



HAL
open science

Topic Identification via Human Interpretation of Word Clouds: The Case of Instagram Hashtags

Stamatios Giannoulakis, Nicolas Tsapatsoulis

► **To cite this version:**

Stamatios Giannoulakis, Nicolas Tsapatsoulis. Topic Identification via Human Interpretation of Word Clouds: The Case of Instagram Hashtags. 17th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Jun 2021, Hersonissos, Crete, Greece. pp.283-294, 10.1007/978-3-030-79150-6_23 . hal-03287661

HAL Id: hal-03287661

<https://inria.hal.science/hal-03287661>

Submitted on 15 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Topic identification via human interpretation of word clouds: The case of Instagram hashtags

Stamatios Giannoulakis^[0000–0003–3020–3717] and Nicolas Tsapatsoulis

Department of Communication and Internet Studies
Cyprus University of Technology
30, Arch. Kyprianos str., CY-3036, Limassol, Cyprus
{s.giannoulakis,nicolas.tsapatsoulis}@cut.ac.cy

Abstract. Word clouds are a very useful tool for summarizing textual information. They can be used to illustrate the most frequent and important words of text documents or a set of text documents. In that respect they can also be used for topic visualisation. In this paper we present an experiment investigating how the crowd understands topics visualised via word clouds. In the experiment we use the topics mined from Instagram hashtags of a set of Instagram images corresponding to 30 different subjects. By subject we mean the research hashtag we use to gather pairs of Instagram images and hashtags. With the aid of an innovative topic modelling method, developed in a previous work, we constructed word clouds for the visualisation of each topic. Then we used a popular crowdsourcing platform (*Appen*) to let users identify the topic they believe each word cloud represents. The results show some interesting variations across subjects which are analysed and discussed in detail throughout the paper. Given that the topics were mined from Instagram hashtags, the current study provides useful insights regarding the appropriateness of hashstags as image annotation tags.

Keywords: Wordclouds · Topic modelling · Instagram hashtags · Image annotation · Visualisation.

1 Introduction

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 [2, 11]. 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 [4]. 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 [1].

Let us now assume a set of Instagram photos grouped together via a common property such as the queried hashtag that was used to collect them. In the context of the current work we name the query hashtag as *subject*. Instagram photos are frequently accompanied by hashtags [3] that the photo owner and other Instagram users use to describe photos' content and, in several cases, their feelings, moments and reactions related with those photos. We can see, therefore, the hashtags of an Instagram image as a textual representation of it and in this way Instagram collections of images can be seen as textual documents and can be analysed via topic modelling techniques once textual preprocessing, such as word splitting is applied first. 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 [19] we can derive a set of terms best describing the set of Instagram photos grouped together within a subject.

In this paper we investigate how the crowd 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 crowd is asked to guess the hidden hashtag providing their best four guesses. The aim of the current work is first to assess whether a word cloud presentation is appropriate for visualising hashtag sets and, second, to investigate any variations in interpreting word clouds corresponding to different subjects. We believe that through this meta-analysis we gain useful insights on whether words mined from Instagram hashtags [6] can be used for image tagging in a collective manner allowing for quick development training sets for Automatic Image Annotation [17, 18, 15].

2 Related Work

Word clouds is an informative data visualisation tool [16] primarily used to summarize textual information but it has been also applied for the analysis of social media data.

Jin [10] 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 [14]. 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 [12].

Vitale *et al.* [20] 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.* [13] 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.* [9] 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.* [8] collected posts related to hashtag *#germas*. They applied topic modelling 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.

The previous discussion shows 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 [19] is quite different and tailored to the specific case of Instagram posts. Photos and associated hashtags are modelled as a bipartite network and the importance of each hashtags is computed via its authority score obtained by applying the HITS algorithm [7].

token corresponding to the associated subject (query hashtag) was excluded in order to examine whether the crowd would guess it correctly (see Section 4 for the details). Word clouds visualization (see an example in Figure 1) was done with the help of WordCloud² Python library.

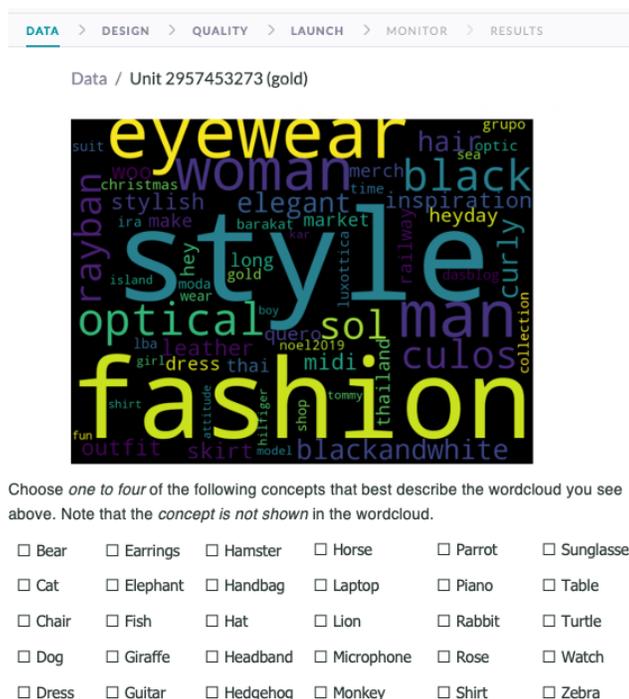


Fig. 2: The way the wordclouds were presented to the crowd through the Appen crowdsourcing platform

4 Crowd-based interpretation of word clouds

Crowd-based interpretation of word clouds was conducted with the aid of the *Appen* (<https://appen.com/>) crowdsourcing platform. The word clouds were presented to the participants as shown in Figure 2 and the participants 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.

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’

² https://amueller.github.io/word_cloud/

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 AIAI purposes [19].

Not all word clouds present the same difficulty in 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. By correct identification we mean that a contributor had selected the right subject within her/his one to four choices (see Figure 2). We see for instance in Table 1 that the accuracy of the first 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.

Table 1: Crowd-based topic identification accuracy.

Subject	Acc.(%)	Subject	Acc.(%)	Subject	Acc.(%)	Subject	Acc.(%)
Guitar	93	Dress	80	Cat	90	Chair	47
Guitar	87	Dress	60	Dog	87	Chair	43
Microphone	67	Shirt	53	Fish	100	Laptop	100
Microphone	57	Shirt	33	Fish	93	Laptop	80
Piano	70	Earrings	90	Hamster	7	Table	77
Piano	47	Handbag	93	Hamster	3	Table	73
		Hat	7	Parrot	90		
Bear	43	Hat	3	Parrot	87	Hedgehog	0
Elephant	37	Headband	30	Rabbit	77	Hedgehog	0
Giraffe	63	Headband	17	Turtle	21	Horse	87
Lion	60	Sunglasses	67	Turtle	20	Rose	63
Lion	67	Watch	87				
Monkey	33						
Zebra	57						

5 Results and discussion

The accuracy of interpretation for all word clouds is presented in Table 1. In order to better facilitate the discussion that follows the subjects (query hash-tags) were divided into six categories: (a) **Music**: Guitar, Piano, Microphone (b) **Wild animals**: Bear, Elephant, Giraffe, Lion, Monkey, Zebra (c) **Fashion**: Dress, Earrings, Handbag, Hat, Headband, Shirt, Sunglasses (d) **Office**: Chair, Laptop, Table, (e) **Pets**: Cat, Dog, Fish, Hamster, Parrot, Rabbit, Turtle (f) **Miscellaneous**: Hedgehog, Horse, Rose.

We see in Table 1 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 the correct interpretation of word clouds. Thus, textual context and key tokens in the word clouds is the second parameter affecting the difficulty of interpretation. This is, obviously, a data related factor and its effect can be minimized by creating word clouds from a larger number of Instagram posts per subject (note that in our case we have used on average 56 Instagram posts per subject). The third parameter that affects interpretation, is the familiarity of people with the concepts. Concepts such as dog, cat and horse are far more familiar to everyday people than concepts such as hedgehog and hamster.

Overall, the word clouds corresponding to the subjects ‘Laptop’ and ‘Fish’ had the highest interpretation accuracy both reaching the absolute 100%. In the case of fish word cloud (see Figure 3a) the prominent presence of the tokens fisherman, aquarium created a strong and clear context and led the annotators to select the concept fish as their single selection. That resulted in a meta word cloud (see Figure 3b) consisting of a single word, the word fish. A similar case is seen in the case of the laptop word cloud (see Figure 3c). The prominent presence of tokens apple, macbook, computer and portable make it clear to the annotators that the right concept choice is laptop. We see in the corresponding meta word cloud that a significant number of contributors chose the concept ‘table’ as well triggered mainly by the strong presence in the cloud of the token coffee (see Figure 3d).

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 4 we see two different word clouds for the subject ‘microphone’. While the word clouds are quite different (see Figures 4b and 4c) the music-singing conceptual context is prominent. Tokens like band, singer, music, hop, hip and stage create a strong and clear textual content. Thus, the annotators chose all music-singing related terms, namely guitar, piano and microphone, as it can be seen in the corresponding meta word clouds (see Figures 4b, 4d).

The monkey word cloud (see Figure 5a) 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 contributors (annotators). We see in the meta word cloud (Figure 5b), however, that the key tokens animal and nature combined with the term gorilla in the upper right corner of the word cloud led the contributors to make selections from the wild animal category including the correct subject (accuracy 33%).

4. Fu, X., Wang, T., Li, J., Yu C., Liu, W.: Improving Distributed Word Representation and Topic Model by Word-Topic Mixture Model. In: Durrant R. J., Kim K.-E.b (eds) Proceedings of the Asian Conference on Machine Learning, vol. 63, pp 190-205 (2016).
5. Giannoulakis, S., Tsapatsoulis, N.: Defining and identifying stophashtags in instagram. 2nd INNS Conference on Big Data. pp. 304-313. Springer Nature, Cham, Switzerland (2016).
6. Giannoulakis, S., Tsapatsoulis, N.: Evaluating the descriptive power of Instagram hashtags. *Journal of Innovation in Digital Ecosystems*, **3**(2), pp. 114-129 (2016).
7. Giannoulakis, S., Tsapatsoulis, N.: Filtering Instagram hashtags through crowdtagging and the HITS algorithm. *IEEE Transactions on Computational Social Systems*, **6**(3), pp. 592-603 (2019).
8. Habibi, M., Priadana, A., Saputra, A., Cahyo, P.: Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA). 2nd International Conference of Health, pp. 260-264. Atlantis Press (2020).
9. Kamil, P., Pratama, A., Hidayatulloh, A.: Did we really #prayfornepal? Instagram posts as a massive digital funeral in Nepal earthquake aftermath. *AIP Conference Proceedings*, **1730**, 090002-1-090002-10 (2016).
10. Jin, X.: Understanding Social-Mediated Disaster and Risk Communication with Topic Model. pp. 159-174. Springer Nature, Switzerland (2021).
11. Lohmann, s., Heimerl, F., Bopp, F., Burch, M., Ertl, T.: ConcentriCloud: Word Cloud Visualization for Multiple Text Documents. In: Banissi E. et al. (eds) Proceedings of the 19th International Conference on Information Visualisation, pp. 114-120. IEEE, Piscataway, NJ (2015).
12. Marcon, A., Bieber, M., Azad, M.: Protecting, promoting, and supporting breastfeeding on Instagram. *Maternal & Child Nutrition*, **15**(1), e12658 (2019).
13. Mittal, V., Kaul, A., Gupta, S., Arora, A.: Multivariate Features Based Instagram Post Analysis to Enrich User Experience. *Procedia Computer Science*, **122**, pp. 138-145 (2017).
14. Nogra, J.A.E.: Text Analysis on Instagram Comments to Better Target Users with Product Advertisements. *International Journal of Advanced Trends in Computer Science and Engineering*, **9**, pp. 175-181 (2020).
15. Ntalianis, K, Tsapatsoulis, N, Doulamis, A., Matsatsinis, N.: Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution. *Multimedia Tools and Applications*, **69**(2), pp.397-421 (2014).
16. Shahid, N., Ilyas, M., Alowibdi, J., Aljohani, N.: Word cloud segmentation for simplified exploration of trending topics on Twitter. *IET Software*. **11**, pp. 214-220 (2017).
17. Theodosiou, Z, Tsapatsoulis, N: Image Retrieval Using Keywords: The Machine Learning Perspective. *Semantic Multimedia Analysis and Processing* (Eds: E. Spyrou, D. Iakovides, P. Mylonas). CRC Press / Taylor & Francis, pp. 3-30 (2014).
18. Tsapatsoulis, N.: Web Image Indexing using WICE and a learning-free Language Model. 12th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI 2016), pp. 131-140 (2016).
19. Tsapatsoulis, N.: Image retrieval via topic modelling of Instagram hashtags. 15th International Workshop on Semantic and Social Media Adaptation & Personalization, pp.1-6. IEEE, Piscataway, NJ (2020).
20. Vitale, P., Mancuso, A., Falco, M.: Museums' Tales: Visualizing Instagram Users' Experience. 14th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 234-245. Springer Nature, Switzerland (2020).