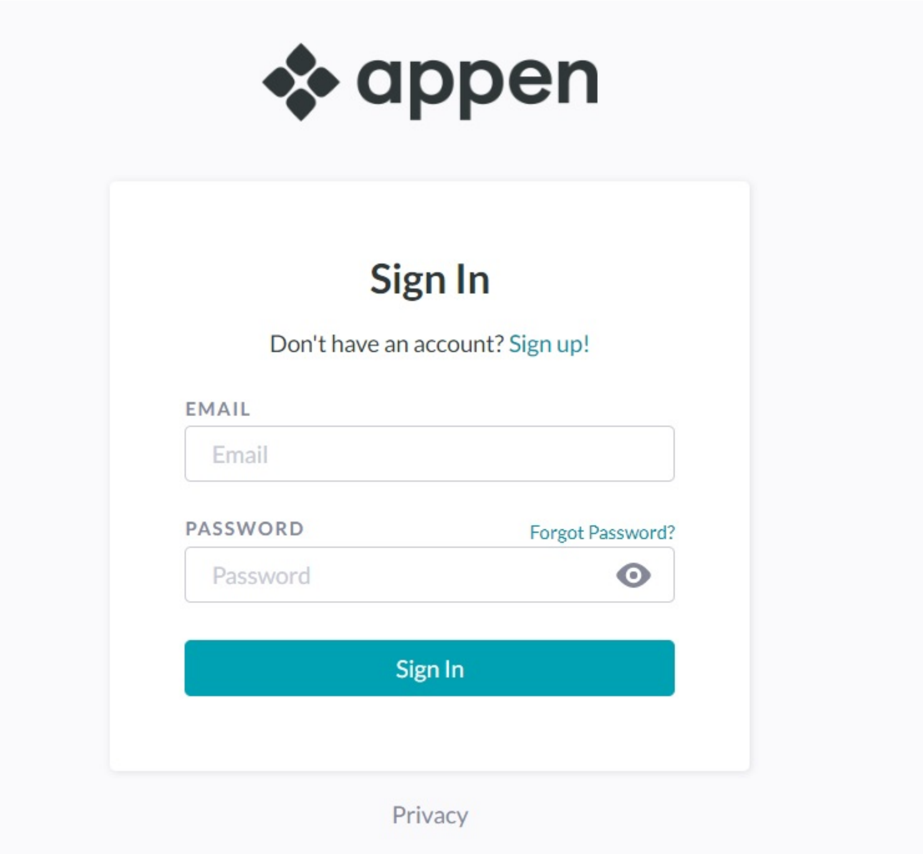
Using crowdsourcing platforms to assign tasks to the crowd can be conducted in return of a small fee and to fulfill the independence criterion of collective intelligence [275]. *Appen* formerly known as *Figure-eight* and *CrowdFlower* is a platform that offers labor as a service[11]. It also gives researchers the possibility "hire" workers to annotate their data. Effective or dishonest workers are evaluated on the basis of their performance on test data, i.e., data annotated by task creator [276]. A number of test questions (test

[11] https://techcrunch.com/2016/06/07/crowdflower-series-d/

data), varied depending on the total number of items to be annotated, are randomly presented to workers to check if the answers they give match the gold standard (annotations by the creator). In case a worker fail on more than 50% of the test data then the system discards all of his/her answers and the worker is excluded from the rest of the annotation task.

Appen[12] offers, among other annotation services, image tagging with low cost in a quick and reliable manner. The pre-conditions of the *Wisdom of Crowds* theory are well-supported in Appen, as in most contemporary crowdsourcing platforms. Thus, for tasks where an different opinions are sought, as in the case of image tagging, Appen can be used to substitute the need of hiring experts. Allowing the crowd to select the appropriate hashtags for a given image allows us to construct image-tag pairs that can be used for AIA tasks.

In order to use the Appen platform it is important to create an account and then login[13] through the interface shown in Fig 5.17. Once the user enters the Appen platform (see Fig 5.18) they can choose the type of tasks they would like to create. The user to complete the task have to upload the data, design the 'user interface', i.e., the guidelines that will be presented to annotators (in the Appen language they are called *contributors*) to help the fulfill the tasks, choose test questions to check the quality of the participants in the task and finally make the task available to the crowd.



Don't have an account? Sign up!

Figure 5.17: Connection with Appen crowdsourcing platform

[12] https://appen.com/
[13] https://client.appen.com/sessions/new

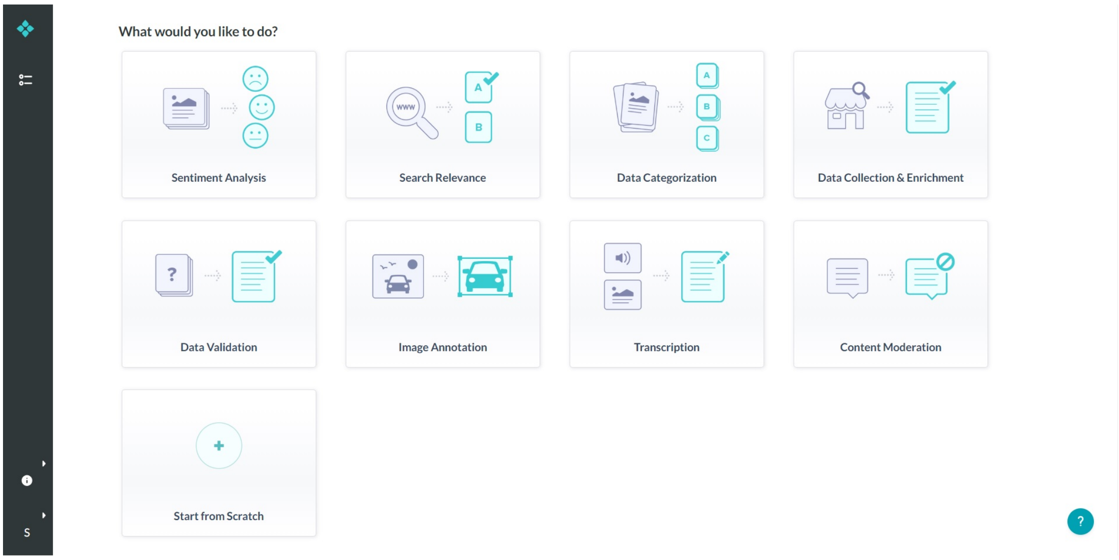


Figure 5.18: Job creation in Appen crowdsourcing platform

5.4.3 The results of the pilot study

The overall Recall, Precision and F-measure scores are shown in Table 5.11. We see there that the average recall value across all images is 0.93 with a standard deviation equal to 0.23. This value is quite impressive if we take into account that the proposed method is learning free. Thus, applying the HITS algorithm for the selection of the appropriate hashtags, for Instagram images, in a crowdsourcing environment is, at least promising.

The obtained precision score is lower than recall. Maximising recall instead of precision is a typical choice in information retrieval applications; it is preferable to have some non-descriptive tags for an image instead of having no tags at all. Our choice to consider the first four tags in terms of authority value, as those describing the image in question, is definitely a heuristic choice causing bias towards recall. In a training set construction scenario, where the aim is to keep only the descriptive tags, precision maximisation should be pursued. In this case thresholding the authority values to discriminate the relevant hashtags from the non-relevant ones is an obvious choice. However, identifying an appropriate threshold value is not always easy.

Table 5.11: Recall, Precision and F-measure scores for M=100 images

(M=100, N=281)	Mean	St. Dev.
Recall (R)	0.93	0.23
Precision (P)	0.57	0.28
F-measure (F)	0.71	0.25

5.4.4 The results of the extended study

A set of 50 Instagram images, along with their hashtags, were automatically crawled with the aid of a Python[14] program. The collected Instagram images were uploaded to *Figure-eight* for crowdtagging

[14] https://www.python.org/

in the form of tag selection as indicated in Figure 4.7 for image #7. To simplify the process all hashtag choices were presented to the annotators as checkboxes. The annotators were invited to select 1-4 hashtags and were given also the opportunity to provide their own tags. Despite these guidelines many annotators select much more than 4 tags and in several cases the extra tags they provided were already among the given choices. Therefore, duplicate tags for the same image were identified and removed. Another important pre-processing step was the splitting of hashtags into their constituting words with the help of the *wordsegment*[15] Python library. For instance, the hashtag *#picoftheday* is decomposed into the words *pic*, *of*, *the* and *day*.

Every image was annotated by 500 annotators for experimentation purposes. In practice much fewer annotations per image are enough while there is absolutely no reason that all images must be assessed by all annotators. Nevertheless, we made those choices to allow us generalize the conclusions of our study as much possible. One of the annotators turned out to be dishonest as indicated by the *_trust* value of *Figure-eight* as well as by the corresponding hub value of the HITS algorithm when it was applied on the full bipartite graph (eq. 4.14), and she / he was excluded from the experiments. Comparison between hub values and *trust* scores are given in Section 5.4.5. The full bipartite graph and the bipartite graphs per image were constructed and analyzed with help of the NetworkX[16] library of Python. We also used the NetworkX implementation of the HITS algorithm to extract the overall hub values (reliability scores for the annotators) and authority scores of the tags of each image.

The 50-Instagram-Image questionnaire was given to the *Figure-eight* annotators. Additionally, two image retrieval experts have acess to the same data set. The annotations of the experts, aggregated together and pre-processed in the same way as the crowdsourced data, consist our gold standard across which the effectiveness of the proposed methodology is evaluated through the measures defined below. In total 145 different tags were proposed by the experts for the 50 images. On the other hand, the 499 annotators proposed a total of 2571 different tags. However, only 135 of the tags proposed by the experts were also proposed by the annotators.

Authority threshold value θ/ FolkRank ranking score threshold value							
Algorithm	**(**M=50**,** N=499**)**	0.25	0.21	0.17	0.15	0.13	0.11
HITS	Recall (R)	0.136	0.223	0.359	0.440	0.527	0.620
AUC = 0.692	Precision (P)	0.962	0.932	0.904	0.862	0.822	0.755
	F_1-measure (F)	0.238	0.360	0.514	0.583	0.642	**0.681**
FolkRank	Recall (R)	0.158	0.261	0.370	0.424	0.504	0.603
AUC = 0.689	Precision (P)	0.935	0.923	0.895	0.876	0.823	0.766
	F_1-measure (F)	0.270	0.407	0.523	0.571	0.626	**0.675**
_trust	Recall (R)	0.168	0.272	0.353	0.424	0.527	0.609
AUC = 0.680	Precision (P)	0.929	0.903	0.877	0.847	0.813	0.772
	F_1-measure (F)	0.286	0.418	0.504	0.565	0.640	**0.681**

Table 5.12: Recall, Precision and F_1-measure scores for M=50 images and various threshold values w.r.t. authority score (HITS), *_trust* weighting and FolkRank ranking score

The Precision, Recall and F_1 measure, as defined in eq. 4.22-4.24, were computed for a variety of author-

[15] http://www.grantjenks.com/docs/wordsegment/

[16] https://networkx.github.io/

Authority threshold value θ/ FolkRank ranking score threshold value						
Algorithm	**(M=50, N=499)**	0.09	0.07	0.05	0.03	0.01
HITS	Recall (R)	0.679	0.712	0.766	0.804	0.842
AUC = 0.692	Precision (P)	0.654	0.604	0.504	0.396	0.265
	F_1-measure (F)	**0.667**	0.653	0.608	0.530	0.403
FolkRank	Recall (R)	0.663	0.707	0.755	0.804	0.832
AUC = 0.689	Precision (P)	0.709	0.613	0.529	0.418	0.277
	F_1-measure (F)	**0.685**	0.657	0.622	0.550	0.415
_trust	Recall (R)	0.652	0.696	0.739	0.798	0.856
AUC = 0.680	Precision (P)	0.698	0.601	0.517	0.412	0.267
	F_1-measure (F)	**0.674**	0.645	0.609	0.543	0.407

Table 5.13: Recall, Precision and F_1-measure scores for M=50 images and various threshold values w.r.t. authority score (HITS), *_trust* weighting and FolkRank ranking score

Number of mined hashtags kept (k)											
(M=50, N=499)	1	2	3	4	5	6	7	8	9	10	11
Recall (R)	0.234	0.467	0.603	0.685	0.750	0.772	0.808	0.815	0.837	0.842	0.848
Precision (P)	0.862	0.858	0.740	0.630	0.552	0.473	0.426	0.375	0.342	0.310	0.284
F_1-measure (F)	0.368	0.605	**0.665**	**0.656**	0.636	0.587	0.558	0.514	0.486	0.453	0.425

Table 5.14: Recall, Precision and F_1-measure scores for M=50 images and various values of the top ranked hashtags based on the authority score

ity threshold values θ and are presented in Tables 5.12 -5.13. Moreover, we present the Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) results according to the eq. 4.25-4.26, in Table 5.15. The corresponding Receiver Operating Characteristic curves[17] (ROC) are shown in Figure 5.19. For convenient juxtaposition with the values presented in Tables 5.12 -5.13, in this ROC curve it is plotted the Precision versus Recall instead of the typical case of ROC curves in which are usually plotted the True Positive Rate versus the False Positive Rate. We observe from both Tables 5.12 -5.13 and Figure 5.19 that the best results in terms of the F_1 measure is obtained for an authority score threshold value θ=0.11. However, as in most information retrieval systems we usually prefer a higher value of Recall, that is identifying more tags even if they are not that accurate, instead of Precision. Thus, an authority score threshold θ=0.09 give us also a reasonable choice.

With a MAP score equal to 0.891 (see Table 5.15) we can conclude that applying the HITS algorithm for the selection of the appropriate hashtags, for Instagram images, in a crowdsourcing environment is, at least promising. Since, MAP ranges [0,1] and the result is close to 1, we can conclude that the algorithm located almost all the relevant hashtags of the collection. Another indication that the proposed methodology is suitable for locating relevant hashtags is the MRR results (see also Table 5.15). Values for MRR range from 0 to 1, with higher values signify that the relevant hashtags are ranked higher. Thus, MRR=0.5 corresponds to the correct hashtags being in the top two returned by the HITS algorithm. Another important metric that is used to evaluate the performance of information retrieval systems is the Area Under the (ROC) Curve (AUC or AUROC). Since both Precision and Recall take values in the

[17] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

	Mean	Min	Max
Average Precision	0.89	0.51	1.00
Reciprocal Rank	0.52	0.16	1.00

Table 5.15: Mean Average Precision and Mean Reciprocal Rank for for M=50 images

range $[0, 1]$, AUC also ranges in $[0, 1]$. The intuition behind this metric is that an AUC of 0.5 represents a random information retrieval system (or, similarly, a uninformative two-class classifier) while an AUC equal to 1 represents the perfect information retrieval system. The AUC corresponding to the ROC curve of Figure 5.19 is equal to 0.692. As we show in the Appendix (*Step 6*) the computation was done with the aid of the *metrics*[18] Python library of *Sklearn*[19].

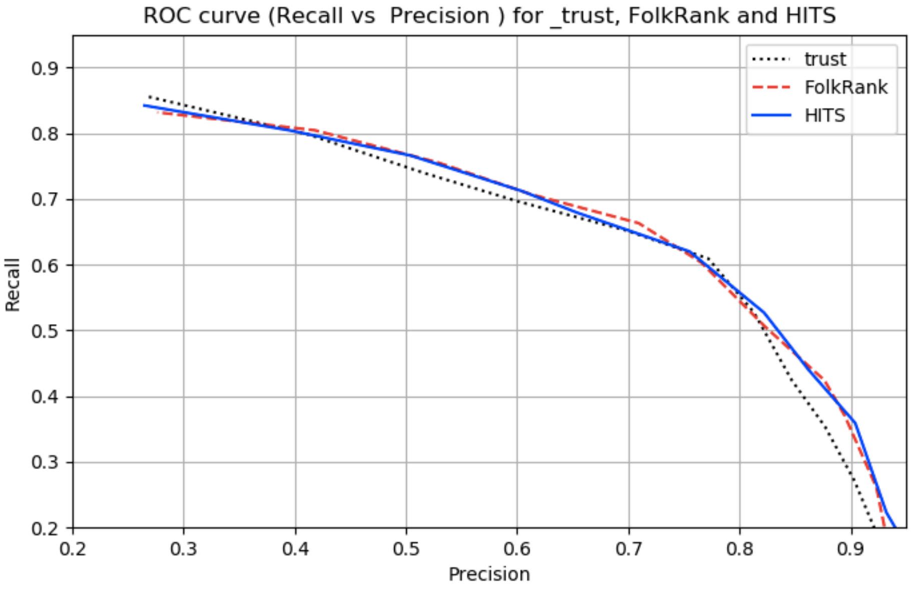


Figure 5.19: Recall vs precision ROC curves for the *_trust* (AUC = 0.680), the FolkRank (AUC = 0.689) and the HITS (AUC = 0.692) weighting schemes

As we seen in the previous section 5.2 we concluded that on average four of the hashtags accompanying each Instagram image are related to its visual content. This conclusion was inline with the findings of Ferrara *et al.* [110] who studied users' behavior while they annotate their photos with hashtags and concluded that users use quite a few hashtags in order to annotate image content. In order to verify these findings we also evaluated, again with the aid of the gold standard set, the effectiveness of hashtags' selection through the HITS algorithm by keeping the k top ranked hashtags per image based on their authority scores. The results, for a variety of k values, are shown in Table 5.14 while the corresponding ROC curve is shown in Figure 5.20. We see that the best F_1 scores are achieved by keeping either the top three or the top four ranked hashtags per image. Keeping four hashtags per image favors the recall value which, as already discussed above, is preferable for the majority of information retrieval systems. We see also in Figure 5.20 that the area under the curve (AUC) is 0.675, which is comparable with the authority score thresholding case. This means that there is no significant variation of the agreed hashtags

[18] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

[19] https://scikit-learn.org/stable/

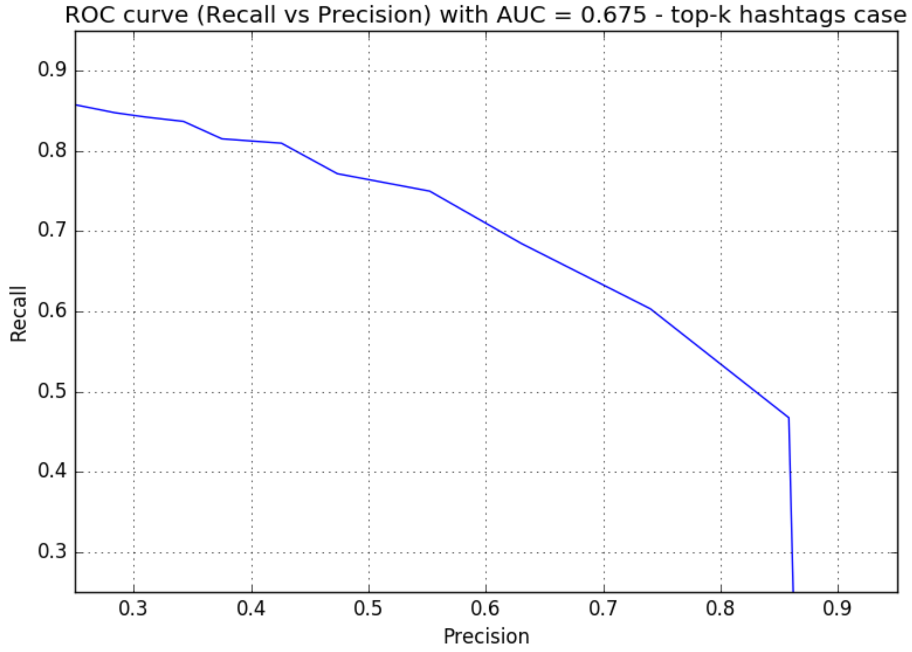


Figure 5.20: Recall vs precision ROC curve with an area under the curve (AUC) equal to 0.675 - the case of top-k hashtags

per image; so keeping the k top ranked hashtags based on the authority score is another option for mining tags from Instagram hashtags accompanying images.

User ID	**hub value** $x10^{-2}$	**hub based ranking**	**FolkRank value** $x10^{-2}$	**FolkRank ranking**	***_trust value***	***_trust* based ranking**
xx7892	0.3195	1	0.1444	1	0.6665	490
xx5795	0.3060	2	0.1372	2	0.7104	462
xx7746	0.3045	3	0.1363	3	0.6688	487
xx9591	0.3020	4	0.1350	4	0.6504	496
xx8610	0.2964	5	0.1320	5	0.7308	419
xx3452	0.2939	6	0.1306	6	0.6547	493
xx0988	0.2931	7	0.1302	7	0.6351	497
xx1052	0.2912	8	0.1291	8	0.7306	422
xx8286	0.2909	9	0.1290	9	0.7367	404
xx2687	0.2888	10	0.1278	10	0.7402	389

Table 5.16: Top 10 users according to the hub value along with their corresponding ranking based on *Figure-eight*'s *_trust* value

.

5.4.5 Reliability measures for the annotators

Figure-eight, as many other crowdsourcing platforms, provides its own measure to identify dishonest annotators. In particular it uses the *_trust* variable which is computed on a subset of the data, known as Gold Test Questions, for which the creators provide the correct answers and which is considered as a type of gold standard. In our case, an additional set of Instagram images corresponding to 10% of the data was assessed (crowdtagged) by the creators. The performance of each one of the annotators is the recall value of the tags used by the creators that the annotator correctly identified.

As already mentioned, in the proposed method the reliability of the annotators is estimated with the aid of the hub value computed on the full graph composed from all images and all tags (see eq. 4.17). So the annotators reliability is based on the total number of image for hub value on contrast to the calculated for all the _trust value that is based on 10% of the data. In Table 5.16 we present the hub values of the top 10 reliable annotators based on our method along with the corresponding *_trust* value as computed by *Figure-eight*. In the same table we show also, the corresponding ranking of the differential FolkRank algorithm. While the rankings of annotators based on the hub scores and the FolkRank algorithm are identical, as they both based on the same principle, we observe large differences between them and the *_trust* values (fifth column) of *Figure-eight*. In fact the *_trust* values, of the top 10 annotators based on the hub scores and FolkRank, are below the average *_trust* value (0.7675) and in almost all cases the corresponding ranking is in the last 100. We remind here that the total number of annotators is N=499.

We observe also, by examining the extreme values of hub and *_trust*, that the hub scores provide a more subtle diversification than the *_trust* scores. Therefore, our choice to weight the bipartite graphs for each image (see eq. 4.18) with the hub scores of the full bipartite graph rather than the *_trust* values seems justified. However, in order to empirically check this assumption we repeated our experiments by using as weights in the bipartite graphs for each image the *_trust* scores of the annotators. The results are summarized in Tables 5.12 -5.13 and illustrated in Figure 5.19. We see a quite similar performance in terms of the F_1 metric although some differentiation between Recall and Precision for the same values of the authority threshold θ do exist. The area under the curve achieved when using the *_trust* scores to weight the bipartite graphs is 0.680, not very much lower than that of the hub score weighting of the bipartite graphs. We further discuss this finding in Section 6.3.

5.4.6 Python Code

Here we provide the full Python code that allows anyone who wishes to re-run the experiments and test their validity. The graphs as Pajek[20] files are also publicly available at https://irci.eu/insta-hashtags/

- *Step 1*: Read the datafile produced through crowdsourcing (already converted to *json*[21] format)

```
>>> import json
>>> with open('../data/F8_data.json', 'r') as fp:
 ... data = json.load(fp)
>>> users = list(data.keys())
>>> data[users[0]].keys()
```

- *Step 2*: Create a full bipartite graph composed by annotators and all available tags in order to rank the annotators.

```
>>> import networkx as nx
>>> import numpy as np
>>> exec(open('csv2imageGraphs.py').read())
>>> G = FullGraph(data,50,no_split, '../data/full499.net')
```

[20] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[21] https://www.json.org/

- *Step 3*: Apply the HITS algorithm and get the hub values (h).

```
>>> [h,a] = nx.hits(G)
```

- *Step 4*: Use the hub values (h) computed in the previous step to initialize the bipartite graphs for each one of the images.

```
>>> ImageGraphs(data,50,h,no_split)
>>> G7 = nx.read_pajek('../data/img7.net')
>>> [annotators,tags] = nx.bipartite.sets(G7)
>>> list(sorted(tags))[:9]
['acosta', 'amigo', 'amores', 'and', 'animal', 'animales', 'baby', 'bau',
    'beautiful']
>>> list(sorted(annotators))[:5]
['3374092858', '3374094788', '3374097114', '3374098976', '3374107231']
>>> G7['3374092858']
{'cat': {'weight': 0.1629}, 'doll': {'weight': 0.1629}, 'white': {'weight':
    0.1629}}
>>> G7['3374098976']
{'cat': {'weight': 0.1248}}
```

- *Step 5*: For each image graph apply the HITS algorithm to rank the tags according to the computed authority value (a).

```
>>> import operator
>>> G7 = nx.DiGraph(G7)
>>> [h7, a7] = nx.hits(G7)
>>> sorted_a7 = sorted(a7.items(), key=operator.itemgetter(1), reverse=True)
>>> sorted_a7[:4]
[('cat', 0.2030), ('doll', 0.1318), ('white', 0.1268), ('cute', 0.1171)]
```

- *Step 6*: Compute various recall and precision values for different authority score thresholds θ and plot the result.

```
>>> Thresholds=[0.25, 0.21, 0.17, 0.15, 0.13, 0.11, 0.09, 0.07, 0.05, 0.03, 0.01]
>>> p = []; r = []
>>> for t in Thresholds:
... [R, P] = computeROC('img', 'data/gold.json', 50, t)
 ... p += [P]; r += [R]
 ...
>>> from sklearn import metrics
>>> metrics.auc(r,p)
>>> import matplotlib.pyplot as plt
>>> plt.plot(p,r); plt.axis([0.2, 0.95, 0.2, 0.95])
>>> plt.title('ROC curve (Recall vs Precision) with AUC = 0.692')
>>> plt.xlabel('Precision'); plt.ylabel('Recall')
>>> plt.grid(True); plt.show()
```

The proprietary Python functions that were developed and used in the experimentation (file *csv2imagGraphs.py*) are listed below:

```
import networkx as nx
import numpy as np
from wordsegment import load, segment
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import TweetTokenizer
load()

def FullGraph(data,M,no_split,file_out):
   G = nx.DiGraph()
   for j in np.arange(M):
      img = str(j+1)+'_choose'
      img1= str(j+1)+'_own'
      users = data.keys()
      for u in users:
         key_list = list(set(tknzr.tokenize(data[u][img])+
             tknzr.tokenize(data[u][img1])))
         keys = []
         for key in key_list:
            if key in no_split:
               keys +=[key]
            else:
               keyX = segment(key)
               keyX = [lemmatizer.lemmatize(w) for w in keyX if len(w)>2]
               keys +=keyX
         keys = sorted(list(set(keys)))
         for key in keys:
            G.add_edge(u, key)
   nx.write_pajek(G,file_out, encoding='UTF-8')
   return G

def ImageGraphs(data,M,h,no_split):
   for j in np.arange(M):
      G1 = nx.DiGraph()
      img = str(j+1)+'_choose'
      img1= str(j+1)+'_own'
      users = data.keys()
      for u in users:
         key_list = list(set(tknzr.tokenize(data[u][img])+
             tknzr.tokenize(data[u][img1])))
         key_list = [w.lower() for w in key_list]
         keys = []
         for key in key_list:
            if key in no_split:
               keys +=[key]
            else:
```

```
          keyX = segment(key)
          keyX = [lemmatizer.lemmatize(w) for w in keyX if len(w)>2]
          keys +=keyX
      keys = sorted(list(set(keys)))
      for key in keys:
        G1.add_edge(u, key, weight=h[u]*100)
    filename = 'img'+str(j+1)+'.net'
    nx.write_pajek(G1,filename, encoding='UTF-8')

def computeROC(filestart, goldfile, N, thresh_level):
  with open(goldfile, 'r') as fp:
    Gold = json.load(fp)
  retrieved = []; matched = []; gold = []
  tp = []; fp = []; fn = []
  for i in np.arange(N):
    filename = filestart+str(i+1)+'.net'
    gold_current = Gold[filestart+str(i+1)]
    G1 = nx.read_pajek(filename, encoding='UTF-8')
    G1 = nx.DiGraph(G1)
    [h, a] = nx.hits(G1)
    keys = [key for key in a.keys() if a[key]>thresh_level]
    tp +=[key for key in keys if key in gold_current]
    fp +=[key for key in keys if key not in gold_current]
    fn +=[key for key in gold_current if key not in keys]
    gold += gold_current
    retrieved += keys
  R = len(tp)/len(gold)
  P = len(tp)/len(retrieved)
  return R, P
```

5.5 Topic models

In this section we refer to the data collection, presentation and analysis of results and conclusions drawn regarding the application of topic modelling (see Section 4.5), for Instagram hashtags filtering and image retrieval. Further details can be found at [20–22].

5.5.1 Data collection and topic coherence evaluation

For the needs of this study and for the evaluation of the proposed methodology we constructed a dataset composed of 1000 Instagram images (see Table 5.17) along with their hashtags by querying with 20 different subjects / hashtags (i.e., #airplane, #ring, etc.). From the retrieved images of each query we manually selected the 50 most visually relevant ones. Those images were uploaded to Instagram by 970 different Instagram users. Owners' hashtags surrounding these images were automatically crawled using the Beautiful Soup library of Python[22]. The crawled hashtags were stored in 20 different files one for

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

each subject. Each file contained 50 rows, each row representing the hashtags of an Instagram image. A total of 17240 hashtags were collected across all subjects.

The Latent Dirichlet Allocation (LDA) method was applied to the hashtags of each subject in an effort to create topic models for each one of the subjects. Since the LDA algorithm requires as input the number of intended topics we had first to identify the optimal number of topics for each one of the subjects. The procedure we followed for that purpose is described in Section 5.5.3. In Table 5.17 we show the optimal number of topics for 10 subjects along with the number of hashtags per subject and the corresponding ACV value (see eq. 5.2).

Table 5.17: Aggregate coherence value for each subject (shown 10 out of 20)

A/A	Topic	# Hashtags	optimal k	ACV
1	HORSE	867	26	0.72
2	ICECREAM	753	20	0.70
3	BIKE	751	20	0.73
4	LAPTOP	913	32	0.73
5	TULIP	683	14	0.72
6	AIRPLANE	1022	26	0.69
7	CAMEL	741	20	0.71
8	RING	1034	32	0.69
9	CROISSANT	752	20	0.71
10	BAG	921	20	0.69

5.5.2 Topic coherence and interpretability

The topic coherence of the models created for each one of the subjects was assessed with the aid of Aggregate Coherence Value (ACV) as shown in Table 5.17. The evaluation was performed by first selecting the optimal number of topics for each subject as explained in the next section. We observe that the aggregate topic coherence for all subjects is quite stable across all subjects ranging from 0.69 to 0.73. This stability is an interesting result on its own regarding the robustness of the proposed method.

Interpretability of topic modelling is usually done through visualization and human assessment. We use the pyLDAvis tool[23] to visualize the fit of our LDA model across the various topics and their top (most frequent) words. pyLDAvis was designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization. Figures 5.22 and 5.23, visualize the topics distribution of the created topic models corresponding to the subjects *#airplane* and *#ring* respectively. It is clear in both cases that in the topic with the highest coherence all the words are tightly associated with the relevant subject.

Interpretability was also qualitatively evaluated by humans through a simple experiment. Two ordinary Instagram users were, independently, shown the topic with the highest coherence through the pyLDAvis tool and were asked to guess the topic title (subject). In all cases they could guess the right subject with less than three attempts. We should mention here, however, that the subjects we used in this experiment correspond to clear and tangible concepts, such as airplane, laptop, dog, etc. Whether humans could

[23] https://pypi.org/project/pyLDAvis/

interpret topics related to abstract terms such as freedom, liberty and democracy is questionable though (see [102] for a related discussion).

5.5.3 Optimal number of topics

When determining how many topics to use, it is important to consider both qualitative and quantitative factors. Qualitatively, you should have domain knowledge of the data you're analyzing and be able to gauge a general ballpark of clusters your data will separate into. There should be enough topics to be able to distinguish between overarching themes in the text but not so many topics that they lose their interpretability. In the case of evaluating Instagram hashtags related to a particular subject, from a qualitative perspective, 10 topics seemed like a reasonable number to start with.

Quantitative evaluation is usually done in an empirical manner as Prabhakaran [277] suggests: In order to find the optimal number k_{opt} of topics in a corpus build many LDA models with different values of k (number of topics) and pick the one that gives the highest Aggregate Coherence Value (ACV). ACV is computed with the aid of eq. 5.2

$$ACV[k] = \frac{1}{k}\sum_{t=1}^{k} C[\mathcal{T}_t] \tag{5.2}$$

$$k_{opt} = argmax\{ACV[k]\} \tag{5.3}$$

Plotting the aggregate coherence value versus k as in Figure 5.21, is, in most cases, useful. Choosing a k that marks the end of a rapid growth (k = 6 in the diagram) of topic coherence usually offers meaningful and interpretable topics. Picking an even higher value can sometimes provide more granular sub-topics. If you see the same keywords being repeated in multiple topics, it's probably a sign that the k is too large. If the aggregate coherence score seems to keep increasing, it may make better sense to pick a k that gave the highest CV before flattening out (k=20 in the diagram).

Table 5.18 shows our experimentation regarding the selection of the optimal number of topics for the subject #AIRPLANE.

Table 5.18: Optimal number of topics for the subject #AIRPLANE

# Topics	2	8	14	20	**26**	32	38
ACV	0.462	0.6101	0.6712	0.6857	**0.6951**	0.6753**26**	0.6366

5.5.4 Human interpretation of topic models through Word Clouds

As mentioned in Section 4.5.5 the purpose of this study was to examine the interpretability of topic models created from the Instagram hashtags as described in the previous sections. We have decided to investigate human interpretation of topic models on the basis of a generic crowd and students of the Cyprus University of Technology.

The topics were shown as word clouds with the queried hashtags (subjects) hidden and the crowd and students were asked to guess the hidden hashtag providing their best four guesses. The aim was to examine

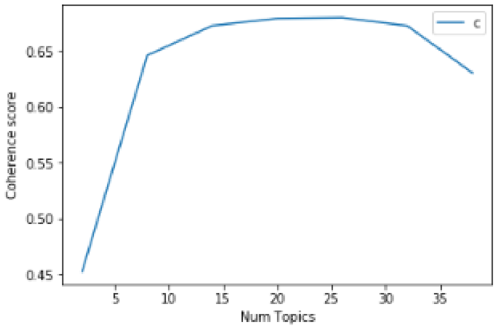


Figure 5.21: Diagrammatic presentation of the results of Table 5.18

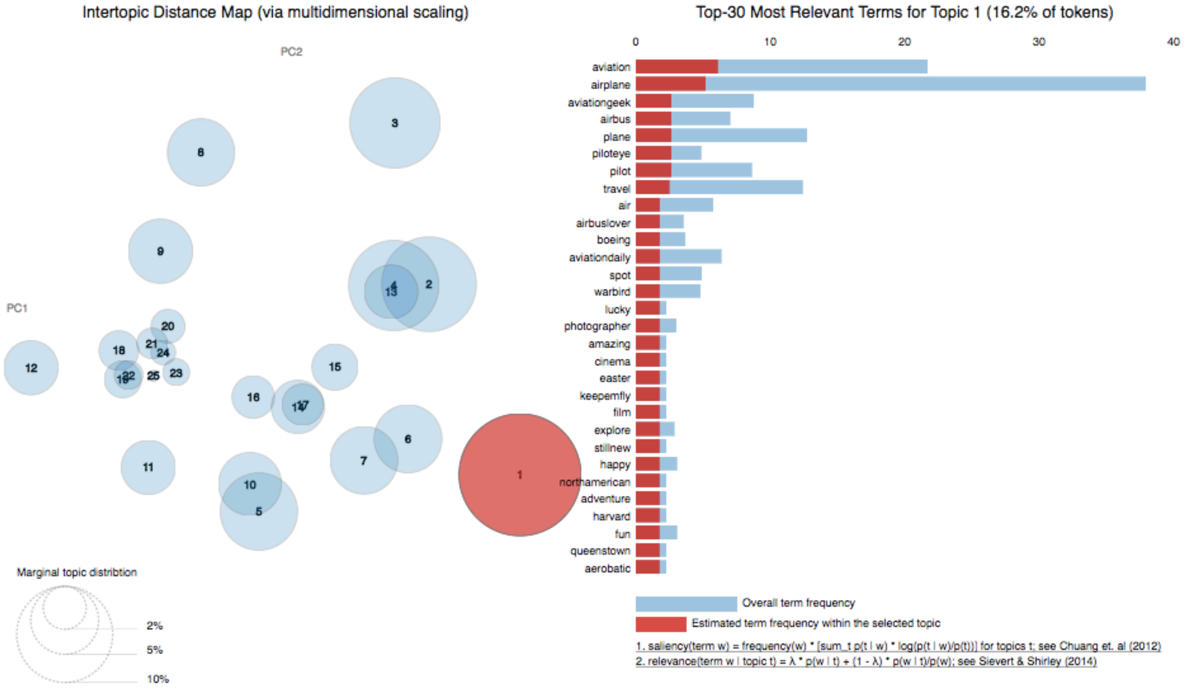


Figure 5.22: #AIRPLANE Topic visualization

the accuracy of topic models interpretation, as well as a comparison between the crowd and the students. Regarding the latter, we aimed also to investigate if there is significant correlation on the way the crowd and the students interpret the word clouds of Instagram hashtags. If crowd and students choice coincide with the subject of the word cloud, we have a good indication that the word cloud words, indeed, related with the subject. We considered that through this meta-analysis we could gain useful insights on whether we can use words mined form Instagram hashtags for AIA purposes.

A dataset of 520 Instagram posts (photos along with their associated hashtags) was created by querying with 26 different hashtags / subjects (see Table 5.19). For each subject we collected 10 visually relevant to the subject Instagram posts (images and associated hashtags) and 10 visually irrelevant ones. This led to a total of 520 (260 relevant and 260 non-relevant) images and 8199 hashtags (2883 for relevant images and 5316 non-relevant images).

All collected hashtags were undergone preprocessing so as to derive meaningful tokens (words in En-

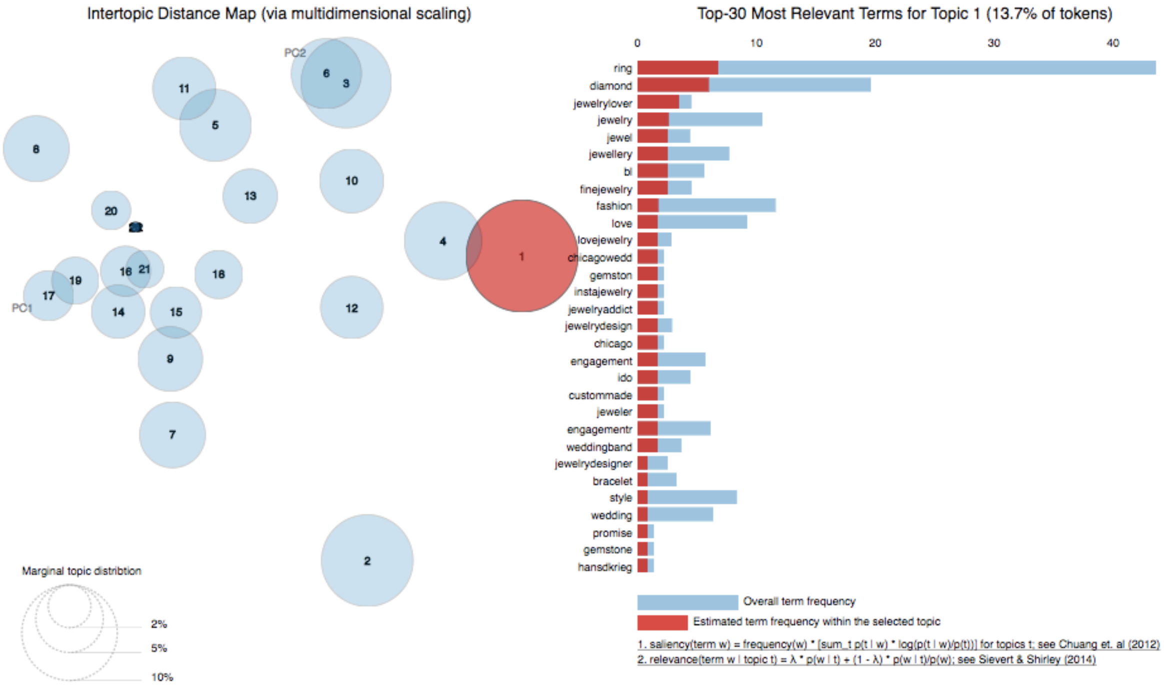
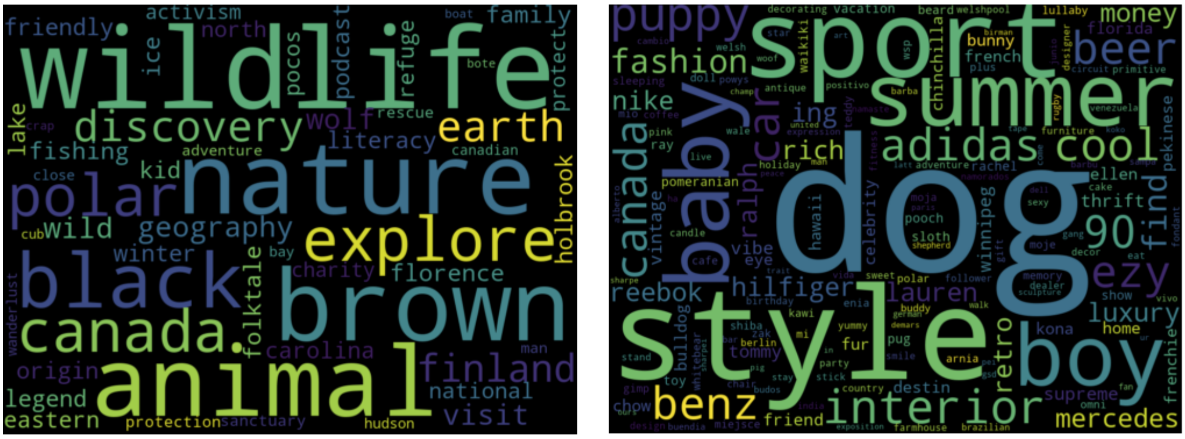


Figure 5.23: #RING Topic visualization

glish). Instragram hashtags, are unstructured and ungrammatical, and it is important to use linguistic processing to (a) remove stophashtags (see Section 5.3), that is hashtags that are used to fool the search results of the Instagram platform, (b) split a composite hashtag to its consisting words (e.g. the hashtag '#spoilyourselfthisseason' should be split into four words: 'spoil', 'yourself', 'this', 'season'), (c) remove stopwords that are produced in the previous stage (e.g. the word 'this' in the previous example), (d) perform spelling checks to account for (usually intentionally) misspelled hashtags (e.g. '#headaband', '#headabandss' should be changed to '#headband'), and (e) perform lemmatization to merge words that share the same or similar meaning.



(a) Bear relevant word cloud

(b) Bear irrelevant word cloud

Figure 5.24: Relevant & irrelevant word clouds for the subject (queried hashtag) bear

5.5.5 Interpretation of word clouds

Crowd-based interpretation of word clouds was conducted with the aid of the *Appen*[24] crowdsourcing platform (see Figure 5.25) and student-based interpretation was performed with the aid of the learning platform *Moodle*[25] (see Figure 5.26). We chose the interpretation from cloud because we wanted to take advantage of the principles of collective intelligence.

The word clouds were presented to crowd participants which were asked to select one to four of the subjects that best match the shown word cloud according to their interpretation. The participants were clearly informed that the token corresponding to the correct subject was not shown in the cloud. The same questions were presented to the students and which also had to choose between one to four subjects that best match the word cloud they saw. Both students and the crowd were informed the correct subject was not included in the word cloud.

Every word cloud was judged by at least 30 annotators (contributors in Appen's terminology) while eight word clouds were also used as 'gold questions' for quality assurance, i.e., identification of dishonest annotators and task difficulty assessment. The correct answer(s) for the gold clouds were provided to the crowdsourcing platform and all participants had to judge those clouds. However, gold clouds were presented to the contributors in random order and they could not know which of the clouds were the gold ones. A total of 165 contributors from more than 25 different countries participated in the experiment. The cost per judgement was set to \$0.01 and the task was completed in less than six hours. A total of 25 students annotations were also collected.

The crowd and students interpretations of each one of the word clouds were also transformed as word clouds, i.e., meta word clouds, for illustration purposes. The importance of each token in a meta word cloud was based on the frequency of its selection by the contributors and students. Meta word clouds are presented in Figures 5.28, 5.29, 5.31, 5.32, 5.34, 5.35. The tokens in a meta word cloud can be seen as the topic model suggested by the crowd and students for the Instagram photos grouped under the corresponding subject. For instance we could say that the topic model for the images grouped under the subject 'microphone' includes also the words 'guitar' and 'piano' and thus all three words can be used for tagging the corresponding photos even for creating training datasets for AIA purposes [195].

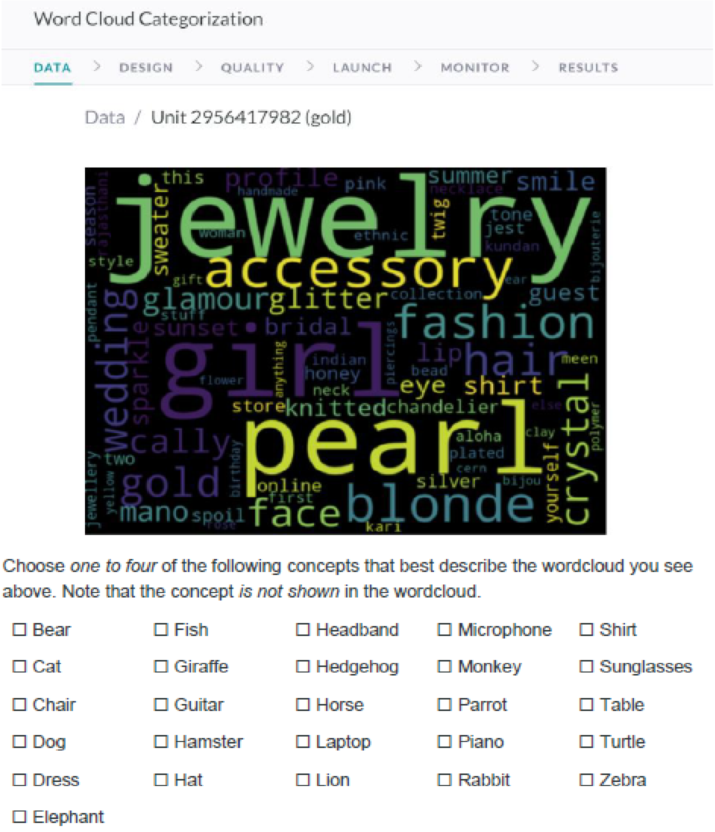
Not all word clouds present the same difficulty of interpretation. Thus, in order to quantitatively estimate that difficulty per subject we used the typical accuracy metric, that is the percentage of correct subject identifications by the crowd and the students. By correct identification we mean that a contributor or a student had selected the right subject within her/his one to four choices. We see for instance in Table 5.19 that the accuracy for crowd of the guitar word cloud is 93%. This means that 93% of the contributors included the word 'guitar' in their interpretation for that word cloud, regardless the number (1 to 4) of contributor choices.

5.5.6 Discussion regarding the interpretatblity of topics

The accuracy of interpretation for all word clouds is presented in Table 5.19 while summary statistics are presented in Table 5.20. In order to better facilitate the discussion that follows the subjects (query

[24] https://appen.com/

[25] https://elearning.cut.ac.cy/



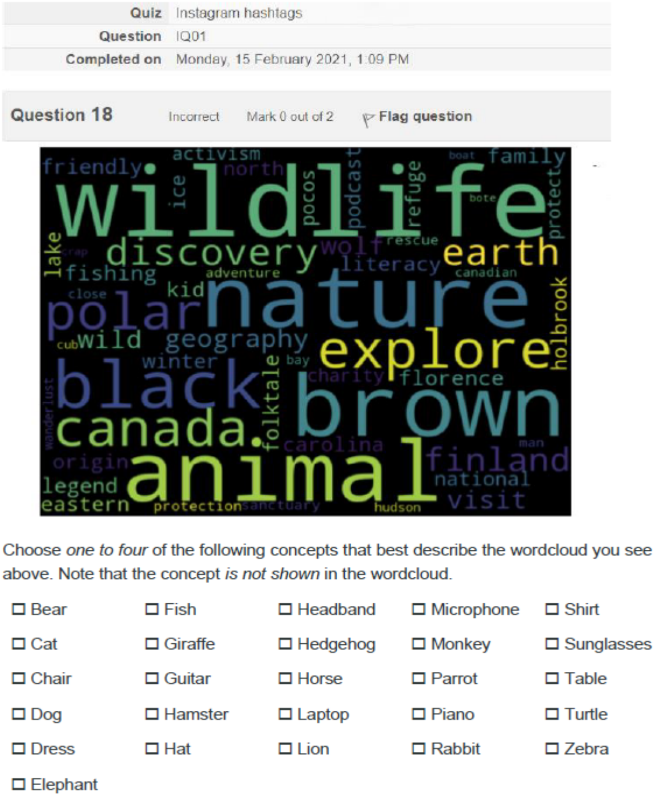
Word Cloud Categorization

Data / Unit 2956417982 (gold)

Choose *one to four* of the following concepts that best describe the wordcloud you see above. Note that the concept *is not shown* in the wordcloud.

☐ Bear
☐ Cat
☐ Chair
☐ Dog
☐ Dress
☐ Elephant
☐ Fish
☐ Giraffe
☐ Guitar
☐ Hamster
☐ Hat
☐ Headband
☐ Hedgehog
☐ Horse
☐ Laptop
☐ Lion
☐ Microphone
☐ Monkey
☐ Parrot
☐ Piano
☐ Rabbit
☐ Shirt
☐ Sunglasses
☐ Table
☐ Turtle
☐ Zebra

Figure 5.25: Word cloud interpretation in the Appen crowdsourcing platform



Quiz	Instagram hashtags
Question	IQ01
Completed on	Monday, 15 February 2021, 1:09 PM

Question 18 Incorrect Mark 0 out of 2 Flag question

Choose *one to four* of the following concepts that best describe the wordcloud you see above. Note that the concept *is not shown* in the wordcloud.

☐ Bear
☐ Cat
☐ Chair
☐ Dog
☐ Dress
☐ Elephant
☐ Fish
☐ Giraffe
☐ Guitar
☐ Hamster
☐ Hat
☐ Headband
☐ Hedgehog
☐ Horse
☐ Laptop
☐ Lion
☐ Microphone
☐ Monkey
☐ Parrot
☐ Piano
☐ Rabbit
☐ Shirt
☐ Sunglasses
☐ Table
☐ Turtle
☐ Zebra

Figure 5.26: Word cloud interpretation in Moodle

Table 5.19: Topic identification accuracy for word clouds created using visually relevant (Relev.) and irrelevant (Irre.) Instagram photos

	Relevant		Irrelevant	
Subject	Crowd (%)	Student (%)	Crowd (%)	Student (%)
Guitar	93	88	87	84
Piano	70	80	47	72
Microphone	57	92	67	80
Bear	43	76	0	0
Elephant	37	48	0	0
Giraffe	63	72	3	16
Lion	60	76	67	44
Monkey	33	36	0	0
Zebra	57	68	3	0
Dress	80	84	60	76
Hat	7	40	3	52
Headband	30	24	17	80
Shirt	33	48	53	56
Sunglasses	67	68	13	36
Chair	43	60	47	80
Laptop	100	96	80	92
Table	73	84	77	84
Cat	90	92	17	60
Dog	87	92	0	0
Fish	100	92	93	84
Hamster	3	40	7	36
Parrot	87	88	90	84
Rabbit	77	72	7	0
Turtle	20	60	20	52
Hedgehog	0	12	0	0
Horse	87	88	7	4

Table 5.20: Summary statistics for the accuracy of identification

Subject	Mean (%)	St. Dev.(%)	Min (%)	Max (%)
Student Relevant	68	23	12	96
Crowd Relevant	58	30	0	100
Student Irrelevant	45	35	0	92
Crowd Irrelevant	33	34	0	93

Table 5.21: Independent samples t-test, N=26 subjects in all cases

Group	Mean (%)	St. Dev. (%)	Stan. Err. (%)	t	p
Relevant (Students)	68	23	5	25	.003
Irrelevant (Students)	45	35	7	25	
Relevant (Crowd)	58	30	6	25	.001
Irrelevant (Crowd)	33	34	6	25	

hashtags) were divided into six categories: (a) **Music**: Guitar, Piano, Microphone (b) **Wild animals**: Bear, Elephant, Giraffe, Lion, Monkey, Zebra (c) **Fashion**: Dress, Hat, Headband, Shirt, Sunglasses (d) **Office**: Chair, Laptop, Table, (e) **Pets**: Cat, Dog, Fish, Hamster, Parrot, Rabbit, Turtle (f) **Miscellaneous**: Hedgehog, Horse.

In order to answer the main research questions of our study formulate three null hypotheses as follows:

H_{01}*: There is no significant difference of word cloud interpretation of hashtags sets mined from relevant and irrelevant images by the trained students.*

H_{02}*: There is no significant difference of word cloud interpretation of hashtags sets, mined from relevant and irrelevant images, by the generic crowd.*

H_{03}*: There is no significant correlation on the way the generic crowd and trained students interpret the word clouds mined from Instagram hashtags.*

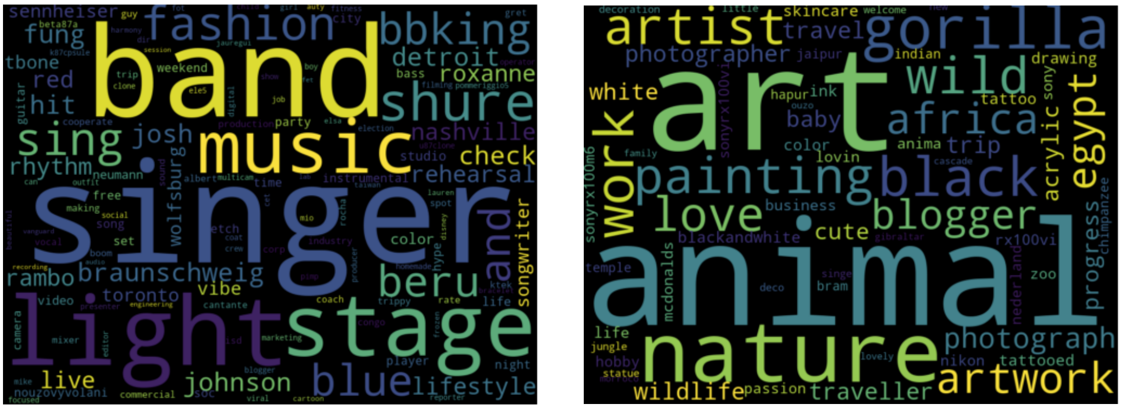
In Table 5.21 we see the paired-sampled t-test which was conducted, with the aid of SPSS, to compare the interpretation in relevant and irrelevant word clouds conditions in both the crowd and students. There is a significant difference in the scores for relevant (Mean Crowd=68%, Mean Student=58%) and irrelevant (Mean Crowd=33%, Mean Student=45%). Thus the null hypotheses H_{01} and H_{02} are rejected at a significance level $a = .003$ for students and $a = .001$ for the crowd.

Regarding the third null hypothesis, for a significant level $a = 0.01$ the critical value for the correlation coefficient (two tail test, $df = 50$) is r_c = 0.354. By computing the correlation coefficient (Pearson rho) of the mean accuracy values per subject of the crowd and the students we find $r = 0.861$. Thus, $r > r_c$ and the null hypothesis(H_{03}) is rejected at a significance level $a = 0.01$, denoting that the way word clouds are interpreted by the trained students and the crowd is highly correlated.

We see in Table 5.19 that the interpretation accuracy varies within and across categories. As we explain later through specific examples, there are three main parameters which affect the difficulty of interpretation. The first one is the conceptual context for a specific term. It is very easy, for instance, to define a clear conceptual context for the term fish but very difficult to define clear conceptual contexts for terms such as *hat* and *hedgehog*. This difficulty is, obviously, reflected in the use of hashtags that accompany photos presenting those terms. As a result the corresponding word clouds do not provide the textual context and hints that allow their correct interpretation. Thus, textual context and key tokens in the word clouds is the second parameter affecting the difficulty of interpretation. Concepts such as *dog*, *cat* and *horse* are far more familiar to everyday people than concepts such as *hedgehog* and *hamster*.

In the following we present and discuss some representative / interesting examples for each one of the six categories mentioned above.

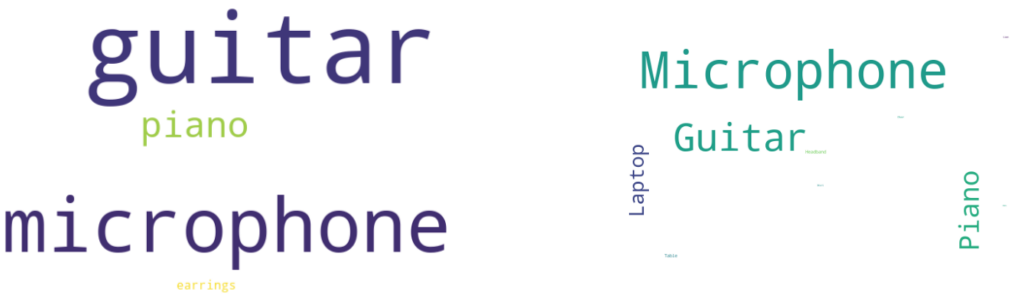
The word clouds in the *Music* category have very high scores of interpretation accuracy. Music related terms share a strong conceptual context which results in clear textual contexts in the Instagram hashtags. In Figure 5.27a we see the word cloud for the subject ‘microphone’ and in Figure 5.28 we see the interpretation word clouds from the crowd and students. Tokens like *band*, *singer*, *music*, *singer* and *stage* create a strong and clear textual content. Thus, the annotators, 57% for crowd and 92% for students, correctly chose ‘microphone’ to interpret the word cloud. Moreover, the ‘microphone’ word cloud tokens had as a result the crowd and the students to choose also *guitar*, *piano* as we see in interpretation word clouds (see Figure 5.28).



(a) Word cloud for the 'microphone' subject

(b) Word cloud for the 'monkey' subject

Figure 5.27: Word clouds for the 'microphone' and 'monkey' subject



(a) Crowd based interpretation

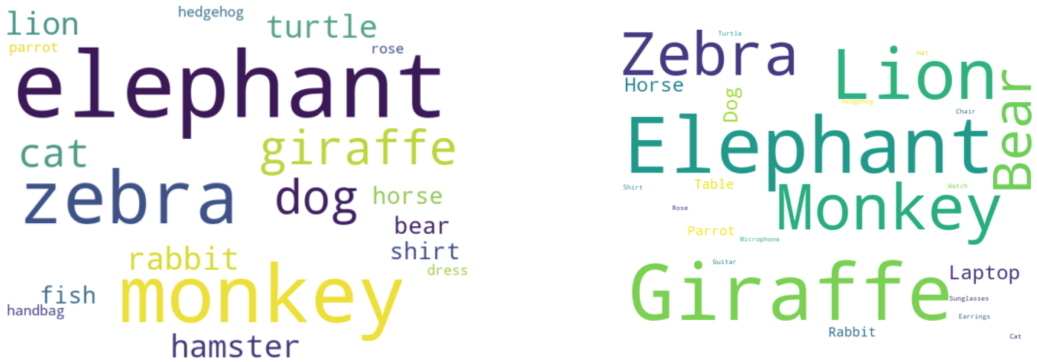
(b) Students' interpretation

Figure 5.28: Crowd and students based interpretation of the 'microphone' word cloud

The 'monkey' word cloud (see Figure 5.27b) was in fact a confusing one. The most prominent tokens were *art*, *animal* and *nature* while some other terms such as *artist*, *artwork*, and *work* could also confuse the crowd and the students. As a result the accuracy for that category is 33% for the crowd and 36% for the students. We see in the meta word clouds for students and crowd (Figure 5.29), however, the key tokens *animal* and *nature* combined with the term *gorilla* in the upper right corner of the word cloud led the contributors and students to make selections from the *wild animal* category including the correct subject.

The case of subject 'hat' (see word cloud in Figure 5.30a) shows a situation where there are many different conceptual contexts. As a result, the hashtags appeared in different Instagram photos differ significantly and the resulting word cloud is confusing. We see that the most prominent tokens in the cloud are *blogger*, *style*, *sun*, and *beach* (obviously these are concepts shown in some of the Instagram photos grouped under the subject 'hat'). There is no doubt that the subject 'hat' fits well with those terms. However, the same terms fit well or even better to other subjects such as 'sunglasses' and dress that had as result the accuracy was not high for students and crowd (7% for crowd and 40% for students - see Figure 5.31).

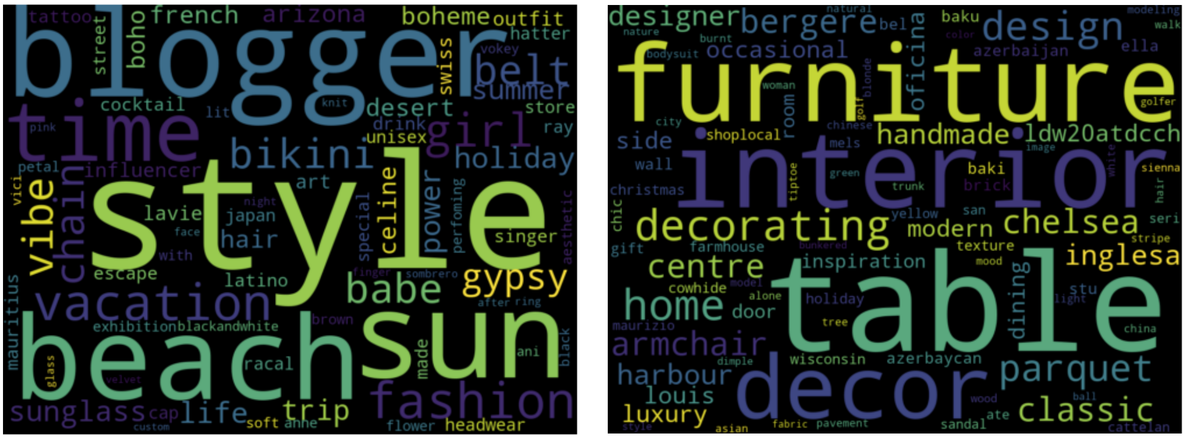
The 'chair' word cloud (see Figure 5.30b) contains words like *furniture*, *table*, *interior* which are all related to the 'table' and 'chair' terms. As, a result the crowd and the students interpret that word cloud



(a) Crowd based interpretation

(b) Students' interpretation

Figure 5.29: Crowd and student based interpretation of the 'monkey' word cloud



(a) Word cloud for the 'hat' subject

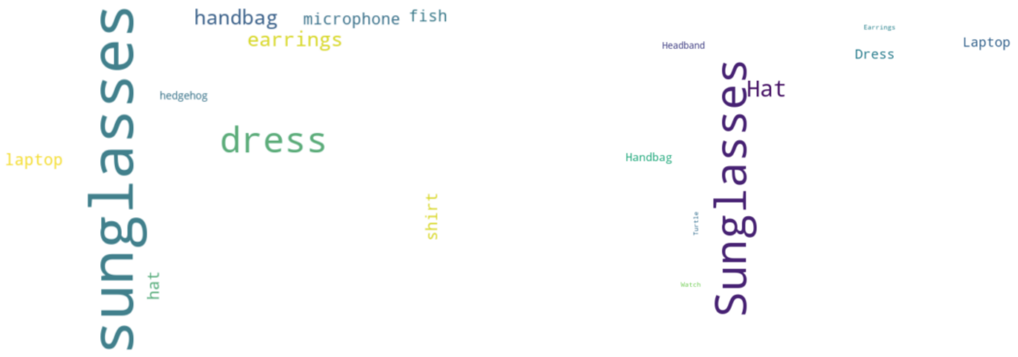
(b) Word cloud for the 'chair' subject

Figure 5.30: Word clouds the subjects 'hat' and 'chair'

with the terms tags 'table 'and 'chair' (43% for crowd and 60% for students - see Figure 5.32).

The case of 'hedgehog' is a classic example showing that the familiarity with a concept affects the difficulty in interpretation of the word cloud derived from Instagram hashtags. While in the word cloud (see Figure 5.33a) the words *pygmy*, *pet* and *animal* are by far the most important ones, none of the participants selected the right subject. It appears that both the crowd and the students were non-familiar with the word *pygmy*. The African pygmy hedgehog is the species often used as pet. By examining the meta word clouds (see Figure 5.34) we see that the Appen contributors and students mixed up concepts related to pets with concepts related to wild animals.

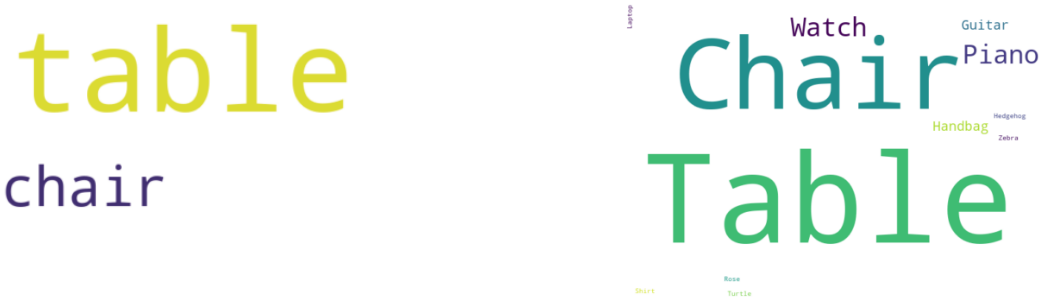
The case of 'hamster' (see Figure 5.33b) represents a situation with different conceptual contexts. The hashtags that appear in that world cloud are indeed related to 'hamster': *dwarf* and *syrian* are hamster species while *life* is a generic word. The crowd and the students were not familiar with these hamster species and as a result the accuracy was quite low: 3% for crowd and 40% for students. By examining the meta word clouds (see Figure 5.35) the word *animal* in the word cloud led the crowd to choose mostly wild animals and the students to choose pets.



(a) Crowd based interpretation

(b) Students' interpretation

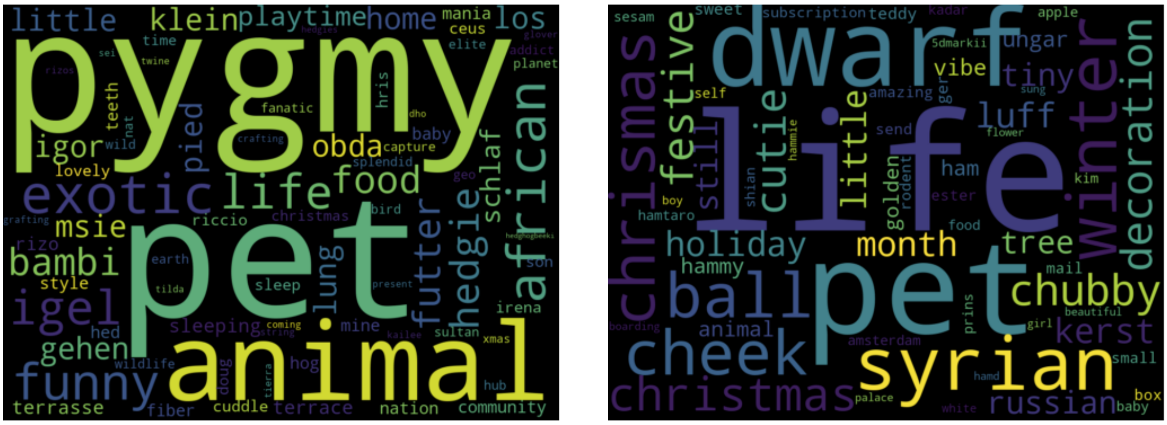
Figure 5.31: Crowd and students based interpretation of the 'hat' word cloud



(a) Crowd based interpretation

(b) Students' interpretation

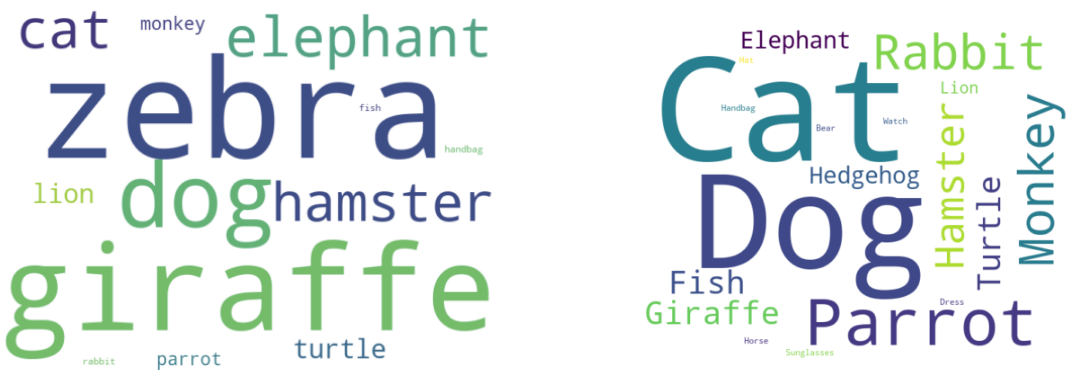
Figure 5.32: Crowd and students based interpretation of the 'chair' word cloud



(a) Word cloud for the 'hedgehog' subject

(b) Word cloud for the 'hamster' subject

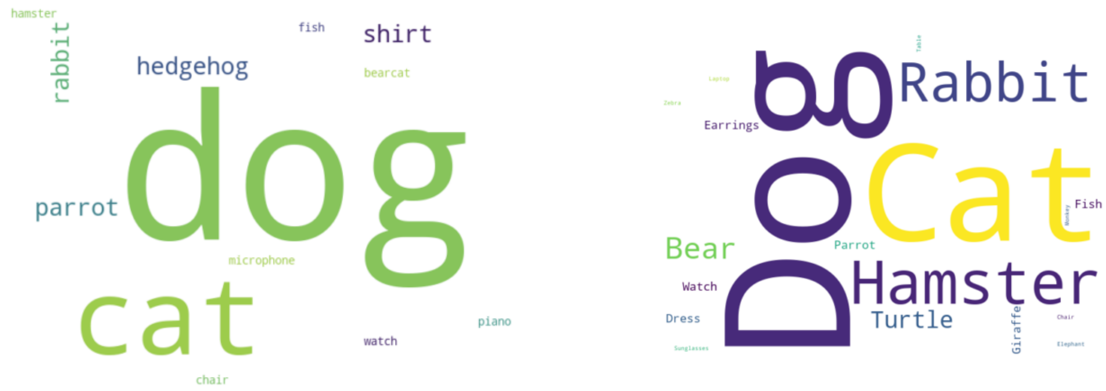
Figure 5.33: Word clouds for the subjects 'hedgehog' and 'hamster'



(a) Crowd based interpretation

(b) Students' interpretation

Figure 5.34: Crowd and students based interpretation of the 'hedgehog' word cloud



(a) Crowd based interpretation

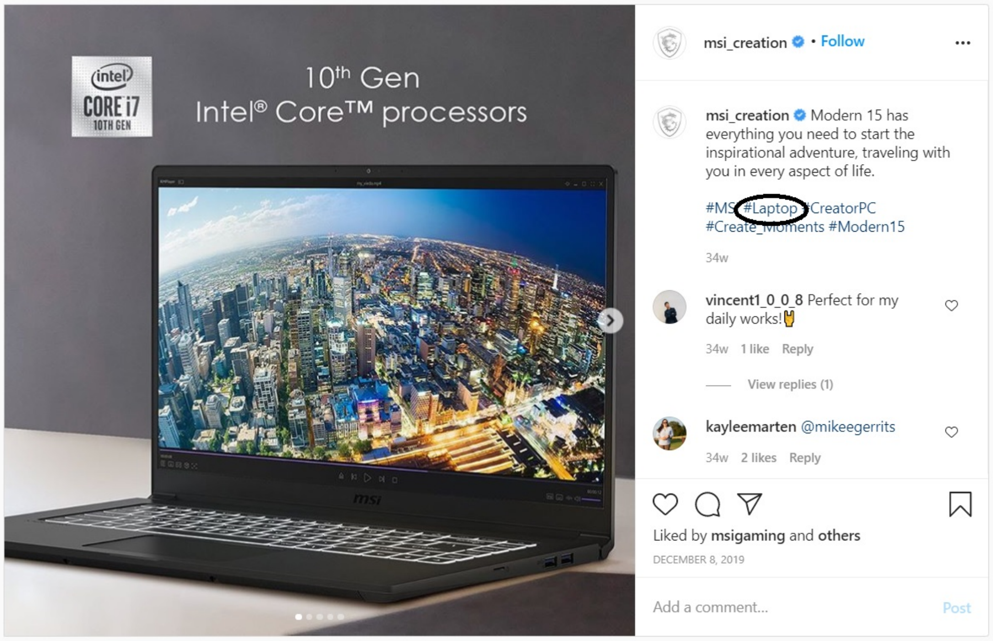
(b) Students' interpretation

Figure 5.35: Crowd and students based interpretation of the 'hamster' word cloud

5.6 Correlation between color histograms and Instagram hashtag sets

In this study we examine whether there is a correlation between the low level visual characteristics of the Instagram images and the (filtered) hashtags appended to them. The methodology we follow was described in Section 4.6 and employs the representation of images via color histograms while the hashtag sets are represented using word embeddings.

The dataset used in the current studies were collected by using 26 independent query hashtags (i.e., #dog, #elephant, etc.), called, as already explained in previous sections, subjects. For each query hashtag we collected 10 relevant and 10 irrelevant image posts (images and associated hashtags) leading to a total of 520 (260 relevant and 260 non-relevant) images and 8199 hashtags (2883 for relevant images and 5316 non-relevant images). An example of a relevant Instagram post for the hashtag subject #laptop is shown in Fig. 5.36 while Fig. 5.37 shows an irrelevant one.



msi_creation Modern 15 has everything you need to start the inspirational adventure, traveling with you in every aspect of life.

#MSI #Laptop #CreatorPC #Create_Moments #Modern15

vincent1_0_0_8 Perfect for my daily works!🤘

kayleemarten @mikeegerrits

Liked by msigaming and others

DECEMBER 8, 2019

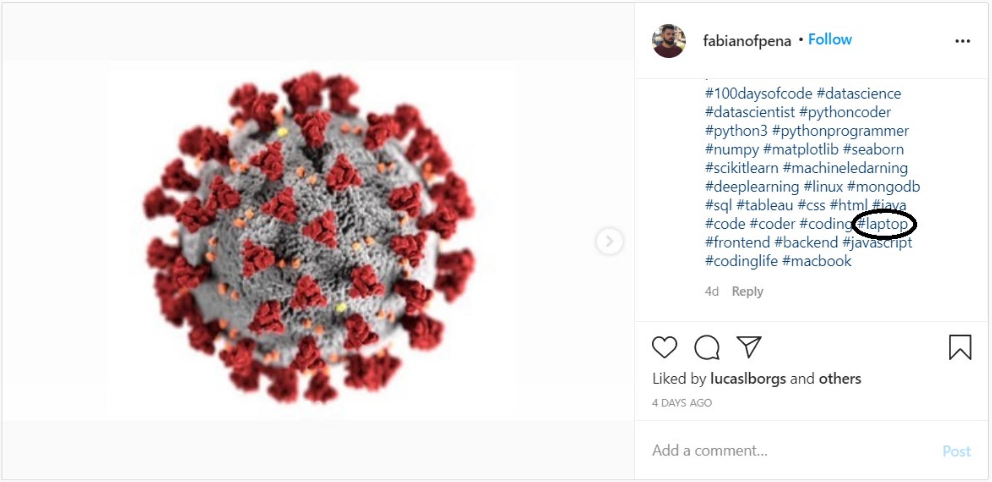
Figure 5.36: Example of a relevant Instagram post for hashtag #laptop

Table 5.22: Average Bhattacharyya similarity scores (relevant posts)

	bear	cat	chair	dog	dress	elephant	fish
Mean	0.722	0.734	0.735	0.749	0.729	0.765	0.668
St. Dev.	0.062	0.046	0.062	0.074	0.057	0.042	0.097
Min	0.602	0.654	0.565	0.592	0.613	0.685	0.529
Max	0.840	0.885	0.840	0.885	0.893	0.855	0.885
	giraffe	guitar	hamster	hat	headband	hedgehog	horse
Mean	0.729	0.674	0.744	0.683	0.701	0.718	0.708
St. Dev.	0.077	0.054	0.054	0.076	0.072	0.066	0.061
Min	0.602	0.581	0.633	0.533	0.575	0.575	0.592
Max	0.840	0.840	0.840	0.870	0.826	0.840	0.833

The data (images and hashtags) were analysed with the aid of Python[26]. In order to compute the color histogram of an image and the Bhattacharyya distance between two images represented via their color

[26] https://www.python.org/



#100daysofcode #datascience #datascientist #pythoncoder #python3 #pythonprogrammer #numpy #matplotlib #seaborn #scikitlearn #machineledarning #deeplearning #linux #mongodb #sql #tableau #css #html #java #code #coder #coding #laptop #frontend #backend #javascript #codinglife #macbook

Liked by lucaslborgs and others

4 DAYS AGO

Add a comment...

Figure 5.37: Example of a non relevant Instagram post for hashtag #laptop

Table 5.23: Average Bhattacharyya similarity scores (relevant posts)

	laptop	lion	mic	monkey	parrot	piano
Mean	0.682	0.767	0.690	0.671	0.693	0.680
St. Dev.	0.043	0.066	0.061	0.096	0.065	0.051
Min	0.602	0.637	0.565	0.521	0.556	0.592
Max	0.800	0.893	0.813	0.847	0.800	0.806
	rabbit	shirt	sunglasses	table	turtle	zebra
Mean	0.719	0.699	0.668	0.719	0.706	0.717
St. Dev.	0.055	0.086	0.055	0.044	0.067	0.070
Min	0.621	0.543	0.565	0.633	0.578	0.606
Max	0.893	0.870	0.813	0.800	0.826	0.906

Table 5.24: Average Bhattacharyya similarity scores (irrelevant posts)

	bear	cat	chair	dog	dress	elephant	fish
Mean	0.728	0.690	0.687	0.671	0.631	0.661	0.665
St. Dev.	0.074	0.061	0.061	0.067	0.065	0.062	0.092
Min	0.581	0.587	0.580	0.548	0.519	0.555	0.510
Max	0.856	0.821	0.785	0.810	0.754	0.841	0.814
	giraffe	guitar	hamster	hat	headband	hedgehog	horse
Mean	0.717	0.626	0.694	0.664	0.665	0.681	0.706
St. Dev.	0.055	0.062	0.079	0.078	0.080	0.071	0.090
Min	0.620	0.523	0.530	0.525	0.518	0.537	0.528
Max	0.846	0.751	0.841	0.798	0.815	0.807	0.889

histograms, we made use the *OpenCV*[27] Python library. For the numeric representation of an Instagram hashtag set we used the Gensim library and specifically the *Doc2Vec*[28] model. We have to note here that the hashtag sets, appended to each one of the Instagram photos, were first filtered described the corresponding methods described in the current thesis and then transformed to numeric representations using the *Doc2Vec* model.

In Table 5.30 we see the global (across all subjects) statistics regarding the color histogram similarities for the relevant and irrelevant posts while in Figure 5.38 we see a more detailed comparison per subject.

[27] https://opencv.org/

[28] https://www.tutorialspoint.com/gensim/gensim_doc2vec_model.htm

Table 5.25: Average Bhattacharyya similarity scores (irrelevant posts)

	laptop	lion	mic	monkey	parrot	piano
Mean	0.682	0.767	0.690	0.671	0.693	0.680
St. Dev.	0.061	0.043	0.066	0.061	0.096	0.065
Min	0.602	0.637	0.565	0.521	0.556	0.592
Max	0.800	0.893	0.813	0.847	0.800	0.806
	rabbit	shirt	sunglasses	table	turtle	zebra
Mean	0.719	0.699	0.668	0.719	0.706	0.717
St. Dev.	0.055	0.086	0.055	0.044	0.067	0.070
Min	0.621	0.543	0.565	0.633	0.578	0.606
Max	0.893	0.870	0.813	0.800	0.826	0.906

Table 5.26: Average similarity of hashtag sets (relevant posts)

	bear	cat	chair	dog	dress	elephant	fish
Mean	0.350	0.255	0.263	0.224	0.250	0.247	0.167
St. Dev.	0.084	0.137	0.109	0.097	0.068	0.115	0.075
Min	0.183	0.049	0.103	0.097	0.107	0.100	0.069
Max	0.575	0.781	0.560	0.471	0.404	0.643	0.356
	giraffe	guitar	hamster	hat	headband	hedgehog	horse
Mean	0.197	0.254	0.257	0.294	0.189	0.160	0.220
St. Dev.	0.047	0.083	0.143	0.110	0.043	0.084	0.117
Min	0.126	0.148	0.087	0.126	0.124	0.046	0.059
Max	0.316	0.534	0.920	0.633	0.297	0.457	0.692

Table 5.27: Average similarity of hashtag sets (relevant posts)

	laptop	lion	mic	monkey	parrot	piano
Mean	0.257	0.225	0.220	0.251	0.151	0.217
St. Dev.	0.042	0.050	0.043	0.090	0.069	0.079
Min	0.162	0.137	0.114	0.044	0.043	0.090
Max	0.337	0.363	0.310	0.445	0.306	0.404
	rabbit	shirt	sunglasses	table	turtle	zebra
Mean	0.719	0.699	0.668	0.719	0.706	0.717
St. Dev.	0.055	0.086	0.055	0.044	0.067	0.070
Min	0.621	0.543	0.565	0.633	0.578	0.606
Max	0.893	0.870	0.813	0.800	0.826	0.906

Table 5.28: Average similarity of hashtag sets (irrelevant posts)

	bear	cat	chair	dog	dress	elephant	fish
Mean	0.148	0.134	0.183	0.209	0.240	0.189	0.138
St. Dev.	0.062	0.072	0.094	0.068	0.115	0.078	0.078
Min	0.051	0.030	0.024	0.103	0.064	0.094	0.019
Max	0.322	0.291	0.395	0.330	0.464	0.507	0.332
	giraffe	guitar	hamster	hat	headband	hedgehog	horse
Mean	0.139	0.238	0.114	0.261	0.187	0.158	0.094
St. Dev.	0.062	0.118	0.058	0.060	0.034	0.059	0.056
Min	0.053	0.068	0.035	0.163	0.120	0.052	0.013
Max	0.291	0.519	0.298	0.417	0.249	0.277	0.206

Additional statistics are presented in Tables 5.22, 5.23 & 5.24, 5.25 for the relevant and irrelevant posts respectively. Despite the fluctuations across various subjects, the global mean for color histogram similarity in relevant posts is significantly higher than that of irrelevant posts ($t = 4.35$, $p < 0.01$, $df = 25$,

Table 5.29: Average similarity of hashtag sets (irrelevant posts)

	laptop	lion	mic	monkey	parrot	piano
Mean	0.161	0.210	0.187	0.121	0.102	0.178
St. Dev.	0.073	0.059	0.048	0.077	0.051	0.083
Min	0.028	0.115	0.107	0.017	0.015	0.054
Max	0.206	0.300	0.361	0.303	0.274	0.257
	rabbit	shirt	sunglasses	table	turtle	zebra
Mean	0.181	0.164	0.104	0.178	0.143	0.119
St. Dev.	0.076	0.066	0.038	0.076	0.040	0.052
Min	0.035	0.081	0.043	0.065	0.058	0.020
Max	0.352	0.367	0.178	0.336	0.220	0.241

Table 5.30: Average statistics of Bhattacharyya similarity across all subjects for photos taken from relevant and irrelevant posts

	Mean	St. Dev.	Min	Max
Relevant Images	0.710	0.028	0.668	0.767
Non Relevant Images	0.672	0.034	0.596	0.728

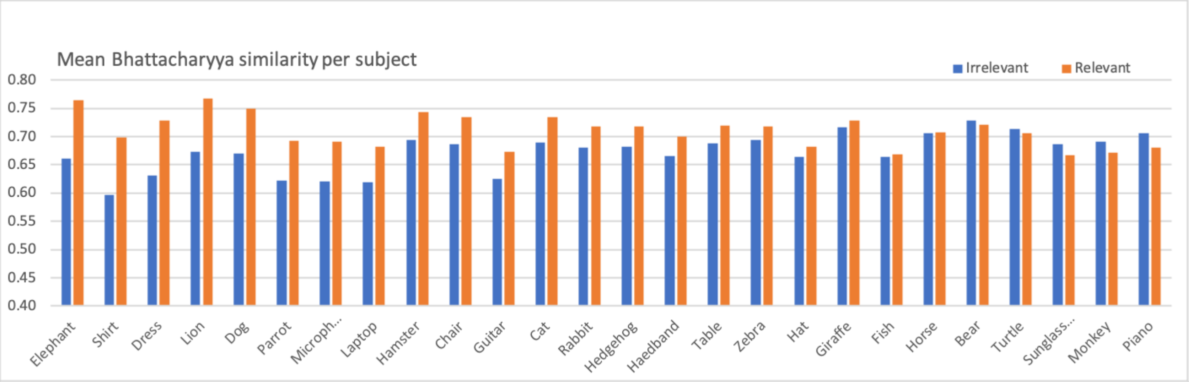


Figure 5.38: Mean Bhattacharyya similarity per subject for relevant & irrelevant posts

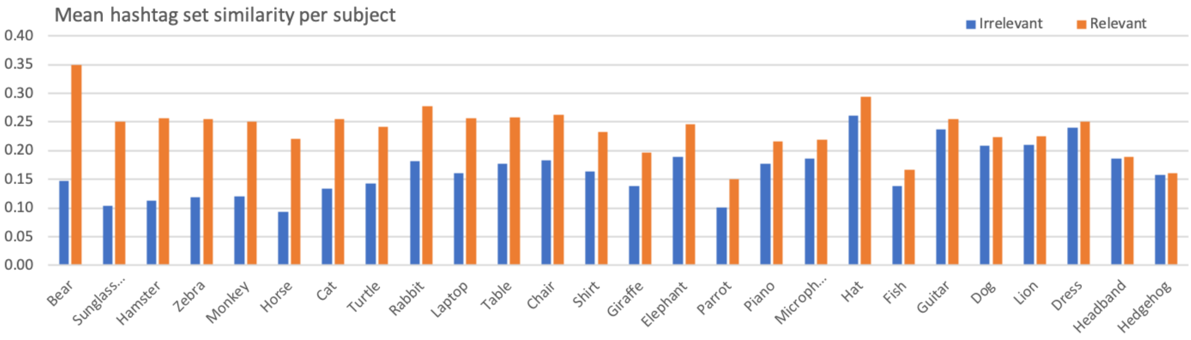


Figure 5.39: Mean hashtag sets similarity per subject for relevant & irrelevant posts

$d = 1.22)$[29].

In Table 5.31 we see the global (across all subjects) statistics regarding the hashtag sets similarities for the relevant and irrelevant posts. while in Figure 5.39 we see a per subject comparison. Detailed statistics are presented in Tables 5.26, 5.27 & 5.28, 5.29 for the relevant and irrelevant posts respectively. Fluctuations

[29] d is the Cohen coefficient denoting the effect size

Table 5.31: Average statistics of hashtag sets similarity across all subjects for relevant and irrelevant posts

	Mean	St. Dev.	Min	Max
Relevant Images	0.237	0.041	0.151	0.350
Non Relevant Images	0.165	0.043	0.094	0.261

across various subjects also appear here, as in the case of color histogram similarities, but they are a bit moderate. Once again the global mean for hashtag sets similarity in relevant posts is significantly higher than that of irrelevant posts ($t = 6.04$, $p < 0.01$, $df = 25$, $d = 1.71$).

In Fig. 5.40 we see the comparison[30] of mean similarity between filtered hashtag sets and color histograms for the relevant posts. The Pearson correlation is r_r=0.242 which is lower than the critical value r_c=0.33 obtained for df = 24 and level of significance a=0.05. Thus, the $H0_1$ null hypothesis that the similarity of color histograms and filtered hashtags sets, in the relevant posts, are significantly correlated cannot be rejected.

Fig. 5.41 examines the case of irrelevant posts. The Pearson correlation in that case is r_i=-0.255 showing that in the irrelevant posts information obtained through the color histograms and the hashtags sets are contradicting as one may expect from the fact that no similar visual content is shared within each subject.

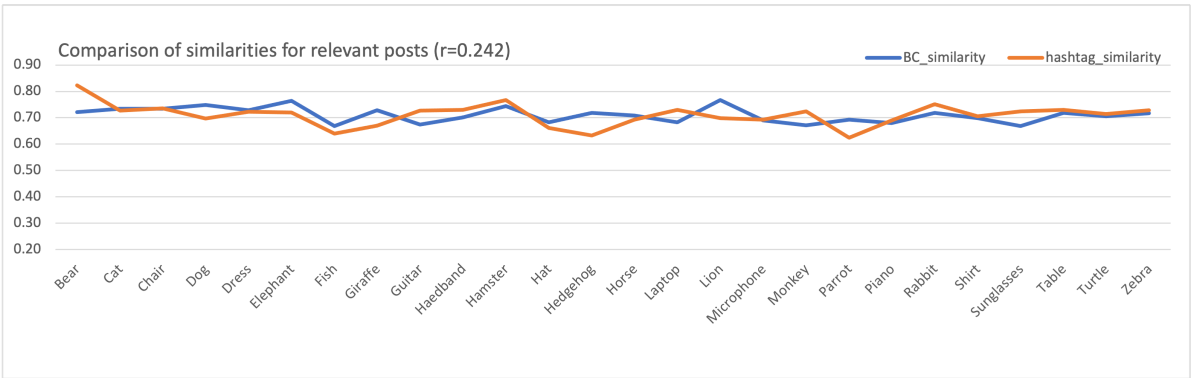


Figure 5.40: Mean similarities across all subjects for the color histogram and the hashtags sets for the relevant posts. For better visualisation equalisation of global means has been performed

In order to examine the second ($H0_2$) null hypothesis we have to compare two correlation coefficients and check the significance of their difference assuming a normal distribution.

The z-score of a correlation coefficient is obtained using the following formula [278]:

$$z_k = 0.5 \cdot log(\frac{1 + r_k}{1 - r_k}) \tag{5.4}$$

Using the previous formula we get the $z_r = 0.247$ and z_i=-0.261. In order to test the significance of z-score difference we need to normalize with the standard error (see eq. 5.6):

$$z = \frac{z_r - z_i}{\sigma_{z_r - z_i}} \tag{5.5}$$

[30] Equalization of means has been performed for better visualisation

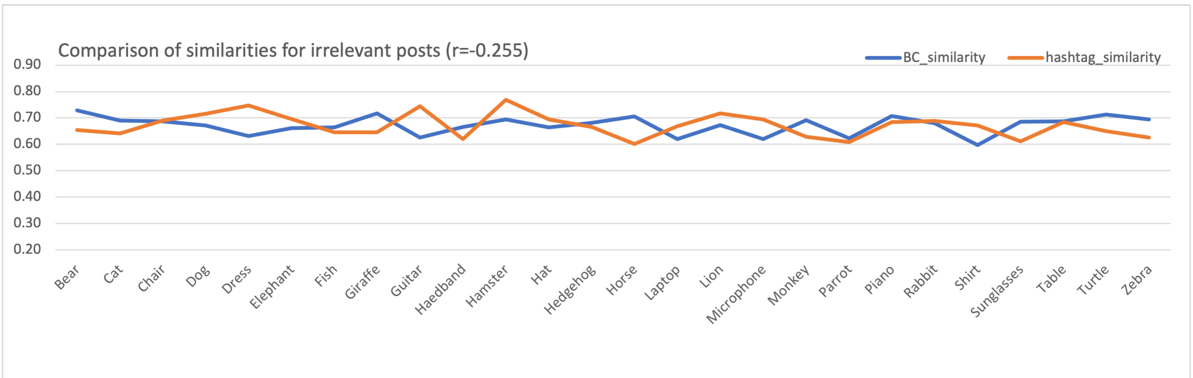


Figure 5.41: Mean similarities across all subjects for the color histogram and the hashtags sets for the irrelevant posts. For better visualisation equalisation of global means has been performed

where

$$\sigma_{z_r - z_i} = \sqrt{\frac{1}{n_r - 3} + \frac{1}{n_i - 3}} \tag{5.6}$$

Given that n_r=n_i = 26 (the number of subjects) from eq. 5.5 & 5.6 we get $z = 1.73$ which gives $p = 0.042$. Thus, the $H0_2$ null hypothesis is rejected at a significant level $a = 0.05$.

5.7 Transfer learning

In the previous section we have seen that there is significant difference between the correlation of visual content (at least in terms of color histograms) among Instagram photos obtained from relevant posts and irrelevant ones. We have also seen that there is significant difference between the Instagram hashtag sets obtained form relevant and irrelevant posts. However, no significant correlation was found between the Instagram hashtags sets and visual content (expressed via the color histogram) of the same Instagram posts. This means that we can not measure the visual similarity between two Instagram images, indirectly, by comparing the associated hashtag sets. On the other hand, wee have already shown that an important proportion of Instagram hashtags that accompany an Instagram photo describe its visual content via semantic terms. By combining this contradicting results we can conclude that probably the color histogram is not an ideal feature set for training concept models.

In the current study we try to examine whether Instagram photos collected through a query hashtag (subject) can be used for adapting concept models with the aid of transfer learning. Comparison with image retrieval using topic modelling of Instagram hashtags, as proposed by Tsapatsoulis [195], could be also done to assess whether there is a correlation between the performance of concept models, developed via transfer learning, and topic models trained on hashtags sets.

5.7.1 Data collection and pre-processing

For the development of concept models for image classification using transfer learning we used the same set of 26 subjects described in Sections 5.5.4 & 5.6. For each subject, we collected 100 images, constructing a dataset of 2600 Instragram photos in total. This dataset was divided into two parts: a training part, consisting of 2080 images (80% of the dataset) used for training, and the testing part, which consists of 520 pictures (20% of the dataset) that were used to evaluate the models we created in the training phase. The Python code we used for image classification was retrieved from GitHub project[31]. Image classification was based on *TensorFlow 2.x* and *TensorFlow Hub*. Convolutional Neural Networks were implemented in Python 3 under Google Colaboratory, referred to as Colab[32], which is an online platform for machine learning models that provides GPU and TPU options. In our study, we used GPU to run the code, which is more appropriate for CNN. The pre-trained ResNet-152[33] model was applied to the image classification process to reduce the training time and improve the classification accuracy. The corresponding training and test set labels of the images were saved in a CSV file with an images file path. In the CNN architecture, we added an output layer corresponding to the 26 labels of our study. We have to mention here that the initial output layer of the Resnet-182 network consists of 1000 labels. Figure 5.42 shows train image examples along with their labels.

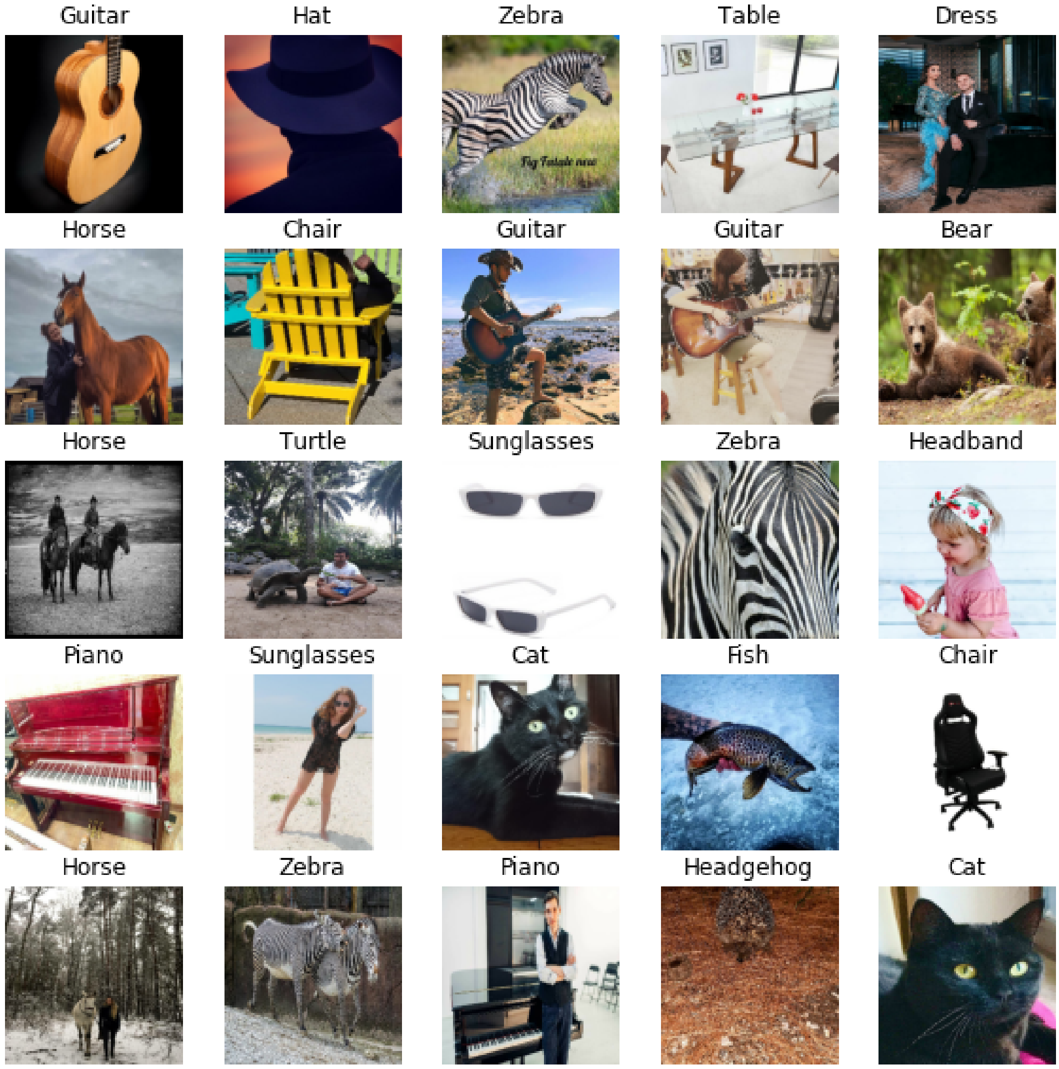


Figure 5.42: Example of training images with labels

[31] https://github.com/mrdbourke/zero-to-mastery-ml/tree/master/section-4-unstructured-data-projects?fbclid=IwAR2hGLqCQhIWN3e4LM7xs5xID4YXT1GeC6AZXoRCUWUrpO8FLLxln2maKO8

[32] https://colab.research.google.com/

[33] https://www.kaggle.com/pytorch/resnet152

5.7.2 Results

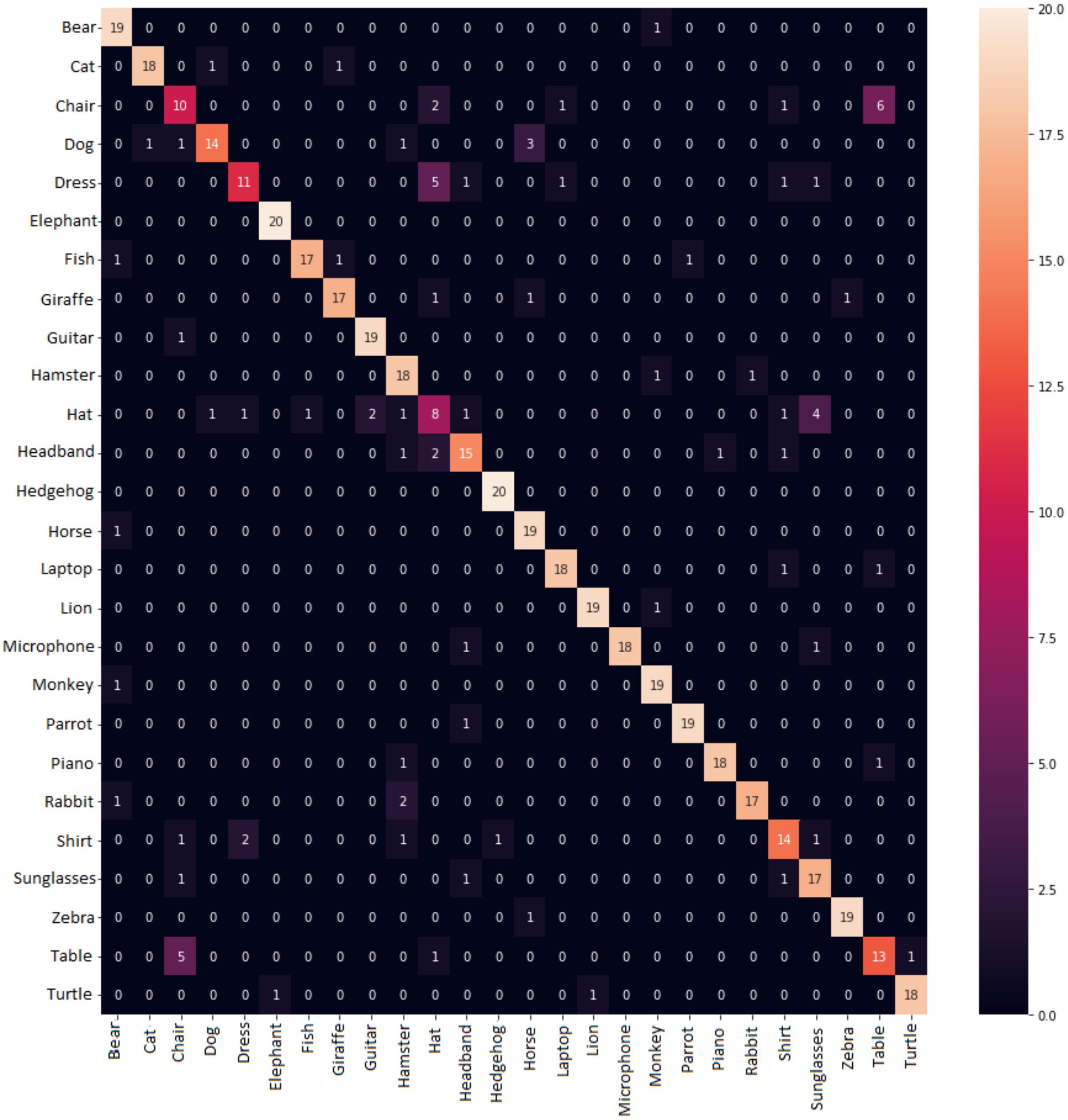
We tested the performance of the trained models using the 520 images (20 images per subject) that were reserved for the evaluation. Table 5.32 illustrates the performance scores of the trained concept models. As we can easily understand, for each subject in the training set were n=20 positive samples and 25xn = 500 negative samples corresponding to the other 25 subjects. We see there that the average recall performance is 0.90 and while the overall accuracy (how many of the 520 images across all subjects were correctly classified) is 0.83.

A careful inspection of the results presented in Table 5.32 shows that the minimum average recall value (0.40) has been obtained for the 'hat' subject while the maximum average recall value (1.00) has been obtained for the 'microphone' subject. As we have seen also in the previous section, Instagram photos containing a hat contain also other concepts, including dress, shirt, sunglasses, and others that do not correspond to trained models. As a result the trained models corresponding to those concepts produce also high output, which in some cases may be higher than that of the model corresponding to the 'hat' concept. This can be easily understood if we check the precision in those models ('dress' precision = 0.65, 'shirt' precision = 0.70) as well as the number of absolute misclassifications of hats into the sunglasses category (see Figure 5.43).

Table 5.32: Evaluation of image classification

Topic	Precision	Recall	F-measure	Accuracy
Bear	0.95	0.83	0.88	0.990
Cat	0.90	0.95	0.92	0.994
Chair	0.50	0.53	0.51	0.963
Dog	0.70	0.88	0.78	0.984
Dress	0.55	0.79	0.65	0.976
Elephant	1.00	0.95	0.98	0.998
Fish	0.85	0.94	0.89	0.992
Giraffe	0.85	0.89	0.87	0.990
Guitar	0.95	0.90	0.93	0.994
Hamster	0.90	0.72	0.80	0.982
Hat	0.40	0.42	0.41	0.955
Headband	0.75	0.75	0.75	0.980
Hedgehog	1.00	0.95	0.98	0.998
Horse	0.95	0.79	0.86	0.988
Laptop	0.90	0.90	0.90	0.992
Lion	0.95	0.95	0.95	0.996
Microphone	0.90	1.00	0.95	0.996
Monkey	0.95	0.86	0.90	0.992
Parrot	0.95	0.95	0.95	0.996
Piano	0.90	0.95	0.92	0.994
Rabbit	0.85	0.94	0.89	0.992
Shirt	0.70	0.70	0.70	0.976
Sunglasses	0.85	0.71	0.77	0.980
Zebra	0.95	0.95	0.95	0.996
Table	0.65	0.62	0.63	0.971
Turtle	0.90	0.95	0.92	0.994
Average	0.90	0.90	0.89	
Total Accuracy				0.83

The confusion matrix, regarding the absolute classifications, is shown in Figure 5.43. It is interesting to examine some specific cases such as the pair of concepts Chair-Table in which a significant mutual missclassification is observed. The chair and table subjects share a common semantic meaning and in several cases they appear together in the same image. We see that 6 out of the 20 testing photos containing chairs were misclassified to the Table category while 5 out of the 20 testing photos containing tables were misclassified to the Chair category. It is very likely that in those photos both concepts appear, thus, the classification is not entirely wrong. It is also interesting to see the triple of concepts Dress - Hat - Sunglasses. we observe in this case that a circular misclassification occurs: As shown in Figure 5.43, 6 out of the 20 testing photos containing dresses were misclassified to the Hat category while 4 out of the 20 testing photos containing hats were misclassified to the Sunglasses category.



	Bear	Cat	Chair	Dog	Dress	Elephant	Fish	Giraffe	Guitar	Hamster	Hat	Headband	Hedgehog	Horse	Laptop	Lion	Microphone	Monkey	Parrot	Piano	Rabbit	Shirt	Sunglasses	Zebra	Table	Turtle
Bear	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Cat	0	18	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chair	0	0	10	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	1	0	0	6	0
Dog	0	1	1	14	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
Dress	0	0	0	0	11	0	0	0	0	0	5	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0
Elephant	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fish	1	0	0	0	0	0	17	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Giraffe	0	0	0	0	0	0	0	17	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
Guitar	0	0	1	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hamster	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
Hat	0	0	0	1	1	0	1	0	2	1	8	1	0	0	0	0	0	0	0	0	0	1	4	0	0	0
Headband	0	0	0	0	0	0	0	0	0	1	2	15	0	0	0	0	0	0	0	1	0	1	0	0	0	0
Hedgehog	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
Horse	1	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0
Laptop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	1	0	0	1	0
Lion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	1	0	0	0	0	0	0	0	0
Microphone	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	18	0	0	0	0	0	1	0	0	0
Monkey	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
Parrot	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	19	0	0	0	0	0	0	0
Piano	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	18	0	0	0	0	1	0
Rabbit	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0
Shirt	0	0	1	0	2	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	14	1	0	0	0
Sunglasses	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	17	0	0	0
Zebra	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	19	0	0
Table	0	0	5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	13	1
Turtle	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	18

Figure 5.43: A confusion matrix for the classification results

The overall conclusion drawn from the previous cases is that the concept models learned from Instagram images via transfer learning are semantically consistent. Taking into account that the filtered hashtags sets accompanying Instagram photos are also semantically consistent (as shown in Section 5.5.4 we can derive to an indirect conclusion that the filtered hashtags sets are semantically correlated with the actual visual content of the Instagram photos that are appended to. This conclusion suggests that the application of topic modelling based image retrieval, as suggested by Tsapatsoulis [195], is consistent with image classification of Instagram photos based on deep learning. Which one of the two methods is better or whether they can be combined is left for further work.

5.8 Summary

In this chapter we have reported the findings of our study aiming to investigate whether Instagram photos, collected using a query hashtag, can be used in the context of transfer learning for the training of effective models for image classification. We have seen that this is indeed the case, and, furthermore, the derived concept models present a semantically sensitive performance with the most of the misclassifications occurred being explainable. In the next chapter we conclude the current thesis proving the overall conclusions drawn from the studies performed so far regarding the suitability of Instagram photos-hashtag pairs for developing concept models for AIA.

Chapter 6

Conclusion and Future Work

Conclusion contents

6.1 Introduction

This study aimed to determine if we can use Instagram images and hashtags for Automatic Image Annotation purposes. Appropriate image-tag pairs are vital in AIA to train concept models that produce reliable automatic tagging results. The current Ph.D. thesis argues that Instagram is a rich source of photo-tag pairs that could serve the AIA purpose through learnable concept models under appropriate pre-processing. The rationale behind this argumentation is that the users/owners of Instagram photos can also describe their content better than anyone. Seven research questions were developed to analyze for that research. The first research question focused on why we chose Instagram as a platform for AIA. The second analyzed the percentage of descriptive hashtags. The third focused on locating common non-descriptive hashtags. The fifth examined the implementation of HITS algorithm as a method of hashtag filtering. In the sixth research question, we used topic modeling to locate a descriptive hashtag for a category of images with the same hashtag. We employed transfer learning for image categorization in the seventh and last research question. A more detailed description of the conducted work follows.

6.2 Summary of the Study

In Chapter 1, we briefly indicated the purpose of the current study and explained the research questions addressed regarding the use of Instagram hashtags for Automatic Image Annotation purposes. We also explained the contribution of the current thesis to the overall body of knowledge related to the field of Automatic Image Annotation.

In Chapter 2, we presented and discussed the image retrieval methods and explained the methodology of automatic image annotation. First, the image retrieval framework is described, and then the basic steps are analyzed. Specifically, content-based image retrieval based on the image features, text-based image retrieval based on text from the web page, and automatic image retrieval is examined. The drawbacks of content-based and text-based are examined, and concluded that automatic image annotation could be a better solution for image retrieval. With that method, we can combine the approaches mentioned above. In addition, we answer the research question of why we chose Instagram for automatic image annotation purposes.

Chapter 3 synthesizes the foundation and recent literature related to the research questions addressed in the current thesis (see Section 1.2). A theoretical framework of hashtags was given, and creating sets for automatic image annotation was briefly described. The literature for common non-descriptive hashtags was also analyzed. Furthermore, the conceptual framework and the research conducted for the two filtering methodologies we propose (graph-based and topic modeling). Moreover, we examined the research related to image retrieval based on color histograms. We examined possible relations of color histograms and hashtag sets to clearly define the gap between high-level image semantics and low-level features. In the final section of the chapter, we surveyed the literature related to transfer learning.

Chapter 4 described the methodology and mathematical formulation for the study and explained the methodological framework for the research questions. Specifically, we described how we collected the Instagram posts and kept images and hashtags. Then we constructed the online questionnaire so the humans could select the hashtags that better describe the image's content. In order to locate stophashtags, we analyzed the framework we developed. Moreover, we described the methodology we followed to filter out hashtags with the help of the crowd and the implementation of the HITS algorithm. In addition, we analyzed the topic model approach for relevant hashtag identification in a category of images. Furthermore, we described the methodology we followed to quantify the similarity of the color histogram and hashtag sets. The last section of the chapter described the transfer learning method we implemented for image classification.

Chapter 5 described the data collection, the data analysis procedure, and the results. In that chapter, we see how with the help of an electronic questionnaire, we collected image annotation from librarians and students in order to check the descriptive value of the hashtags accompanying images in Instagram. Moreover, we explored stophashtag identification, analyzed data collection, and described the human evaluation for stophashtag location. In order to locate those hashtags that are suitable for AIA purposes, we have seen the implementation of the HITS algorithm in data from the crowd. We used the Appen crowdsourcing platform, and we used the HITS algorithm to examine if, with the help of that algorithm, we could select appropriate hashtags for AIA purposes. Topic model was another method we exploited to locate related hashtags based on a category of images. We created topic models based on a common subject (e.g. #dog), and we used topic coherence and human judgment to evaluate those models. Also, we present how we created the corpus and results related to our effort to use the color histogram for automatic image annotation purposes. We end that chapter with the results of transfer learning classification.

In Chapter 6, we describe the conclusion made from the findings and future directions of the research. In the final chapter, we summarize the study, conclude the research questions and future research.

6.3 Summary of Findings and Conclusion

Research question 1: The purpose of Research Question 1 (see Section 1.2) was related to the reasons we chose Instagram for AIA purposes instead of other social media platforms. Although we could consider other social media platforms for AIA purposes, Instagram is ideal. Instagram focuses on images, contains hashtags, is a more popular photo-oriented platform, and based on Google Scholar; we see high interest in the research community for that subject.

Research question 2: The purpose of Research Question 2 (see Section 1.2) was an attempt to investigate whether Instagram hashtags are suitable for image-tag pair in AIA and the portion of hashtags that describe the visual content of accompanying images. The experiment was conducted in two stages the preliminary and extended research. The participants in both stages were Librarians and undergraduate and undergraduate and postgraduate students. Results in the extended research confirmed those in the preliminary investigation and revealed that approximately 20% of Instagram hashtags are related to Instagram images' visual content. The results show also that an essential portion of image hashtags in Instagram are not directly related to the concept depicted by the image. We have also found that 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. So the results show that hashtags are suitable for image-tag pair in AIA, and 20% of Instagram hashtags are related to the visual content of accompanying image.

Research question 3: The purpose of Research Question 3 (see Section 1.2) was to propose a methodology for locating stophashtags, common non-descriptive hashtags. We defined a stophashtag score which proved to be very effective in modeling the likelihood for an Instagram hashtag to be a stophashtag. In an attempt to evaluate the effectiveness of the proposed method, 30 subjects/hashtags were chosen, and one photo containing descriptive hashtags and stophashtags was selected from each subject. Then eight hashtags, 2–3 descriptive and the rest stophashtags, were chosen among the hashtags used by the image owner. Two online questionnaires containing 15 images each were distributed to evaluators so they could choose the best suitable hashtag for every image according to their interpretation. Our hypothesis was that evaluators would not select stophashtags as descriptive to the questionnaire photos. We find that 26 out of 45 as stophashtags identified by the proposed algorithm coincide with -indirect- human judgment from the results and the evaluation process. Three (sun, girl, food) out of the 45 were erroneously considered stophashtags, while for 16 out of 45 we were unable to conclude because the users did not evaluate them as options in the questionnaires.

Overall, it appears that thresholding the stophashtag score for identifying the list of hashtags was not so effective at least when compared to human judgement. This is caused partially by the fact that Instagram users tend to consider as descriptive many hashtags that clearly lack descriptive power; thus, benchmarking the proposed method against indirect human evaluation, as we did in this study, may not be the most appropriate way of evaluation. On the other hand, examining the stophashtags we managed to locate, we can easily conclude that the majority of these hashtags are not descriptive and are used to fool the search results of the Instagram platform.

Research question 4: The purpose of Research Question 4 (see Section 1.2) was to propose a methodology for filtering Instagram hashtags based on the HITS algorithm and the wisdom of the crowd. To examine the proposed methodology we conducted an experiment in two stages the preliminary and the main research. In the preliminary stage we estimated the reliability of image annotators per image since the HITS algorithm is applied on bipartite networks created, each time, for a single image. The results were impressive and showed an indication that we can apply the method in a real crowdtagging environment. In the extended research we have empirically shown that the application of a two-step HITS algorithm in a crowdtagging context provides an easy and effective way to locate pairs of Instagram images and hashtags that can be used as training sets for content based image retrieval systems in the learning by example paradigm. As a proof of concept we have used 25000 evaluations (500 annotations for each one of 50 images) collected from the *Figure-eight* crowdsourcing platform to create a bipartite graph composed of users (annotators) and the tags they selected to describe the 50 images. The hub scores of the HITS algorithm applied on this graph, called hereby full bipartite graph, give us a measure of reliability of the annotators. The aforementioned approach is based on the findings of Theodosiou *et al.* [144] who claim that the reliability of annotators better approximated if we consider all the annotations they have performed rather than the subset of Gold Test Questions. In a second step a weighted bipartite graph for each image is composed in the same way as the full bipartite graph. The weights of these graphs are the hub scores computed in the previous step. By thresholding the authority scores of the per image graphs, obtained by the application of the HITS algorithm on the weighted graphs, we can rank and then effectively locate the hashtags that are relevant to their visual content as per the annotators evaluation.

Some important findings of the current work are briefly summarized here. The first refers to the value of crowdtagging itself. As in several studies before we found that the crowd can substitute the experts in the evaluation of images w.r.t. relevant tags. However, even with a large number of annotators (499 in our case) it seems that a perfect agreement between annotators and experts cannot be achieved. In particular, it was found that from the 145 different tags suggested for the 50 images used in this study by the two experts, only 135 were also identified by the 499 annotators. This leads to a maximum achievable recall value equal to 0.931. Thus, in subjective evaluation tasks, such as those referring to the identification of tags that are related with the visual content of images, no perfect agreement between the experts and the crowd should be expected.

A second finding is that crowdtagging of images can be effectively modeled through user-tag bipartite graphs, one per image. Thresholding the authority score of the HITS algorithm applied on these graphs is a robust way to identify the tags that characterize the visual content of the corresponding images. Getting the top ranked tags based on the authority score is an alternative solution, but, with a little bit lower effectiveness.

A final remark of the current study refers to the importance of using weighted user-tag bipartite graphs for the crowdtagged images. It appears that weighting the bipartite graphs with the hub scores of the annotators provides the best results. However, even in the case that the reliability metric of the crowdsourcing platform itself (the *_trust* variable of *Figure-eight* in our case) is used to weight the bipartite graphs the results are not significantly worse. We are a little bit reluctant to generalize this conclusion because in the current study we have used too many annotations (499) per image.

Research question 5: The purpose of Research Question 5(see Section 1.2) was to study the idea of using topic modelling as a means to filter out irrelevant hashtags that accompany Instagram images. The hashtags were grouped into subjects based on the results of Instagram queries and then topic models were created for each subject. The relevant hashtags of an Instagram image are the ones that coincide with the best matching topic. We have evaluated the effectiveness of topic modelling through coherence metrics. In that research we used both topic coherence as well as comparison between the topics, referring to the same subject, created from Instagram hashtags and Wikipedia content. From the results we can easily conclude that the words are associated with the subject.

Another topic evaluation was preformed with the help of human judgement. we have presented a crowd-based and student-based interpretation of word clouds created from Instagram hashtags. The main purpose was to examine if we can locate appropriate tags from Instagram photos that share (and grouped together) a common hashtag (called subject in the current work) for image metadata description. A statistical significant difference between the interpretation accuracy of relevant and irrelevant word clouds was found. This mean that Instagram images of similar visual content share hashtags that are related to the subject. In addition to these we concluded that there is correlation in interpretation of train students and the generic crowd, denoting that no specific training is mandatory to mine relevant tags from Instagram to describe photos. Moreover, since there is no difference in the interpretation accuracy performance of generic crowd and trained students we have an indication that indeed these hashtags can describe an image. In the results analysis we concluded that there is significant variation in the difficulty of interpretation of word clouds corresponding to different terms and we named three parameters affecting this interpretation: conceptual context, textual context and familiarity with concept. Terms that have a clear conceptual context (‘fish’, ‘guitar’, ‘laptop’), can be easily identified. On the contrary, term without clear conceptual context like ‘hat’ had as a result to confuse students and the crowd. In addition, terms like ‘hedgehog’ that students and crowd were no familiar had a difficult to interpret. The main conclusion is that we can use topic model to mine information from Instagram tags for image description metadata.

Research question 6: The purpose of Research Question 6 (see Section 1.2) was to examine the correlation between the color similarity of Instagram images and their filtered hashtag sets. While no statistical significant correlation between color histogram and hashtag sets similarity was found, the information seems complementary for image retrieval. This is supported by the fact that the difference in correlation between the similarity of color histograms and hashtag sets in relevant and irrelevant posts is both high and significant. This means that Instagram images of similar visual content, i.e., relevant posts share similar hashtags as well. This is not the case for Instagram images of varying visual content that share few (at least one) hashtags.

The purpose was to examine if can bridge the gap between low level feature and high-level semantic content. To achieve the aforementioned purpose, we calculated the correlation between color similarity of Instagram images and filtered hashtags. Proposing searching based algorithm and compute the computational complexity was out of the scope of the study. The results showing that the semantic gap was not fully covered. Although, concluding that color histogram and hashtag sets are complementary for image retrieval, especially for relevant posts, we have a strong indication to continue research to lower the semantic gap. Another important finding (expected though) is that both color histogram and hashtag sets similarities are significantly higher in relevant posts than in irrelevant ones. Thus, it is confirmed that both color histograms and hashtag sets provide important information related to the visual content of Instagram images. Comparing the effect size, as indicated by the Cohen d coefficient, for the color histograms and hashtags sets one may see that hashtag sets provide more rich information regarding the visual content of Instagram images within a specific subject.

Research question 7: The aim of Research Question 7 (see Section 1.2) was to investigate if the photo-tags we created from the previous steps are appropriate for automatic image annotation. The HITS algorithm and topic modeling methods we have already analyzed to filter out the irrelevant hashtags. So, it is essential to examine a methodology for image classification. To achieve the goal mentioned above, we used transfer learning to train the deep learning model for image classification. Specifically, with the help of Colab and the pre-trained model ResNet 152, we trained the model on Instagram images from 26 different subjects. The results show that the classification performance in 20 out of 26 was very high. The rest 6 subjects, as we previously analyzed(see section 5.7.2), have different shapes, sizes and, in some cases, have a visual relationship, and that had resulted in the low performance. The main conclusion is that we can use pre-trained models to classify with high recall images from Instagram.

6.4 Future Work

This thesis investigated whether we can use Instagram images and hashtags as pairs for automatic image annotation. Different methodologies were explored, and new ideas and techniques were proposed. This section highlights some directions for possible future research based on the findings and insights we gained through this research.

Instagram images and hashtags seem to be an attractive solution for automatic image annotation. In the research, we focused on hashtags in the English language since English is the most common language used in Social Media and the Web(25,9% of the Internet users write in English[1]). It is necessary to expand the research for non-English hashtags and check if we can imply the methodologies we suggested in hashtags from other languages. It would also be interesting to identify any cultural differences in how Instagram hashtags are used.

Mask R-CNN [279] and YOLO [280] are the most frequently used object detection techniques that can locate instances of semantic objects of a specific class in digital images and videos. These methods can detect the region of the object and then assign an object class for each of the proposed regions. Implementing the aforementioned object detection methods in Instagram images would be interesting and compared with the hashtag annotation results we concluded in the current thesis.

A fully automated system for image annotation based on Instagram images and hashtags as per the methodologies / ideas presented in the current thesis is a highly desirable continuation of the current work. Towards this direction, an Application Programming Interface would automatically acquire images from Instagram along with their hashtags filter out irrelevant hashtags and images to create datasets for automatic image annotation purposes, possibly through deep learning.

We strongly believe that topic modeling of Instagram hashtags, coined in this thesis, is a field where additional research can be conducted, and new ideas can be investigated. The results of this thesis were more than promising, and investing in this topic is highly recommended.

As in every Ph.D. thesis, many others arise during the investigation of one research problem. But this is the perpetual research cycle ...

[1] https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

Bibliography

[1] M. Alkhawlani, M. Elmogy, and H. El Bakry, “Text-based, content-based, and semantic-based image retrievals: A survey,” *International Journal of Computer and Information Technology*, vol. 4, no. 1, pp. 58–66, January 2015. [Online]. Available: https://www.ijcit.com/archives/volume4/issue1/Paper040109.pdf

[2] Z. Theodosiou, “Image retrieval: modelling keywords via low-level features,” Ph.D. dissertation, Cyprus University of Technology, Cyprus, 2014. [Online]. Available: https://ktisis.cut.ac.cy/handle/10488/3546

[3] V. Maihami and F. Yaghmaee, “Fuzzy neighbor voting for automatic image annotation,” *International Journal of Multimedia and Ubiquitous Engineering*, vol. 4, no. 1, pp. 1–8, 2016. [Online]. Available: https://doi.org/10.22061/jecei.2016.526

[4] M. Zappavigna, “Twitter,” in *Pragmatics of Social Media*, C. R. Hoffmann and W. Bublitz, Eds. Berlin/Boston: De Gruyter Mouton, 2017, ch. 8, p. 201–224. [Online]. Available: https://doi.org/10.1515/9783110431070-008

[5] V. Dey, Y. Zhang, and M. Zhong, “A review on image segmentation techniques with remote sensing perspective,” in *ISPRS TC VII Symposium – 100 Years ISPRS*. Vienna University of Technology, July 2010, pp. 31 – 42. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.466.8028&rep=rep1&type=pdf

[6] A. Yu. How to teach a computer to see with convolutional neural networks. [Online]. Available: https://towardsdatascience.com/how-to-teach-a-computer-to-see-with-convolutional-neural-networks-96c120827cd1

[7] P. Gudikandula. Deep view on transfer learning with iamge classification pytorch. [Online]. Available: https://purnasaigudikandula.medium.com/deep-view-on-transfer-learning-with-iamge-classification-pytorch-5cf963939575

[8] T. Berners-Lee, R. Cailliau, J.-F. Groff, and B. Pollermann, “World-wide web: the information universe,” *Internet Research*, vol. 20, no. 4, pp. 461–471, 2010. [Online]. Available: https://doi.org/10.1108/10662241011059471

[9] O. O. Ajayi and D. M. Elegbeleye, “Performance evaluation of selected search engines,” *IOSR Journal of Engineering*, vol. 4, no. 2, pp. 1–12, February 2014. [Online]. Available: https://doi.org/10.9790/3021-04210112

[10] E. Çakir, H. Bahceci, and Y. Bitirim, “An evaluation of major image search engines on various query topics,” in *The Third International Conference on Internet Monitoring and*

Protection. Los Alamitos, California: IEEE, July 2008, pp. 161 – 165. [Online]. Available: https://doi.org/10.1109/ICIMP.2008.9

[11] F. Kang, "Automatic image annotation," Ph.D. dissertation, Michigan State University, Michigan, USA, 2007. [Online]. Available: https://www.proquest.com/docview/304839065/DFA12295822C4C3CPQ/1?accountid=36246

[12] N. Tsapatsoulis, "Web image indexing using wice and a learning-free language model," in *Artificial Intelligence Applications and Innovations: 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016 Thessaloniki, Greece, September 16–18, 2016: Proceedings*. Switzerland: Springer Nature, September 2016, pp. 131–140. [Online]. Available: https://doi.org/10.1007/978-3-319-44944-9_12

[13] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Lia, "A survey and analysis on automatic image annotation," *Pattern Recognition*, vol. 79, pp. 242–259, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2018.02.017

[14] Y. Chen, X. Zeng, X. Chen, and W. Guo, "A survey on automatic image annotation," *Applied Intelligence*, vol. 50, no. 10, pp. 3412 – 3428, October 2020. [Online]. Available: https://doi.org/10.1007/s10489-020-01696-2

[15] S. Giannoulakis and N. Tsapatsoulis, "Evaluating the descriptive power of instagram hashtags," *Journal of Innovation in Digital Ecosystems*, vol. 3, no. 2, pp. 114–129, December 2016. [Online]. Available: https://doi.org/10.1016/j.jides.2016.10.001

[16] ——, "Instagram hashtags as image annotation metadata," in *Artificial Intelligence Applications and Innovations : 11th IFIP WG 12.5 International Conference, AIAI 2015, Bayonne, France, September 14-17, 2015, Proceedings*. Cham: Springer, September 2015, pp. 206–220. [Online]. Available: https://doi.org/10.1007/978-3-319-23868-5_15

[17] ——, "Defining and identifying stophashtags in instagram," in *Advances in Big Data : Proceedings of the 2nd INNS Conference on Big Data, October 23-25, 2016, Thessaloniki, Greece*. Cham: Springer, October 2016, pp. 304–313. [Online]. Available: https://doi.org/10.1007/978-3-319-47898-2_31

[18] S. Giannoulakis, N. Tsapatsoulis, and K. S. Ntalianis, "Identifying image tags from instagram hashtags using the hits algorithm," in *2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing/2017 IEEE 15th International Conference on Pervasive Intelligence and Computing/2017 IEEE 3rd International Conference on Big Data Intelligence and Computing/2017 IEEE Cyber Science and Technology Congress DASC-PICom-DataCom-CyberSciTec 2017 : 6-11 November 2017, Orlando, Florida : proceedings*. Piscataway, NJ: IEEE, November 2017, pp. 89–94. [Online]. Available: https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.29

[19] S. Giannoulakis and N. Tsapatsoulis, "Filtering instagram hashtags through crowdtagging and the hits algorithm," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, June 2019. [Online]. Available: https://doi.org/10.1109/TCSS.2019.2914080

[20] ——, “Topic identification of instagram hashtag sets for image tagging: An empirical assessment,” in *15th International Conference on Metadata and Semantics Research*, November 2021.

[21] A. Argyrou, S. Giannoulakis, and N. Tsapatsoulis, “Topic modelling on instagram hashtags: An alternative way to automatic image annotation?” in *13th International Workshop on Semantic and Social Media Adaptation and Personalization SMAP 2018 : September 6-7, 2018, Zaragoza, Spain : proceedings*. Piscataway, NJ: IEEE, September 2018, pp. 61–67. [Online]. Available: https://doi.org/10.1109/SMAP.2018.8501887

[22] S. Giannoulakis and N. Tsapatsoulis, “Topic identification via human interpretation of word clouds: The case of instagram hashtags,” in *17th International Conference on Artificial Intelligence Applications and Innovations*. Springer, June 2021, pp. 283–294.

[23] T. Berners-Lee. (1989) Information management: A proposal. [Online]. Available: https://www.w3.org/History/1989/proposal.html

[24] C. Aced Toledano, “Web 2.0: the origin of the word that has changed the way we understand public relations,” in *Proceedings International PR Conference: Images of Public Relations*. World Scientific, July 2013, pp. 1–12.

[25] M. McLuhan, *The Gutenberg galaxy : the making of typographic man*. Toronto: University of Toronto Press, 1962.

[26] N. Negroponte, *Being Digital*. NY, USA: Random House Inc., 1995.

[27] M. L. Dertouzos, *What will be: how the new world of information will change our lives*. USA: HarperEdge, 1998.

[28] T. O’Reilly. What is web 2.0: Design patterns and business models for the next generation of software. [Online]. Available: https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

[29] B. Marr. How much data do we create every day? the mind-blowing stats everyone should read. [Online]. Available: https://www.bernardmarr.com/default.asp?contentID=1438

[30] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009.

[31] V. Tyagi, *Content-Based Image Retrieval: Ideas, Influences, and Current Trends*. Singapore: Springer, 2017. [Online]. Available: https://doi.org/10.1007/978-981-10-6759-4

[32] Y. Rui, T. S. Huanga, and S.-F. Chang, “Image retrieval: Current techniques, promising directions, and open issues,” *Journal of Visual Communication and Image Representation*, vol. 10, no. 1, March 1999. [Online]. Available: https://doi.org/10.1006/jvci.1999.0413

[33] J. Li and J. Z. Wang, “Real-time computerized annotation of pictures,” in *Fourteenth ACM International Conference on Multimedia*. New York: ACM, October 2006. [Online]. Available: https://doi.org/10.1145/1180639.1180841

[34] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Lia, “Image retrieval using multiple evidence ranking,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 408–417, 2004. [Online]. Available: https://doi.org/10.1109/TKDE.2004.1269666

[35] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, April 2008. [Online]. Available: https://doi.org/10.1145/1348246.1348248

[36] J. Fauqueur and N. Boujemaa, "New image retrieval paradigm: logical composition of region categories," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*. IEEE, September 2003, pp. 601 – 604. [Online]. Available: https://doi.org/10.1109/ICIP.2003.1247316

[37] Y. Tan, "Applications," in *Gpu-Based Parallel Implementation of Swarm Intelligence Algorithms*, W.-K. Chen, Ed. Amsterdam: Elsevier, 2016, ch. 11, pp. 167–177. [Online]. Available: https://doi.org/10.1016/B978-0-12-809362-7.50011-X

[38] A. P. Vartak and V. Mankar, "Colour image segmentation - a survey," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 2, pp. 681 – 688, February 2013. [Online]. Available: https://ijetae.com/files/Volume3Issue2/IJETAE_0213_115.pdf

[39] Y. Aslandogan and C. Yu, "Techniques and systems for image and video retrieval," *Techniques and systems for image and video retrieval*, vol. 11, no. 1, pp. 56 – 63, February 1999. [Online]. Available: https://doi.org/10.1109/69.755615

[40] Y. Ramadevi, T. Sridevi, B. Poornima, and B. Kalyani, "Segmentation and object recognition using edge detection techniques," *International Journal of Computer Science & Information Technology*, vol. 2, no. 6, pp. 153–161, December 2010. [Online]. Available: https://doi.org/10.5121/ijcsit.2010.2614

[41] O. N. Baumann, "Connected operators for unsupervised image segmentation," Ph.D. dissertation, University of Southampton, Southampton, UK, 2004. [Online]. Available: https://eprints.soton.ac.uk/66319/

[42] B. Sathya and R. Manavalan, "Image segmentation by clustering methods: Performance analysis," *International Journal of Computer Applications*, vol. 29, no. 11, pp. 27 – 32, September 2011. [Online]. Available: https://research.ijcaonline.org/volume29/number11/pxc3875127.pdf

[43] R. Albatal, P. Mulhem, T.-J. Chin, and Y. Chiaramella, "Comparing image segmentation algorithms for content based image retrieval systems," in *Proceedings Of The Singaporean-French Ipal Symposium 2009: SinFra'09*. Singapore: World Scientific, February 2009, pp. 66–75. [Online]. Available: https://doi.org/10.1142/9789814277563_0007

[44] N. Hervé and N. Boujemaa, "Automatic image annotation," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer, 2009, pp. 180–187. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_1010

[45] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas, "Automatic image annotation using group sparsity," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, June 2010, p. 3312–3319. [Online]. Available: https://doi.org/10.1109/CVPR.2010.5540036

[46] K. Amiri and M. Farah, "Graph of concepts for semantic annotation of remotely sensed images based on direct neighbors in rag," *Canadian Journal of Remote Sensing*, vol. 44, no. 6, pp.

551–574, 2018. [Online]. Available: https://doi.org/10.1080/07038992.2019.1569507

[47] Y. Hou and Z. Lin, “Image tag completion and refinement by subspace clustering and matrix completion,” in *IEEE Visual Communications and Image Processing*. Los Alamitos, California: IEEE, December 2015. [Online]. Available: https://doi.org/10.1109/VCIP.2015.7457875

[48] L. Wu, R. Jin, and A. K. Jain, “Tag completion for image retrieval,” *Tag Completion for Image Retrieval*, vol. 35, no. 3, pp. 716 – 727, March 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.124

[49] C. R. Hoffmann, “Log in: Introducing the pragmatics of social media,” in *Pragmatics of Social Media*, C. R. Hoffmann and W. Bublitz, Eds. Berlin/Boston: De Gruyter Mouton, 2017, ch. 1, p. 1–28. [Online]. Available: https://doi.org/10.1515/9783110431070-001

[50] K. Skemp, “Facebook,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2020.

[51] A. Rosen. Tweeting made easier. [Online]. Available: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

[52] C. Wolan. The real story of twitter. [Online]. Available: https://www.forbes.com/sites/christianwolan/2011/04/14/the-real-story-of-twitter/?sh=2026bf8166af

[53] M. Johansson, “Youtube,” in *Pragmatics of Social Media*, C. R. Hoffmann and W. Bublitz, Eds. Berlin/Boston: De Gruyter Mouton, 2007, ch. 7, p. 173–200. [Online]. Available: https://doi.org/10.1515/9783110431070-007

[54] K. Smith. 57 fascinating and incredible youtube statistics. [Online]. Available: https://www.brandwatch.com/blog/youtube-stats/

[55] P. G. Cooper, “Social media,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2018.

[56] ——, “Flickr,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2020.

[57] A. Smock, “Automatic image annotation,” Ph.D. dissertation, Michigan State University, Michigan, USA, 2012. [Online]. Available: https://d.lib.msu.edu/etd/1127

[58] P. G. Cooper, “Pinterest,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2020.

[59] Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel, “Visual search at pinterest,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM, 2015, p. 1889–1898. [Online]. Available: https://doi.org/10.1145/2783258.2788621

[60] C. Sanchez, *Pinterest : for business & you*. Boca Raton, FL: BarCharts, Inc., 2015.

[61] J. Elliott, “Using snapchat to connect with generation z and millennials,” *Parks & Recreation*, vol. 54, no. 7, p. 16 – 19, 2019. [Online]. Available: https://www.nrpa.org/parks-recreation-magazine/2019/july/using-snapchat-to-connect-with-generation-z-and-millennials/

[62] M. Mazzei, “Snapchat,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2020.

[63] B. Stanley, “Uses and gratifications of temporary social media: A comparison of snapchat and facebook,” Master’s thesis, California State University, USA, 2015. [Online]. Available: https://www.proquest.com/docview/1681985191/47BF61B6B4F14EDBPQ/1?accountid=36246

[64] S. Ta, “The role of visual media in image-based social networking sites in regard to self-expression and social communication; an empirical-qualitative investigation of image usage in communication via instagram,” Master’s thesis, Vienna University of Technology, Vienna, Austria, 2018. [Online]. Available: https://repositum.tuwien.at/handle/20.500.12708/7313

[65] B. Holak. Instagram. [Online]. Available: https://searchcio.techtarget.com/definition/Instagram

[66] M. Garber. Instagram was first called ‘burbn’: Yes, after the drink. [Online]. Available: https://www.theatlantic.com/technology/archive/2014/07/instagram-used-to-be-called-brbn/373815/

[67] S. Sengupta, N. Perlroth, and J. Wortham. Behind instagram’s success, networking the old way. [Online]. Available: https://www.cnbc.com/id/47049161

[68] M. Baranovic. (2013) What #hashtags mean to mobile photography. [Online]. Available: https://www.dpreview.com/post/1256293279/hastag-photography

[69] P. G. Cooper, “Instagram,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2018.

[70] J. Constine. Instagram launches “stories,” a snapchatty feature for imperfect sharing. [Online]. Available: https://techcrunch.com/2016/08/02/instagram-stories/

[71] L. Kolowich Cox. The history of hashtags. [Online]. Available: https://blog.hubspot.com/marketing/history-of-hashtags

[72] T. A. Small, “What the hashtag?: A content analysis of canadian politics on twitter,” *Information, Communication & Society*, vol. 14, no. 6, p. 872 –895, September 2011. [Online]. Available: https://doi.org/10.1080/1369118X.2011.554572

[73] A. Ulges, M. Worring, and T. Breuel, “Learning visual contexts for image annotation from flickr groups,” *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 330 – 341, April 2011. [Online]. Available: https://doi.org/10.1109/TMM.2010.2101051

[74] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, and G. Thallinger, “Automatic image annotation using visual content and folksonomies,” *Multimedia Tools and Applications*, vol. 42, p. 97–113, March 2009. [Online]. Available: https://doi.org/10.1007/s11042-008-0247-7

[75] L.-C. Hsieh and W. H. Hsu, “Search-based automatic image annotation via flickr photos using tag expansion,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Los Alamitos, California: IEEE, March 2010, pp. 2398–2401.

[76] H. Xu, X. Zhou, M. Wang, Y. Xiang, and B. Shi, “Exploring flickr’s related tags for semantic annotation of web images,” in *Proceeding of the ACM International Conference on Image and Video Retrieval ACM-CIVR 2009*. New York: ACM, July 2009, p. 1–8. [Online]. Available: https://doi.org/10.1145/1646396.1646450

[77] L. J. Bynum, “Crowdsourcing,” in *Salem Press Encyclopedia*. Amenia, NY: Salem Press, 2019.

[78] M. Hirth, T. Hoßfeld, and P. Tran-Gia, “Anatomy of a crowdsourcing platform - using the example of microworkers.com,” in *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Los Alamitos, California: IEEE, July 2011, pp. 322–329. [Online]. Available: https://doi.org/10.1109/IMIS.2011.89

[79] J. Surowiecki, *The Wisdom of Crowds*. New York: Anchor, 2005.

[80] H. Oinas-Kukkonen, “Network analysis and crowds of people as sources of new organisational knowledge,” in *Knowledge Management: Theoretical Foundation*. Santa Rosa, CA, US: Informing Science Press, 2008, ch. 6, pp. 173–189.

[81] M. Everman and C. Leider, “Toward digital musical instrument evaluation using crowd-sourced tagging techniques,” in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2013. [Online]. Available: https://www.nime.org/proceedings/2013/nime2013_251.pdf

[82] N. J. Nilsson. Introduction to machine learning. [Online]. Available: https://ai.stanford.edu/~nilsson/MLBOOK.pdf

[83] R. F. De Mello and M. A. Ponti, *Machine Learning:A Practical Approach on the Statistical Learning Theory*. Switzerland: Springer Nature, 2018.

[84] P. P. Angelov and X. Gu, “Brief introduction to statistical machine learning,” in *Empirical Approach to Machine Learning*. Cham, Switzerland: Springer Nature, 2019, ch. 2. [Online]. Available: https://doi.org/10.1007/978-3-030-02384-3_2

[85] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, “Subject independent facial expression recognition with robust face detection using a convolutional neural network,” *Neural Networks*, vol. 16, no. 5-6, pp. 555–559, 2003. [Online]. Available: https://doi.org/10.1016/S0893-6080(03)00115-1

[86] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, p. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[87] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, “Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment,” in *Inclusive Smart Cities and Digital Health: 14th International Conference on Smart Homes and Health Telematics, ICOST 2016 Wuhan, China, May 25–27, 2016: Proceedings*. Switzerland: Springer Nature, May 2016, p. 37–48. [Online]. Available: https://doi.org/10.1007/978-3-319-39601-9_4

[88] B. Zhang, “Gaussian processes based transfer learning for online multiple-person tracking and building blocks for deep learning,” Ph.D. dissertation, Brunel University London, London, UK, 2020. [Online]. Available: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.814454

[89] E. Christinaki, “Bayesian transfer learning for personalised well-being forecasting from scarce, sporadic observations,” Ph.D. dissertation, University of Essex, Essex, UK, 2021. [Online]. Available: http://repository.essex.ac.uk/id/eprint/29568

[90] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345 – 1359, October 2010. [Online]. Available: https://doi.org/10.1109/TKDE.2009.191

[91] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep learning for computer vision: A brief review,” *Computational Intelligence and Neuroscience*, vol. 2018, 2018. [Online]. Available: https://doi.org/10.1155/2018/7068349

[92] W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review,” *Neural Computation*, vol. 29, no. 9, p. 2352–2449, September 2018. [Online]. Available: https://doi.org/10.1162/neco_a_00990

[93] R. Ek, “Automatic image annotation using transfer learning on convolutional neural networks,” Master’s thesis, Åbo Akademi, Finland, 2018. [Online]. Available: https://www.doria.fi/handle/10024/158799

[94] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh, “Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks,” in *Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications - IPTA 2017 Montreal, Canada, November 28-December 1*. Los Alamitos, CA: IEEE, December 2017. [Online]. Available: https://doi.org/10.1109/IPTA.2017.8310149

[95] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, “Automatic image annotation via label transfer in the semantic space,” *Pattern Recognition*, vol. 71, pp. 144–157, November 2017. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.05.019

[96] A. Singla, L. Yuan, and T. Ebrahimi, “Food/non-food image classification and food categorization using pre-trained googlenet model,” in *MADiMa’16 : proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management : October 16, 2016, Amsterdam, the Netherlands*. New Yok: ACM, October 2016, p. 3–11. [Online]. Available: https://doi.org/10.1145/2986035.2986039

[97] J. Ma, F. Wu, J. Zhu, D. Xu, and D. Kong, “A pre-trained convolutional neural network based method for thyroid nodule diagnosis,” *Ultrasonics*, vol. 73, pp. 221–230, 2017. [Online]. Available: https://doi.org/10.1016/j.ultras.2016.09.011

[98] B. A. . M. Ashqar and S. S. Abu-Naser, “Identifying images of invasive hydrangea using pre-trained deep convolutional neural networks,” *International Journal of Academic Engineering Research*, vol. 3, no. 3, pp. 28–36, March 2019. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3369016

[99] A. Hanbury, “A survey of methods for image annotation,” *Journal of Visual Languages & Computing*, vol. 19, no. 5, pp. 617–627, October 2008. [Online]. Available: https://doi.org/10.1016/j.jvlc.2008.01.002

[100] C. Jin and S.-W. Jin, “Automatic image annotation using feature selection based on improving quantum particle swarm optimization,” *Signal Processing*, vol. 109, no. 2, pp. 172–181, April 2015. [Online]. Available: https://doi.org/10.1016/j.sigpro.2014.10.031

[101] C. G. M. Snoek and M. Worring, “Concept-based video retrieval,” *Foundations and Trends® in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2009. [Online]. Available: https://doi.org/10.1561/1500000014

[102] Z. Theodosiou and N. Tsapatsoulis, “Crowdsourcing annotation: Modelling keywords using low level features,” in *IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*. Piscataway, NJ: IEEE, December 2011. [Online]. Available: https://doi.org/10.1109/IMSAA.2011.6156351

[103] Z. Zdziarski, C. Bourgès, J. Mitchell, P. Houdyer, D. Johnson, and R. Dahyot, “On summarising the ‘here and now’ of social videos for smart mobile browsing,” in *2014 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM) 1-2 Nov. 2014, Paris*. Los Alamitos, CA: IEEE, November 2014. [Online]. Available: https://doi.org/10.1109/IWCIM.2014.7008797

[104] A. R. Daer, R. F. Hoffman, and S. Goodman, “Rhetorical functions of hashtag forms across social media applications,” *Communication Design Quarterly Review*, vol. 3, no. 1, pp. 12–16, 2014. [Online]. Available: https://doi.org/10.1145/2721882.2721884

[105] B. Chacon. 5 ways to rock a branded instagram hashtag. [Online]. Available: https://later.com/blog/branded-instagram-hashtag-guide

[106] S. M. Mohammad and S. Kiritchenko, “Using hashtags to capture fine emotion categories from tweets,” *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, May 2015. [Online]. Available: https://doi.org/10.1111/coin.12024

[107] F. Kunneman, C. Liebrecht, and 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)*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 26–34. [Online]. Available: https://www.aclweb.org/anthology/W14-1304

[108] D. M. Carmean and M. E. Morris. Selfie examinations: Applying computer vision, hashtag scraping and sentiment analysis to finding and interpreting selfies. [Online]. Available: http://nebula.wsimg.com/27bab6eda0e75b69fcab8a5cdc4e22af?AccessKeyId=A6A4DAF733A0F616E396

[109] Y. Zhang, F. Baghirov, H. Hashim, and J. Murphy, “Gender and instagram hashtags: A study of #malaysianfood,” in *Proceedings of the ‘eTourism: Empowering Places’*, February 2016. [Online]. Available: http://agrilife.org/ertr/files/2016/01/ENTER2016_submission_118_.pdf

[110] E. Ferrara, R. Interdonato, and A. Tagarelli, “Online popularity and topical interests through the lens of instagram,” in *HT ’14 : proceedings of the 25th ACM Conference on Hypertext and Social Media : September 1-4, 2014, Santiago, Chile*. Association for Computing Machinery, September 2014, p. 24–34. [Online]. Available: https://doi.org/10.1145/2631775.2631808

[111] F. Fauzi, J.-L. Hong, and M. Belkhatir, “Webpage segmentation for extracting images and their surrounding contextual information,” in *MM ’09 : proceedings of the 2009 ACM Multimedia Conference & co-located workshops : October 19-24, 2009, Beijing, China : AMC ’09, CEA ’09, EiMM ’09, IMCE ’09, LS-MMRM ’09, MiFor ’09, MSIADU ’09, MTDL ’09, SSCS ’09, WSM ’09, & WSMC ’09*. Association for Computing Machinery, October 2009, p. 649–652. [Online]. Available: https://doi.org/10.1145/1631272.1631379

[112] P. M. Joshi and S. Liu, “Web document text and images extraction using dom analysis and natural language processing,” in *DocEng’09 : proceedings of the 2009 ACM Symposium on Document Engineering, Munich, Germany, September 15-18, 2009 00*. New York, USA: ACM, September 2009, p. 218–221. [Online]. Available: https://doi.org/10.1145/1600193.1600241

[113] G. Tryfou, Z. Theodosiou, and N. Tsapatsoulis, “Web image context extraction based on semantic representation of web page visual segments,” in *SMAP 2012 : proceedings : 2012*

Seventh International Workshop on Semantic and Social Media Adaptation and Personalization : 3-4 December 2012 : Luxembourg, Luxembourg. Piscataway, NJ: IEEE, December 2012, pp. 63–67. [Online]. Available: https://doi.org/10.1109/SMAP.2012.13

[114] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 754 – 766, April 2011. [Online]. Available: https://doi.org/10.1109/TPAMI.2010.133

[115] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proceedings of the 17th International Conference on World Wide Web, April 21 - 25, 2008, Beijing, China*. USA: ACM, April 2008, p. 327–336. [Online]. Available: https://doi.org/10.1145/1367497.1367542

[116] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conferenceon Research and Development in Information Retrieval : SIGIR 2005 ; August 15-19, 2005, Salvador, Brazil*. New York, USA: ACM, 2005, p. 154–161. [Online]. Available: https://doi.org/10.1145/1076034.1076063

[117] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on : date, 7-12 June 2015*. Los Alamitos, CA: IEEE, June 2015, p. 3128–3137. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298932

[118] W. Wang and Z.-H. Zhou, "Crowdsourcing label quality: a theoretical analysis," *Science China Information Sciences*, vol. 58, p. 1–12, November 2015. [Online]. Available: https://doi.org/10.1007/s11432-015-5391-x

[119] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel, "Learning automatic concept detectors from online video," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 429 – 438, April 2010. [Online]. Available: https://doi.org/10.1016/j.cviu.2009.08.002

[120] X.-J. Wang, W.-Y. Ma, and X. Li, "Exploring statistical correlations for image retrieval," *Multimedia Systems*, vol. 11, no. 4, p. 340–351, April 2006. [Online]. Available: https://doi.org/10.1007/s00530-006-0013-5

[121] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel, "Content analysis meets viewers: linking concept detection with demographics on youtube," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 2, pp. 145 – 157, June 2013. [Online]. Available: https://doi.org/10.1007/s13735-012-0029-x

[122] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, August 2009, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[123] X. Chen, A. Shrivastava, and A. Gupta, "Neil: Extracting visual knowledge from web data," in *2013 IEEE International Conference on Computer Vision : 1-8 December 2013*. IEEE, December 2013, pp. 1409 – 1416. [Online]. Available: https://doi.org/10.1109/ICCV.2013.178

[124] N. H. Do and K. Yanai, “Automatic construction of action datasets using web videos with density-based cluster analysis and outlier detection,” in *Image and video technology : PSIVT 2015 workshops : RV 2015, GPID 2013, VG 2015, EO4AS 2015, MCBMIIA 2015, and VSWS 2015, Auckland, New Zealand, November 23-27, 2015 : revised selected papers*. Springer, November 2015, pp. 160–172. [Online]. Available: https://doi.org/10.1007/978-3-319-29451-3_14

[125] K. Ntalianis, N. Tsapatsoulis, A. Doulamis, and N. Matsatsinis, “Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution,” *Multimedia Tools and Applications*, vol. 69, no. 2, p. 397–421, March 2014. [Online]. Available: https://doi.org/10.1007/s11042-012-0995-2

[126] J. Cui, L. Liu, H. Wang, C. Du, and W. Song, “Tagged image clustering via topic models,” in *2015 27th Chinese Control and Decision Conference*. IEEE, May 2015, pp. 4424 – 4429. [Online]. Available: https://doi.org/10.1109/CCDC.2015.7162653

[127] Z. Xia, X. Feng, J. Peng, and J. Fan, “Content-irrelevant tag cleansing via bi-layer clustering and peer cooperation,” *Journal of Signal Processing Systems*, vol. 81, no. 1, pp. 29 – 44, October 2015. [Online]. Available: https://doi.org/10.1007/s11265-014-0895-y

[128] T. Tsikrika, C. Diou, A. P. De Vries, and A. Delopoulos, “Image annotation using clickthrough data,” in *CIVR 2009 : proceedings of the ACM international conference on image and video retrieval, Santorini, Greece, 08-10.07.2009*. New York: ACM, October 2009. [Online]. Available: https://doi.org/10.1145/1646396.1646415

[129] C. Macdonald and I. Ounis, “Usefulness of quality click-through data for training,” in *Proceedings of Workshop on Web Search Click Data (WSCD09) : Barcelona, Spain, February 9, 2009*. New York: ACM, February 2009, p. 75–79. [Online]. Available: https://doi.org/10.1145/1507509.1507521

[130] X. He, O. King, W.-Y. Ma, M. Li, and H.-J. Zhang, “Learning a semantic space from user’s relevance feedback for image retrieval,” *Journal of Visual Languages & Computing*, vol. 13, no. 1, pp. 39–48, February 2003. [Online]. Available: https://doi.org/10.1109/TCSVT.2002.808087

[131] W. Jiang, G. Er, Q. Dai, and J. Gu, “Hidden annotation for image retrieval with long-term relevance feedback learning,” *Pattern Recognition*, vol. 38, no. 11, pp. 2007–2021, November 2005. [Online]. Available: https://doi.org/10.1016/j.patcog.2005.03.007

[132] T. Tsikrika, C. Diou, A. P. De Vries, and A. Delopoulos, “Reliability and effectiveness of clickthrough data for automatic image annotation,” *Multimedia Tools and Applications*, vol. 55, no. 1, pp. 27–52, October 2011. [Online]. Available: https://doi.org/10.1007/s11042-010-0584-1

[133] I. Sarafis, C. Diou, T. Tsikrika, and A. Delopoulos, “Weighted svm from clickthrough data for image retrieval,” in *IEEE International Conference on Image Processing (ICIP), 2014 27-30 Oct. 2014, Paris, France*. Piscataway, NJ: IEEE, October 2014. [Online]. Available: https://doi.org/10.1109/ICIP.2014.7025609

[134] I. Sarafis, C. Diou, and A. Delopoulos, “Building effective svm concept detectors from clickthrough data for large-scale image retrieval,” *International Journal of Multimedia Information*

Retrieval, vol. 4, no. 2, pp. 129 – 142, June 2015. [Online]. Available: https://doi.org/10.1007/s13735-015-0080-5

[135] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *CML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*. PMLR, June 2014, pp. 595–603. [Online]. Available: https://dl.acm.org/doi/10.5555/3044805.3044959

[136] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *29th IEEE Conference on Computer Vision and Pattern Recognition CVPR 2016 : proceedings : 26 June-1 July 2016, Las Vegas, Nevada*. Los Alamitos, California: IEEE, 2016, p. 4565–4574. [Online]. Available: https://doi.org/10.1109/CVPR.2016.494

[137] R. Socher, C. Chiung-Yu Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *ICML'11: Proceedings of the 28th International Conference on International Conference on Machine Learning*. USA: ACM, June 2011, p. 129–136. [Online]. Available: https://dl.acm.org/doi/10.5555/3104482.3104499

[138] S. Nowak and S. Röger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the 2010 ACM SIGMM International Conference on Multimedia Information Retrieval*. New York: ACM, March 2010, p. 557–566. [Online]. Available: https://dl.acm.org/doi/10.1145/1743384.1743478

[139] B. Ionescu, A. Popescu, A.-L. Radu, and H. Müller, "Result diversification in social image retrieval: a benchmarking framework," *Multimedia Tools and Applications*, vol. 75, no. 2, p. 1301–1331, 2016. [Online]. Available: https://doi.org/10.1007/s11042-014-2369-4

[140] A. A. Veloso, J. A. d. Santos, and K. Nogueira, "Learning to annotate clothes in everyday photos: Multi-modal, multi-label, multi-instance approach," in *SIBGRAPI 2014 : 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images : proceedings : Rio de Janeiro, Brazil, 27-30 August 2014*. Los Alamitos, CA: IEEE Computer Society, October 2014. [Online]. Available: https://doi.org/10.1109/SIBGRAPI.2014.37

[141] N. R. Asheghi, S. Sharoff, and K. Markert, "Crowdsourcing for web genre annotation," *Language Resources and Evaluation*, vol. 50, no. 3, p. 603–641, January 2016. [Online]. Available: https://doi.org/10.1007/s10579-015-9331-6

[142] R. Di Salvo, D. Giordano, and 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)*. Association for Computing Machinery, July 2013. [Online]. Available: https://doi.org/10.1145/2501105.2501113

[143] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, no. C, pp. 141–156, July 2015. [Online]. Available: https://doi.org/10.1016/j.neucom.2014.10.082

[144] Z. Theodosiou, O. Georgiou, and N. Tsapatsoulis, "Evaluating annotators consistency with the aid of an innovative database schema," in *2011 Sixth International Workshop on Semantic Media Adaptation and Personalization : SMAP 2011 : 1-2 December 2011, Vigo,*

Pontevedra, Spain : proceedings. Piscataway, NJ: IEEE, December 2011. [Online]. Available: https://doi.org/10.1109/SMAP.2011.22

[145] Y. Baba and H. Kashima, "Statistical quality estimation for general crowdsourcing tasks," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, January 2013, pp. 554–562. [Online]. Available: https://doi.org/10.1145/2487575.2487600

[146] Q. Li, F. Ma, J. Gao, L. Su, and C. J. Quinn, "Crowdsourcing high quality labels with a tight budget," in *WSDM'16 : proceedings of the Ninth ACM International Conference on Web Search and Data Mining : February 22-25, 2016 San Francisco, CA, USA*. New York: February, October 2016. [Online]. Available: https://doi.org/10.1145/2835776.2835797

[147] Q. Hu, Q. He, H. Huang, K. Chiew, and Z. Liu, "A formalized framework for incorporating expert labels in crowdsourcing environment," *Journal of Intelligent Information Systems*, vol. 47, no. 3, pp. 403–425, December 2016. [Online]. Available: https://doi.org/10.1007/s10844-015-0371-6

[148] D. Mitry, K. Zutis, B. Dhillon, T. Peto, S. Hayat, K.-T. Khaw, J. E. Morgan, W. Moncur, and P. J. Trucco, Emanuele Foster, "The accuracy and reliability of crowdsource annotations of digital retinal images," *Translational Vision Science and Technology*, vol. 5, no. 6, September 2016. [Online]. Available: https://doi.org/10.1167/tvst.5.5.6

[149] M. V. Giuffrida, F. Chen, H. Scharr, and S. A. Tsaftaris, "Citizen crowds and experts: observer variability in image-based plant phenotyping," *Plant Methods*, vol. 14, no. 1, 2018. [Online]. Available: https://doi.org/10.1186/s13007-018-0278-7

[150] L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel, "Can masses of non-experts train highly accurate image classifiers?: A crowdsourcing approach to instrument segmentation in laparoscopic images," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II*. Switzerland: Springer Nature, September 2014, p. 438–445. [Online]. Available: https://doi.org/10.1007/978-3-319-10470-6_55

[151] C. D. Cabralla, Z. Lu, M. Kyriakidis, L. Manca, C. Dijksterhuis, R. Happee, and J. Wintera, "Validity and reliability of naturalistic driving scene categorization judgments from crowdsourcing," *Accident Analysis & Prevention*, vol. 114, p. 25–33, 2018. [Online]. Available: https://doi.org/10.1016/j.aap.2017.08.036

[152] W. Geyser. (2021) The most popular instagram hashtags on the planet in 2021 (+ free tool). [Online]. Available: https://influencermarketinghub.com/most-popular-instagram-hashtags/

[153] C. Newberry. (2021) 2021 instagram hashtag guide: How to get more reach. [Online]. Available: https://blog.hootsuite.com/instagram-hashtags/

[154] Y. Zhang, M. Ni, W. Han, and J. Pang, "Does #like4like indeed provoke more likes?" in *Proceedings 2017 IEEE/WIC/ACM International Conference on Web Intelligence WI 2017 Leipzig, Germany 23-26 August 2017*. New York: ACM, August 2017, p. 179–186. [Online]. Available: https://doi.org/10.1145/3106426.3106460

[155] G. Armano, F. Fanni, and A. Giuliani, “Stopwords identification by means of characteristic and discriminant analysis,” in *Proceedings Of The 7th International Conference on Agents Artificial Intelligence*. Portugal: SCITEPRESS, January 2015, p. 353–360. [Online]. Available: https://doi.org/10.5220/0005194303530360

[156] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “Nus-wide: A real-world web image database from national university of singapore,” in *Proceeding of the ACM International Conference on Image and Video Retrieval ACM-CIVR 2009*. IEEE, July 2009, pp. 368 – 375. [Online]. Available: https://doi.org/10.1145/1646396.1646452

[157] J. Fan, Y. Shen, N. Zhou, and Y. Gao, “Harvesting large-scale weakly-tagged image databases from the web,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, June 2010, pp. 802 – 809. [Online]. Available: https://doi.org/10.1109/CVPR.2010.5540135

[158] N. Drewe. The hilarious list of hashtags instagram won’t let you search. [Online]. Available: http://thedatapack.com/banned-instagram-hashtags-update/#more-171

[159] S. Sedhai and A. Sun, “Hspam14: A collection of 14 million tweets for hashtag-oriented spam research,” in *SIGIR 2015: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. USA: ACM, August 2015, p. 223–232. [Online]. Available: https://doi.org/10.1145/2766462.2767701

[160] H.-C. Yang and C.-H. Lee, “Identifying spam tags by mining tag semantics,” in *ICMIA 2011 : proceedings : the 3rd International Conference on Data Mining and Intelligent Information Technology Applications : Macao, October 24-26, 2011*. Piscataway, N.J.: IEEE, October 2011, p. 263–268. [Online]. Available: https://ieeexplore.ieee.org/document/6108441

[161] R. Tang, Z. Jie, K. Xu, J. Zheng, and Y. Wang, “An intelligent semantic-based tag cleaner for folksonomies,” in *International Conference on Intelligent Computing and Integrated Systems*. Piscataway, NJ: IEEE, 2010. [Online]. Available: https://doi.org/10.1109/ICISS.2010.5657118

[162] G. Zhu, S. Yan, and Y. Ma, “Image tag refinement towards low-rank, content-tag prior and error sparsity,” in *MM’10:Proceedings of the ACM Multimedia 2010 International Conference*. New York: ACM, October 2010, p. 461–470. [Online]. Available: https://doi.org/10.1145/1873951.1874028

[163] J. C. Miller, G. Rae, F. Schaefer, L. A. Ward, and T. LoFaro, “Modifications of kleinberg’s hits algorithm using matrix exponentiation and web log records,” in *Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, September 9-13, 2001*. New York: ACM, September 2001, p. 444–445. [Online]. Available: https://doi.org/10.1145/383952.384086

[164] “Bipartite graph,” in *Encyclopedia of Operations Research and Management Science*, S. I. Gass and M. C. Fu, Eds. Boston, MA: Springer, 2013. [Online]. Available: https://doi.org/10.1007/978-1-4419-1153-7_200987

[165] M. Gao, L. Chen, B. Li, Y. Li, W. Liu, and Y.-C. Xu, “Projection-based link prediction in a bipartite network,” *Information Sciences*, vol. 376, pp. 158–171, January 2017. [Online].

Available: https://doi.org/10.1016/j.ins.2016.10.015

[166] N. Tsapatsoulis and O. Georgiou, “Investigating the scalability of algorithms, the role of similarity metric and the list of suggested items construction scheme in recommender systems,” *International Journal on Artificial Intelligence Tools*, vol. 21, no. 4, August 2012. [Online]. Available: https://doi.org/10.1142/S0218213012400180

[167] J. Mao, K. Lu, G. Li, and M. Yi, “Profiling users with tag networks in diffusion-based personalized recommendation,” *Journal of Information Science*, vol. 42, no. 5, p. 711–722, 2016. [Online]. Available: https://doi.org/10.1177/0165551515603321

[168] J. Wang, J. Zhou, H. Xu, T. Mei, X.-S. Hua, and S. Li, “Extracting and ranking product features in opinion documents,” in *COLING ’10: Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Beijing, China: Chinese Information Processing Society of China, August 2010, p. 1462–1470. [Online]. Available: https://dl.acm.org/doi/10.5555/1944566.1944733

[169] T. Tri Nguyen and J. J. Jung, “Exploiting geotagged resources to spatial ranking by extending hits algorithm,” *Computer Science and Information Systems*, vol. 12, no. 1, pp. 185 – 201, 2015. [Online]. Available: https://doi.org/10.2298/CSIS141015091T

[170] H. Cui, Q. Li, H. Li, and Z. Yan, “Healthcare fraud detection based on trustworthiness of doctors,” in *IEEE TrustCom/BigDataSE/ISPA 2016 proceedings : 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications : 10th IEEE International Conference on Big Data Science and Engineering : 14th IEEE International Symposium on Parallel and Distributed Processing with Applications : 23-26 August 2016, Tianjin, China*. IEEE, August 2016, pp. 74 – 81. [Online]. Available: https://doi.org/10.1109/TrustCom.2016.0048

[171] A. London and T. Csendes, “Hits based network algorithm for evaluating the professional skills of wine tasters,” in *8th International Symposium on Applied Computational Intelligence and Informatics*. Los Alamitos, California: IEEE, May 2013, p. 197–200. [Online]. Available: https://doi.org/10.1109/SACI.2013.6608966

[172] V. S. Tseng, J.-C. Ying, C.-W. Huang, Y. Kao, and K.-T. Chen, “Fraudetector: A graph-mining-based framework for fraudulent phone call detection,” in *KDD ’15 : proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining : August 10-13, 2015, Sydney, NSW, Australia*. New York, USA: ACM, September 2015, p. 2157–2166. [Online]. Available: https://doi.org/10.1145/2783258.2788623

[173] T. Sunahase, Y. Baba, and H. Kashima, “Pairwise hits: Quality estimation from pairwise comparisons in creator - evaluator crowdsourcing process,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. Palo Alto, California: AAAI press, February 2017, pp. 977 – 983. [Online]. Available: https://www.aaai.org/Conferences/AAAI/2017/PreliminaryPapers/08-Sunahase-14353.pdf

[174] D. Schall, B. Satzger, and H. Psaier, “Crowdsourcing tasks to social networks in bpel4people,” *World Wide Web*, vol. 17, no. 1, pp. 1 – 32, January 2014. [Online]. Available: https://doi.org/10.1007/s11280-012-0180-6

[175] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas, "Crowdsourcing for multiple-choice question answering," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. New York: ACM, July 2014, p. 2946–2953. [Online]. Available: https://cse.buffalo.edu/~demirbas/publications/wwtbam.pdf

[176] M. Gupta, R. Li, Z. Yin, and J. Han, "Survey on social tagging techniques," *SIGKDD Explorations*, vol. 12, no. 1, pp. 58 – 72, June 2010. [Online]. Available: https://www.kdd.org/exploration_files/v12-1-p58-gupta-sigkdd.pdf

[177] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Folkrank: A ranking algorithm for folksonomies," in *University Of Hildesheim, Institute Of Computer Science*, 2006, p. 111–114. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.5271

[178] J. Wang, J. Zhou, H. Xu, T. Mei, X.-S. Hua, and S. Li, "Image tag refinement by regularized latent dirichlet allocation," *Computer Vision and Image Understanding*, vol. 124, p. 61–70, July 2014. [Online]. Available: https://doi.org/10.1016/j.cviu.2014.02.011

[179] F. Pérez, R. Lapeña, A. C. Marcén, and C. Cetina, "Topic modeling for feature location in software models: Studying both code generation and interpreted models," *Information and Software Technology*, vol. 140, December 2021. [Online]. Available: https://doi.org/10.1016/j.infsof.2021.106676

[180] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012. [Online]. Available: https://doi.org/10.1145/2133806.2133826

[181] V. A. Rohani, S. Shayaa, and G. Babanejaddehaki, "Topic modeling for social media content: A practical approach," in *2016 3rd International Conference On Computer And Information Sciences (ICCOINS)*. IEEE, August 2016, p. 397–402. [Online]. Available: https://doi.org/10.1109/ICCOINS.2016.7783248

[182] S. Liu and P. Jansson, "City event identification from instagram data using word embedding and topic model visualization," *Arcada Working Papers*, vol. 7, 2017. [Online]. Available: https://www.theseus.fi/handle/10024/140582

[183] ——, "Topic modelling analysis of instagram data for the greater helsinki region," *Arcada Working Papers*, vol. 3, 2017. [Online]. Available: https://www.theseus.fi/handle/10024/140608

[184] A. Fiallos, K. Jimenes, C. Fiallos, and S. Figueroa, "Detecting topics and locations on instagram photos," in *International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, April 2018, pp. 246–250. [Online]. Available: https://doi.org/10.1109/ICEDEG.2018.8372314

[185] L. Manikonda, V. V. Meduri, and S. Kambhampati, "Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media," in *Tenth International AAAI Conference on Web and Social Media*. Palo Alto, California: The AAAI Press, May 2016, p. 639–642. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13166/12817

[186] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval : 33rd European Conference on IR Researh, ECIR 2011, Dublin, Ireland, April 18-21, 2011, Proceedings*. Berlin,

Heidelberg: Springer, April 2011, pp. 338–349. [Online]. Available: https://doi.org/10.1007/978-3-642-20161-5_34

[187] S. A. Alkhodair, B. C. M. Fung, O. Rahman, and P. C. K. Hung, "Improving interpretations of topic modeling in microblogs," *International Journal of Computer and Information Technology*, vol. 69, no. 4, pp. 528–540, April 2018. [Online]. Available: https://doi.org/10.1002/asi.23980

[188] A. Fang, "Analysing political events on twitter: topic modelling and user community classification," *ACM SIGIR Forum*, vol. 53, no. 1, p. 38–39, 2019. [Online]. Available: https://doi.org/10.1145/3458537.3458542

[189] I. Uglanova and E. Gius, "The order of things. a study on topic modelling of literary texts," in *Proceedings of the Workshop on Computational Humanities Research (CHR 2020) Amsterdam, the Netherlands, November 18-20, 2020.* CEUR Workshop Proceedings, November 2020. [Online]. Available: http://ceur-ws.org/Vol-2723/long7.pdf

[190] N. Shahid, M. U. Ilyas, J. S. Alowibdi, and N. R. Aljohani, "Word cloud segmentation for simplified exploration of trending topics on twitter," *IET Software*, vol. 11, no. 5, p. 214 – 220, October 2017. [Online]. Available: https://doi.org/10.1049/iet-sen.2016.0307

[191] R. Atenstaedt, "Word cloud analysis of the bjgp: 5 years on," *British Journal of General Practice*, vol. 67, no. 658, pp. 231–232, May 2017. [Online]. Available: https://doi.org/10.3399/bjgp17X690833

[192] S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl, "Concentricloud: Word cloud visualization for multiple text documents," in *Information visualisation: computer graphics, imaging and visualisation biomedical visualization, visualisation in built and rural environments & geometric modelling and imaging : IV 2015 : 22-24 July 2015, Barcelona, Spain : proceedings*. Piscataway, NJ: IEEE, July 2015, p. 114–120. [Online]. Available: https://doi.org/10.1109/iV.2015.30

[193] X. Fu, T. Wang, J. Li, C. Yu, and W. Liu, "Improving distributed word representation and topic model by word-topic mixture model," in *Asian Conference on Machine Learning, 16-18 November 2016, The University of Waikato, Hamilton, New Zealand.* MLResearchPress, November 2016, pp. 190–205. [Online]. Available: http://proceedings.mlr.press/v63/Fu60.html

[194] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic arabic text summarization using document clustering and topic modeling," *Journal of Physics: Conference Series*, vol. 172, June 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.114652

[195] N. Tsapatsoulis, "Image retrieval via topic modelling of instagram hashtags," in *Proceedings of the 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP).* Piscataway, NJ: IEEE, October 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/SMAP49528.2020.9248465

[196] X. Jin, "Understanding social-mediated disaster and risk communication with topic model," in *Integrated Research on Disaster Risks: Contributions from the IRDR Young Scientists Programme*,

R. Djalante, M. B. F. Bisri, and R. Shaw, Eds. Cham: Springer, 2021, ch. 9, pp. 159–174. [Online]. Available: https://doi.org/10.1007/978-3-030-55563-4_19

[197] J. A. E. Nogra, "Text analysis on instagram comments to better target users with product advertisements," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.3, pp. 175 – 181, 2020. [Online]. Available: https://doi.org/10.30534/ijatcse/2020/2691.32020

[198] A. R. Marcon, M. Bieber, and M. B. Azad, "Protecting, promoting, and supporting breastfeeding on instagram," *Maternal & Child Nutrition*, vol. 15, no. 1, January 2019. [Online]. Available: https://doi.org/10.1111/mcn.12658

[199] P. Vitale, A. Mancuso, and M. Falco, "Museums' tales: Visualizing instagram users' experience," in *Advances on P2P, Parallel, Grid, Cloud and Internet Computing : Proceedings of the 14th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2019)*. Cham: Springer, November 2020, pp. 234–245. [Online]. Available: https://doi.org/10.1007/978-3-030-33509-0_21

[200] V. Mittal, A. Kaul, S. S. Gupta, and A. Arora, "Multivariate features based instagram post analysis to enrich user experience," *Procedia Computer Science*, vol. 122, pp. 138–145, September 2017. [Online]. Available: https://doi.org/10.1016/j.procs.2017.11.352

[201] P. I. Kamil, A. H. Pratama, and A. Hidayatulloh, "Did we really #prayfornepal? instagram posts as a massive digital funeral in nepal earthquake aftermath," in *The 5th International Symposium on Earthhazard and Disaster Mitigation : the Annual Symposium on Earthquake and Related Geohazard Research for Disaster Risk Reduction : Bandung, Indonesia, 19-20 October 2015*. Melville, New York: AIP Publishing, October 2016, pp. 090 002–1–090 002–10. [Online]. Available: https://doi.org/10.1063/1.4947419

[202] M. Habibi, A. Priadana, A. B. Saputra, and P. W. Cahyo, "Topic modelling of germas related content on instagram using latent dirichlet allocation (lda)," in *Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020)*. Atlantis Press, 2021, pp. 260–264. [Online]. Available: https://doi.org/10.2991/ahsr.k.210127.060

[203] A. Latif, A. Rasheed, U. Sajid, J. Ahmed, N. Ali, N. I. Ratyal, B. Zafar, S. H. Dar, M. Sajid, and T. Khalil, "Content-based image retrieval and feature extraction: A comprehensive review," *Mathematical Problems in Engineering*, vol. 2019, 2019. [Online]. Available: https://doi.org/10.1155/2019/9658350

[204] A. Nazir, R. Ashraf, T. Hamdani, and N. Ali, "Content based image retrieval system by using hsv color histogram, discrete wavelet transform and edge histogram descriptor," in *2018 International Conference on Computing, Mathematics and Engineering Technologies: iCoMET 2018: Invent, Innovate and Integrate for Socioeconomic Development: March 3 - 4, 2018: Conference Proceedings*. Piscataway, New Jersey: IEEE, March 2018. [Online]. Available: https://doi.org/10.1109/ICOMET.2018.8346343

[205] S. Sergyán, "Color histogram features based image classification in content-based image retrieval systems," in *6th International Symposium on Applied Machine Intelligence and Informatics*.

Piscataway, NJ: IEEE, January 2008, pp. 221–224. [Online]. Available: https://doi.org/10.1109/SAMI.2008.4469170

[206] M. Takeishi, D. Oguro, H. Kikuchi, and J. Shin, "Histogram-based image retrieval keyed by normalized hsy histograms and its experiments on a pilot dataset," in *IEEE International Conference on Consumer Electronics - Asia*. Piscataway, NJ: IEEE, June 2018. [Online]. Available: https://doi.org/10.1109/ICCE-ASIA.2018.8552118

[207] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognition*, vol. 46, no. 1, pp. 188–198, January 2013. [Online]. Available: https://doi.org/10.1016/j.patcog.2012.06.001

[208] H. Zhang, M. Jiang, and Q. Kou, "Color image retrieval algorithm fusing color and principal curvatures information," *IEEE Access*, vol. 8, pp. 184 945 – 184 954, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3030056

[209] F. A. Mufarroha and A. G. Anamisa, Devie Rosa Hapsani, "Content based image retrieval using two color feature extraction," *Journal of Physics: Conference Series*, vol. 1569, no. 3, July 2020. [Online]. Available: https://doi.org/10.1088/1742-6596/1569/3/032072

[210] W. H. Alwan, E. Fazl-Ersi, and A. Vahedian, "Identifying influential users on instagram through visual content analysis," *IEEE Access*, vol. 8, pp. 169 594–169 603, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3020560

[211] E. Deza and M. M. Deza, "Image distances," in *Encyclopedia of Distances*. Dordrecht: Springer-Verlag, 2009, p. 349–362. [Online]. Available: https://link.springer.com/book/10.1007%2F978-3-642-00234-2

[212] C. Li and S. Qian, "Measuring image similarity based on shape context," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 3, pp. 127–134, March 2015. [Online]. Available: https://doi.org/10.14257/ijmue.2015.10.3.13

[213] K. Arai, "Image retrieval method based on back-projection," in *Advances in Computer Vision : Proceedings of the 2019 Computer Vision Conference*. Cham, Switzerland: Springer Nature, April 2019, pp. 689–698. [Online]. Available: https://doi.org/10.1007/978-3-030-17798-0_54

[214] M. G. Forero, C. Arias-Rubio, and B. T. González, "Analytical comparison of histogram distance measures," in *23rd Iberoamerican Congress*. Switzerland: Springer Nature, November 2019, pp. 81–90. [Online]. Available: https://doi.org/10.1007/978-3-030-13469-3_10

[215] M. S. Prieto and A. R. Allen, "A similarity metric for edge images," *A Similarity Metric for Edge Images*, vol. 25, no. 10, pp. 1265 – 1273, October 2003. [Online]. Available: https://doi.org/10.1109/TPAMI.2003.1233900

[216] R. Chacon-Quesada and F. Siles-Canales, "Evaluation of different histogram distances for temporal segmentation in digital videos of football matches from tv broadcast," in *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*. IEEE, July 2017. [Online]. Available: https://doi.org/10.1109/IWOBI.2017.7985543

[217] K. M. Ong, P. Ong, C. K. Sia, and E. S. Low, "Effective moving object tracking using modified flower pollination algorithm for visible image sequences under complicated

background," *Applied Soft Computing Journal*, vol. 83, October 2019. [Online]. Available: https://doi.org/10.1016/j.asoc.2019.105625

[218] H. Abidi, M. Chtourou, K. Kaaniche, and H. Mekki, "Visual servoing based on efficient histogram information," *International Journal of Control, Automation and Systems*, vol. 15, no. 4, pp. 1746–1753, August 2017. [Online]. Available: https://doi.org/10.1007/s12555-016-0070-2

[219] A. Doulah and E. Sazonov, "Clustering of food intake images into food and non-food categories," in *Bioinformatics and Biomedical Engineering : 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26-28, 2017, Proceedings*. Springer Nature, April 2010, pp. 454–463. [Online]. Available: https://doi.org/10.1007/978-3-319-56148-6_40

[220] N. Tsapatsoulis and C. Djouvas, "Opinion mining from social media short texts: Does collective intelligence beat deep learning?" *Frontiers in Robotics and AI*, vol. 5, no. 138, January 2019. [Online]. Available: https://doi.org/10.3389/frobt.2018.00138

[221] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145 614 – 145 633, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2945911

[222] D. Ganguly, "Learning variable-length representation of words," *Pattern Recognition*, vol. 102, July 2020. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107306

[223] R. Gomez, L. Gomez, J. Gibert, and D. Karatzas, "Learning from #barcelona instagram data what locals and tourists post about its neighbourhoods," in *Computer Vision – ECCV 2018 Workshops*. Switzerland: Springer Nature, September 2018, pp. 530–544. [Online]. Available: https://doi.org/10.1007/978-3-030-11024-6_41

[224] K. Jiang, S. Feng, R. A. Calix, and G. R. Bernard, "Assessment of word embedding techniques for identification of personal experience tweets pertaining to medication uses," in *Precision Health and Medicine: A Digital Revolution in Healthcare*, A. Shaban-Nejad and M. Michalowski, Eds. Switzerland: Springer Nature, 2020, pp. 45–55. [Online]. Available: https://doi.org/10.1007/978-3-030-24409-5_5

[225] H. Zhang, "Dynamic word embedding for news analysis," Master's thesis, University of California, USA, 2019. [Online]. Available: https://www.proquest.com/docview/2248657601/FFB0B0D8CAA74F59PQ/1?accountid=36246

[226] J. Weston, S. Chopra, and K. Adams, "#tagspace: Semantic embeddings from hashtags," in *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, October 2014, p. 1822–1827. [Online]. Available: https://www.aclweb.org/anthology/D14-1194

[227] K. Hammar, S. Jaradat, N. Dokoohaki, and M. Matskin, "Deep text mining of instagram data without strong supervision," in *18th IEEE/WIC/ACM International Conference on Web Intelligence*. Los Alamitos, California: IEEE, December 2018, pp. 158–165. [Online]. Available: https://doi.org/10.1109/WI.2018.00-94

[228] F. Prabowo and A. Purwarianti, "Instagram online shop's comment classification using statistical approach," in *2nd International conferences on Information Technology, Information Systems and*

Electrical Engineering. Piscataway, NJ: IEEE, November 2017, pp. 282–287. [Online]. Available: https://doi.org/10.1109/ICITISEE.2017.8285512

[229] A. Akbar Septiandri and O. Wibisono, "Detecting spam comments on indonesia's instagram posts," *Journal of Physics: Conference Series*, vol. 801, no. 1, March 2017. [Online]. Available: https://doi.org/10.1088/1742-6596/801/1/012069

[230] D. Serafimov, M. Mirchev, and I. Mishkovski, "Friendship paradox and hashtag embedding in the instagram social network," in *11th International Conference, ICT Innovations*. Switzerland: Springer Nature, October 2019, pp. 121–133. [Online]. Available: https://doi.org/10.1007/978-3-030-33110-8_11

[231] Y. Xu, H. Nguyen, and Y. Li, "A semantic based approach for topic evaluation in information filtering," *IEEE Access*, vol. 8, pp. 66 977–66 988, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2985079

[232] J. Cao, A. Zhao, and Z. Zhang, "Automatic image annotation method based on a convolutional neural network with threshold optimization," *PLOS ONE*, vol. 15, no. 9, September 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0238956

[233] Abdullah and M. S. Hasan, "An application of pre-trained cnn for image classification," in *20th International Conference on Computer and Information Technology (ICCIT)*. IEEE, December 2017. [Online]. Available: https://doi.org/10.1109/ICCITECHN.2017.8281779

[234] S. Chaib, H. Yao, Y. Gu, and M. Amrani, "Deep feature extraction and combination for remote sensing image classification based on pre-trained cnn models," in *Ninth International Conference on Digital Image Processing (ICDIP 2017)*, C. M. Falco and X. Jiang, Eds. Bellingham, Washington: SPIE, May 2017. [Online]. Available: https://doi.org/10.1117/12.2281755

[235] Y. Shima, "Image augmentation for object image classification based on combination of pre-trained cnn and svm," *Journal of Physics: Conference Series*, vol. 1004, no. 1, April 2018. [Online]. Available: https://doi.org/10.1088/1742-6596/1004/1/012001

[236] A. Nasiri, A. Taheri-Garavand, D. Fanourakis, Y.-D. Zhang, and N. Nikoloudakis, "Automated grapevine cultivar identification via leaf imaging and deep convolutional neural networks: A proof-of-concept study employing primary iranian varieties," *Plants*, vol. 10, no. 8, August 2021. [Online]. Available: https://doi.org/10.3390/plants10081628

[237] A. Taheri-Garavand, A. Nasiri, D. Fanourakis, S. Fatahi, M. Omid, and N. Nikoloudakis, "Automated in situ seed variety identification via deep learning: A case study in chickpea," *Plants*, vol. 10, no. 7, July 2021. [Online]. Available: https://doi.org/10.3390/plants10071406

[238] N. Tsapatsoulis and K. Diakoumopoulou, "Face verification in practice: The case of greek artist leonidas arniotis," in *Proceedings of the 19th International Conference on Computer Analysis of Images and Patterns (CAIP 2021)*. Springer, October 2021, pp. 1–6.

[239] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden. Social media update 2014. [Online]. Available: https://www.pewresearch.org/internet/2015/01/09/social-media-update-2014/

[240] W. Hersh, “Terms, models, resources, and evaluation,” in *Information Retrieval*. New York: Springer, 2009, pp. 3–39. [Online]. Available: https://doi.org/10.1007/978-0-387-78703-9_1

[241] A. Kulshrestha, “On the hamming distance between base-n representations of whole numbers,” *Canadian Young Scientist Journal*, vol. 2, p. 14–16, 2012. [Online]. Available: http://journal.fsst.ca/index.php/jsst/issue/viewIssue/20/pdf_15

[242] M. Bramer, *Principles of Data Mining*. London: Springer Nature, 2007.

[243] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM*, vol. 46, no. 5, p. 604–632, 1999. [Online]. Available: https://doi.org/10.1145/324133.324140

[244] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, “The web as a graph: Measurements, models, and methods,” in *Computing and Combinatorics 5th Annual International Conference, COCOON’99 Tokyo, Japan, July 26-28, 1999 Proceedings*. Berlin Heidelberg: Springer-Verlag, July 1999. [Online]. Available: https://doi.org/10.1007/3-540-48686-0_1

[245] P. Tsaparas, “Link analysis ranking,” Ph.D. dissertation, University of Toronto, Torondo, Canada, 2004. [Online]. Available: https://www.cs.uoi.gr/~tsap/publications/PhD.Thesis.pdf

[246] I. Nagasinghe, “Computing principal eigenvectors of large web graphs: Algorithms and accelerations related to pagerank and hits,” Ph.D. dissertation, Southern Methodist University, Dallas, USA, 2010. [Online]. Available: https://www.proquest.com/docview/506682643

[247] J. M. Fletcher and T. Wennekers, “From structure to activity: Using centrality measures to predict neuronal activity,” *International Journal of Neural Systems*, vol. 28, no. 2, p. 16 – 19, March 2018. [Online]. Available: https://doi.org/10.1142/S0129065717500137

[248] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, “Trend detection in folksonomies,” in *Semantic multimedia : First International Conference on Semantic and Digital Media Technologies, SAMT 2006, Athens, Greece, December 6-8, 2006 ; proceedings*. Berlin, Heidelberg: Springer, December 2006, pp. 56–70. [Online]. Available: https://doi.org/10.1007/11930334_5

[249] N. Craswell, “Mean reciprocal rank,” in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer, 2009, p. 1703–1703. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_488

[250] Q. Chen, L. Yao, and J. Yang, “Short text classification based on lda topic model,” in *2016 International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE, July 2016, pp. 749 – 753. [Online]. Available: https://doi.org/10.1109/ICALIP.2016.7846525

[251] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic evaluation of topic coherence,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics Proceedings of the Main Conference: June 2–4, 2010 Los Angeles, California*. New York: ACM, June 2010, p. 100–108. [Online]. Available: https://dl.acm.org/doi/10.5555/1857999.1858011

[252] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *WSDM’15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. NY, USA: ACM, February 2015, p. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324

[253] D. Newman, S. Karimi, and L. Cavedon, “External evaluation of topic models,” in *Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009*, December 2009, p. 11–18. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.529.7854

[254] A. Fan, F. Doshi□Velez, and L. Miratrix, “Assessing topic model relevance: Evaluation and informative priors,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 12, no. 3, pp. 210–222, 2019. [Online]. Available: https://doi.org/10.1002/sam.11415

[255] A. Fang, C. Macdonald, I. Ounis, and P. Habel, “Topics in tweets: A user study of topic coherence metrics for twitter data,” in *Advances in information retrieval 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016 : proceedings*. Cham: Springer, March 2016, pp. 492–504. [Online]. Available: https://doi.org/10.1007/978-3-319-30671-1_36

[256] N. Niraula, R. Banjade, D. Ştefănescu, and V. Rus, “Experiments with semantic similarity measures based on lda and lsa,” in *Statistical Language and Speech Processing First International Conference, SLSP 2013, Tarragona, Spain, July 29-31, 2013. Proceedings*, A.-H. Dediu, C. Martín-Vide, R. Mitkov, and B. Truthe, Eds. Berlin, Heidelberg: Springer, July 2013, p. 188–199. [Online]. Available: https://doi.org/10.1007/978-3-642-39593-2_17

[257] A. Fang, C. Macdonald, I. Ounis, and P. Habel, “Using word embedding to evaluate the coherence of topics from twitter data,” in *SIGIR’16 : the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval: Pisa, Italy, July 17-21, 2016*. New York, NY, USA: Association for Computing Machinery, July 2016, p. 1057–1060. [Online]. Available: https://doi.org/10.1145/2911451.2914729

[258] D. Zhang, “Statistical part-based models: Theory and applications in image similarity, object detection and region labeling,” Ph.D. dissertation, Columbia University, Columbia, USA, 2006. [Online]. Available: https://www.proquest.com/docview/305355754/F6B6C1A20D0F4E60PQ/1?accountid=36246

[259] D. Han, “Particle image segmentation based on bhattacharyya distance,” Master’s thesis, Arizona State University, USA, 2015. [Online]. Available: https://keep.lib.asu.edu/items/153947

[260] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors forword representation,” in *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, October 2014, p. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[261] S. Puglisi, J. Parra-Arnau, J. Forné, and D. Rebollo-Monedero, “On content-based recommendation and user privacy in social-tagging systems,” *Computer Standards & Interfaces*, vol. 41, pp. 17–27, September 2015. [Online]. Available: https://doi.org/10.1016/j.csi.2015.01.004

[262] Y. Zhang, X. Wang, Y. Sakai, and T. Yamasaki, “Measuring similarity between brands using followers’ post in social media,” in *MMAsia ’19: Proceedings of the ACM Multimedia Asia*. New York: Association for Computing Machinery, December 2019. [Online]. Available: https://doi.org/10.1145/3338533.3366600

[263] B. Diedenhofen and J. Musch, "cocor: A comprehensive solution for the statistical comparison of correlations," *PLoS ONE*, vol. 10, no. 4, April 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0121945

[264] N. Younes and U.-D. Reips, "Guideline for improving the reliability of google ngram studies: Evidence from religious terms," *PLoS ONE*, vol. 14, no. 3, March 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0213554

[265] D. Bhattacharjee, S. Vracar, P. G. Round, Rachel A. amd Nightingale, J. A. Williams, G. V. Gkoutos, I. M. Stratton, R. Parker, S. D. Luzio, J. Webber, S. E. Manley, and G. A. Roberts, "Utility of hba1c assessment in people with diabetes awaiting liver transplantation," *Diabetic Medicine*, vol. 36, pp. 1444 – 1452, November 2019. [Online]. Available: https://doi.org/10.1111/dme.13870

[266] R. Schreiber, V. R. Bellinazzi, A. C. Sposito, J. é. G. Mill, J. E. Krieger, A. C. Pereira, and W. Nadruz Jr, "Influence of the c242t polymorphism of the p22-phox gene (cyba) on the interaction between urinary sodium excretion and blood pressure in an urban brazilian population," *PLoS ONE*, vol. 8, no. 12, December 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0081054

[267] S. Ložnjak, T. Kramberger, I. Cesar, and R. Kramberger, "Automobile classification using transfer learning on resnet neural network architecture," *Polytechnick and Design*, vol. 8, no. 1, pp. 59–64, 2020. [Online]. Available: https://doi.org/10.19279/TVZ.PD.2020-8-1-18

[268] M. Habibzadeh, M. Jannesari, Z. Rezaei, H. Baharvand, and M. Totonchi, "Automatic white blood cell classification using pre-trained deep learning models: Resnet and inception," in *Tenth International Conference on Machine Vision*. SPIE, 2018. [Online]. Available: https://doi.org/10.1117/12.2311282

[269] A. Maskey and M. Cho, "Cubesatnet: Ultralight convolutional neural network designed for on-orbit binary image classification on a 1u cubesat," *Engineering Applications of Artificial Intelligence*, vol. 96, November 2020. [Online]. Available: https://doi.org/10.1016/j.engappai.2020.103952

[270] T. Carneiro, R. V. M. Da Nóbrega, T. G. Nepomuceno, G. Bian, V. H. C. De Albuquerque, and P. P. Rebouças Filho, "Performance analysis of google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61 677 – 61 685, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2874767

[271] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *29th IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos, California: IEEE, June 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[272] M. Jannesari, M. Habibzadeh, H. Aboulkheyr, P. Khosravi, O. Elemento, M. Totonchi, and I. Hajirasouliha, "Breast cancer histopathological image classification: A deep learning approach," in *Proceedings, 2018 IEEE International Conference on Bioinformatics and Biomedicine : 3-6 Dec. 2018, Madrid, Spain*. Los Alamitos, California: IEEE, December 2018, pp. 2405–2412. [Online]. Available: https://doi.org/10.1109/BIBM.2018.8621307

[273] X. Liu and H. Song, “Automatic identification of fossils and abiotic grains during carbonate microfacies analysis using deep convolutional neural networks,” *Sedimentary Geology*, vol. 410, December 2020. [Online]. Available: https://doi.org/10.1016/j.sedgeo.2020.105790

[274] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62 – 66, January 1979. [Online]. Available: https://doi.org/10.1109/TSMC.1979.4310076

[275] J. De Winter, M. Kyriakidis, D. Dodou, and R. Happee, “Using crowdflower to study the relationship between self-reported violations and traffic accidents,” *Procedia Manufacturing*, vol. 3, p. 2518 – 2525, 2015. [Online]. Available: https://doi.org/10.1016/j.promfg.2015.07.514

[276] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci, “An italian twitter corpus of hate speech against immigrants,” in *Proceedings 11th International Conference on Language Resources and Evaluation*. Stroudsburg, PA: Association for Computational Linguistics, 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1443

[277] V. Prabhakaran, A. Arora, and O. Rambow, “Staying on topic: An indicator of power in political debates,” in *EMNLP 2014: The 2014 Conference on Empirical Methods In Natural Language Processing: Proceedings of the Conference: October 25-29, 2014 Doha, Qatar*. Stroudsburg, PA: Association for Computational Linguistics, October 2014, p. 1481–1486. [Online]. Available: https://doi.org/10.3115/v1/D14-1157

[278] Z. Yuan, H. Liu, X. Zhang, F. Li, J. Zhao, F. Zhang, and F. Xue, “From interaction to co-association —a fisher r-to-z transformation-based simple statistic for real world genome-wide association study,” *PLoS ONE*, vol. 8, no. 7, July 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0070774

[279] K. He, G. Gkioxari, P. Dollàr, and R. Girshick, “Mask r-cnn,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386 – 397, February 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2844175

[280] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *29th IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, June 2016, pp. 779–788. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91