A NOVEL APPROACH OF FILTERING HASH INSTAGRAM TAGS USING HITS ALGORITHM

BOLLA SANTHI GANESH [#1], K.RAMBABU [#2]

[#1] MCA Student, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

[#2] Head & Assistant Professor, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

Abstract

In this proposed work we try to discuss about analyzing or filtering Instagram hashtags given by crowds to detect whether hash tag is correct or not which is given by crowds. To identify correctness of tags we are using HITS algorithm. Now-a-days online social network users are posting messages with related pictures and the hash tags will be assigning to that picture. This related hash tags make other users to search that image easily. Sometime some user's assigns unrelated hash tags to images which make searching process difficult. To overcome from this issue author has introduce hash tags filtering technique using which we will filter hash tags to determine whether hash tag is relevant or irrelevant by matching content of both main hash tag and the annotator hash tags.Using HITS algorithm we can determine whether that hash tags is used more frequently or not, if it's less frequent or unrelated hash tag then we will consider as stop hash tag.

1. INTRODUCTION

Social media are online communication channels dedicated to community-based input, interaction, content sharing and collaboration. These media give the users the opportunity to share their content such as, text, video and images [31]. Users usually accompany the content they post with text such as comments or hashtags. That alternative text (comment, hashtags etc.) provide valuable information about the users posts and other information. Preece et al. [32] to construct a Sentinel platform that can enhance social media data in order to understand different situations they based also in Youtube video comments. Sagduyu et al. [33] present a novel system that can present large-scale synthetic data from social media. The users in several of those media, e.g. Twitter, Instagram and Facebook, use hashtags to annotate the digital content they upload. Hahshtags are, usually, words or nonspaced phrases preceded by the symbol # that allow creators / content contributors to apply tagging that makes it easier for other users to locate their posts.

A great portion of the digital content shared on social media platforms consists of images and short videos. Thus, effective retrieval of images from social media and the web in general, becomes harder and more challenging day by day. Contemporary search engines are basically based on text descriptions to retrieve images, however, inaccurate text descriptions and the plethora of non-textually annotated images, led to extended research for content-based image retrieval techniques [23]. The main problem of content-based image retrieval is the socalled semantic gap [30, 35, 37, 42]: Content-based retrieval is associated with low-level features while humans use highlevel concepts for their search. To overcome this problem, Automatic Image Annotation (AIA) methods were developed, that is, processes by which computing systems automatically assign metadata in the form of captions or keywords to images [4].

Among the AIA methods those based on the learning by example paradigm are probably the most common [21]. A small set of manually annotated training images are used to train models that learn the correlation between image features and textual words (high level concepts) and then, allow automatic annotation of other (unseen) images. Obviously, good training examples, i.e., representative and accurate pairs of images and related tags are vital in this case [38]. Social media, and especially the Instagram, provide a rich source of image - tag pairs [8, 12]. Mining the right ones, automatically or semi-automatically, so as to be used as training examples is extremely important. We have to consider, however, that, in many cases, hashtags that accompany images in social media are not related with the image's content but serve several other purposes such as the expression of user's emotional state, the increase of user's clicks and find ability, and the beginning of a new communication or discussion [7].

In our previous research we have shown that the percentage of the Instagram hashtags that describe the visual content of the image they are associated with, does not exceed 25% [12]. We have also noticed that many Instagram hashtags are used across images that have nothing in common, just for searchability enhancement. We named those hashtags as stophashtags [13]. Thus, filtering the Instagram hashtags in terms of the visual content of the image they accompany is required. HITS is a ranking algorithm than we could use to filter Instagram hashtags and locate the most relevant. 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 [27]. We have started experimenting with the HITS algorithm for mining informative Instangram hashtags in one of our previous works [14] and we extend this study here by considering the application of HITS algorithm in a real crowdtagging environment facilitated by the Figure-eight, formerly known as Crowdflower, crowdsourcing platform. In addition, we have increased the number of annotations per image to 500, we formed the bipartite graphs for all images and we calculated the performance of annotators across all those images. Moreover, FolkRank is used as baseline to evaluate the performance of the proposed method.

2. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language used for developing the tool. Once the programmers start building the tool, the programmers need lot of external support. This support obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into for developing the proposed system.

Zhang et al. [47] 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 [40] 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 geo-tagged 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. [6] 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 [22] 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. [44] 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.

Sunahase et al. [36] 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.

Aydin et al. [2] 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.

3. EXISTING SYSTEM

To the best of our knowledge, there is no concept like finding hash tag relevant or not in the existing system .Hence the users cannot able to filter out the information in a proper manner from the social networks. Also there is no concept like clustering the tweets and their re-tweets as per the several social influence categories like: Positive, Negative and Neutral.

LIMITATION OF EXISTING SYSTEM

The following are the limitation of existing system. They is as follows:

1. There is no Information filtering for social computing.
2. There is no automatic approach to display hash tag based on filter approach.
3. There is no single method to gather all the individual users interest into one location.

4. PROPOSED SYSTEM

In the proposed system, To overcome from this issue author has introduce hash tags filtering technique using which we will filter hash tags to determine whether hash tag is relevant or irrelevant by matching content of both main hash tag and the annotator hash tags. If annotator assigns related hash tags then it will be relevant and supervisor will give good score to that annotator. Using HIT algorithm we can determine whether that hash tags is used more frequently or not, if it’s less frequent or unrelated hash tag then we will consider as stop hash tag

ADVANTAGES OF THE PROPOSED SYSTEM

The following are the advantages of the proposed system:

1. We show that our proposed approach exhibits better results for filtering hash tags.
2. We can able to find the images which are posted by friends and we can also give review for those images.
3. Here we can able to Percentage of prevalent information based on topic wise

5. SOFTWARE PROJECT MODULES

Implementation is a stage where the theoretical design is automatically converted into programmatically manner. Here we divide the application into number of modules and try to code each and every module for deployment. The current application is mainly divided into 2 modules:

1) Instagram Server/Admin Module

2) User Module

Now let us discuss about each and every module in detail as follows:

5.1 INSTAGRAM SERVER MODULE

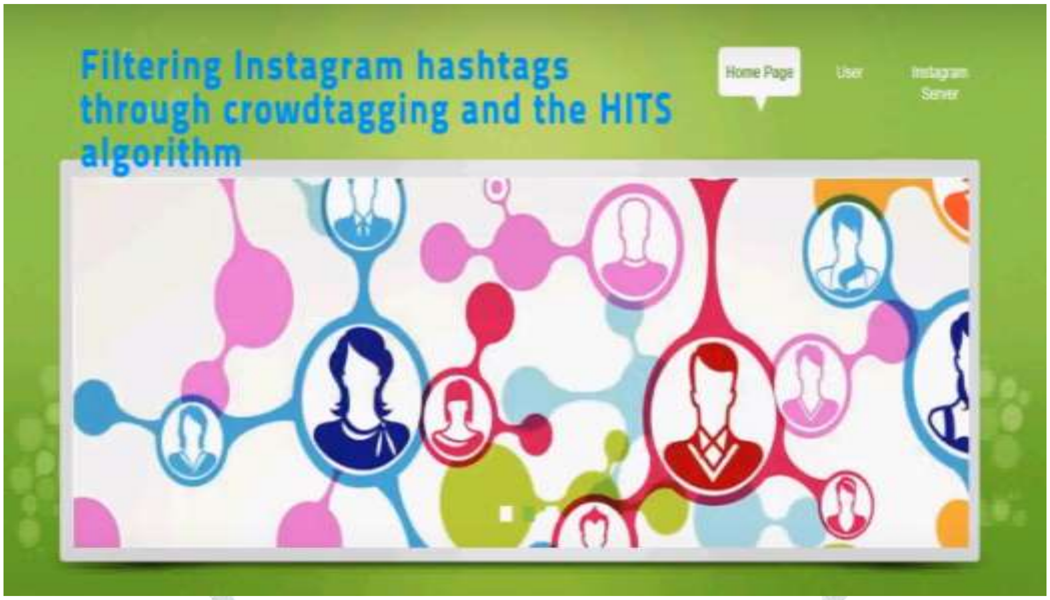
In this module, the Admin has to login by using valid user name and password. After login successful he can perform some operations such as View All users, View all Friend Request and Response ,View all users images and view all recommendation images, view image reviews, view dislikes posted by the other users, view HITS on Images result.

5.2 USER MODULE

In this module, there are n numbers of users are present. User should register before performing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user can perform some operations like send friend request or accept, upload images, search images, give comments, recommend for others, view recommended images from others.

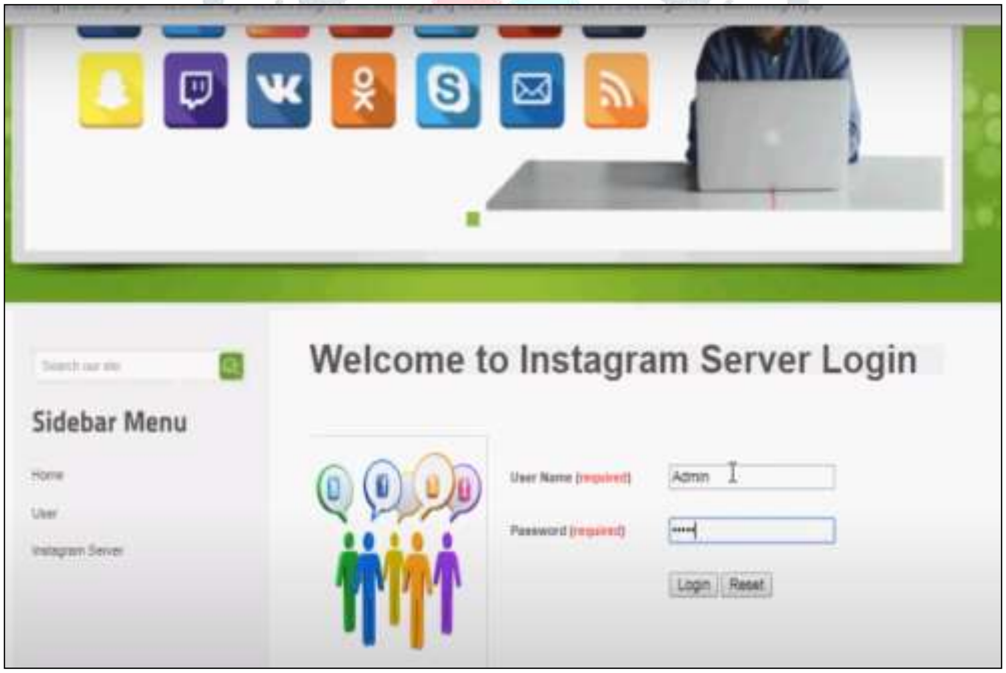
6. RESULTS (OUTPUT SCREENS)

MAIN SCREEN



Filtering Instagram hashtags through crowdtagging and the HITS algorithm

ADMIN LOGIN

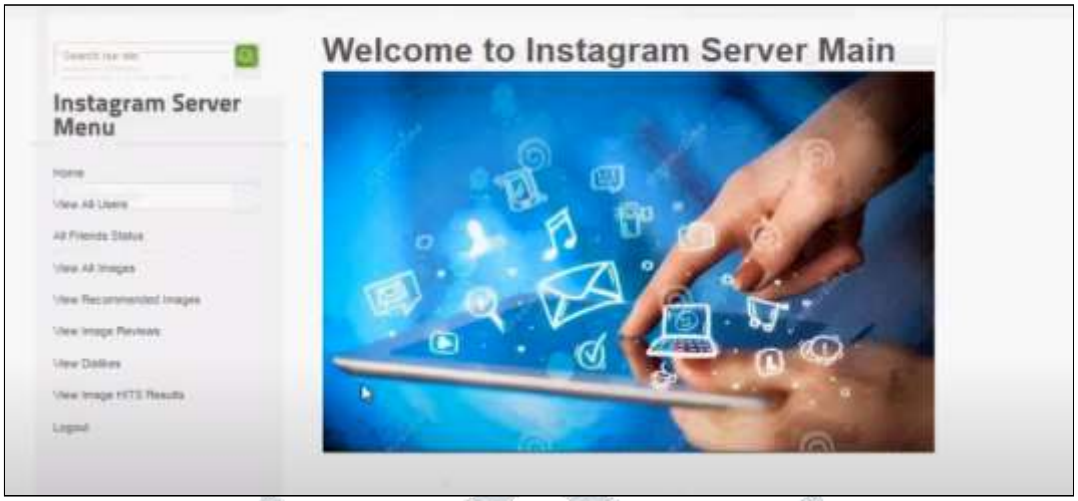


Welcome to Instagram Server Login

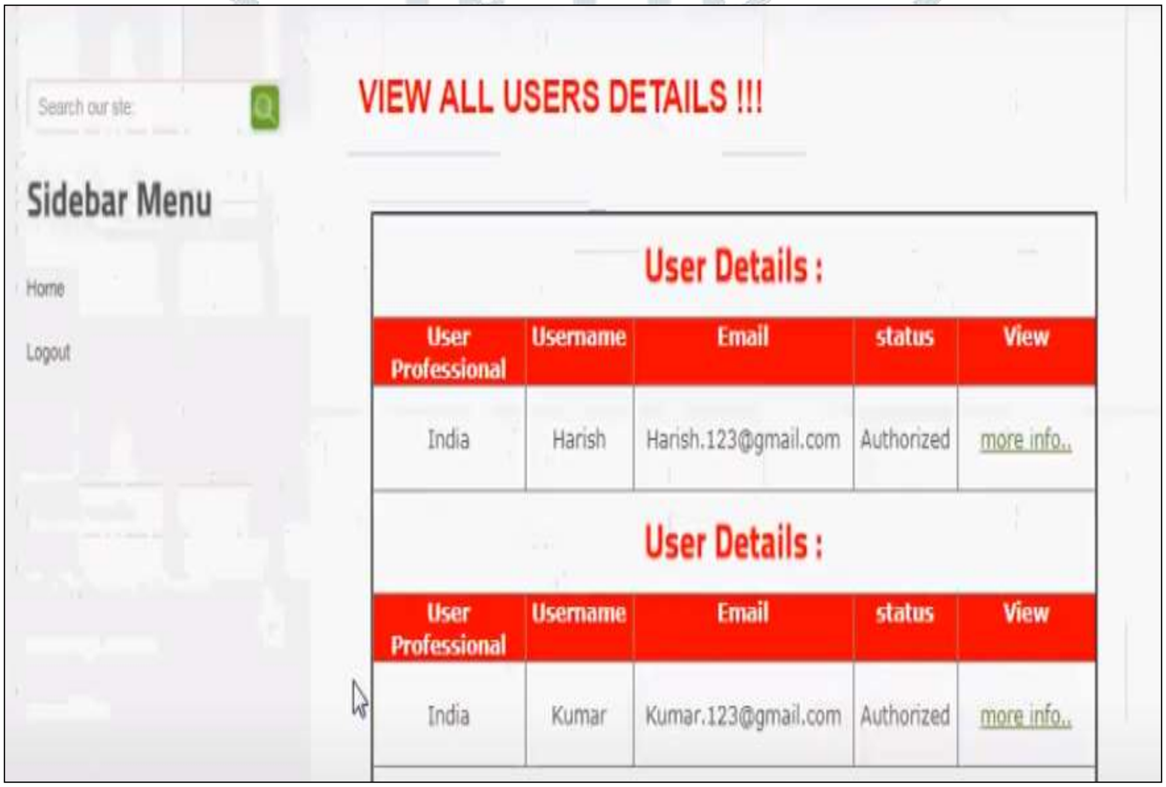
User Name (required) Admin

Password (required) •••••

ADMIN HOME PAGE



VIEW USER DETAILS



VIEW ALL USERS DETAILS !!!

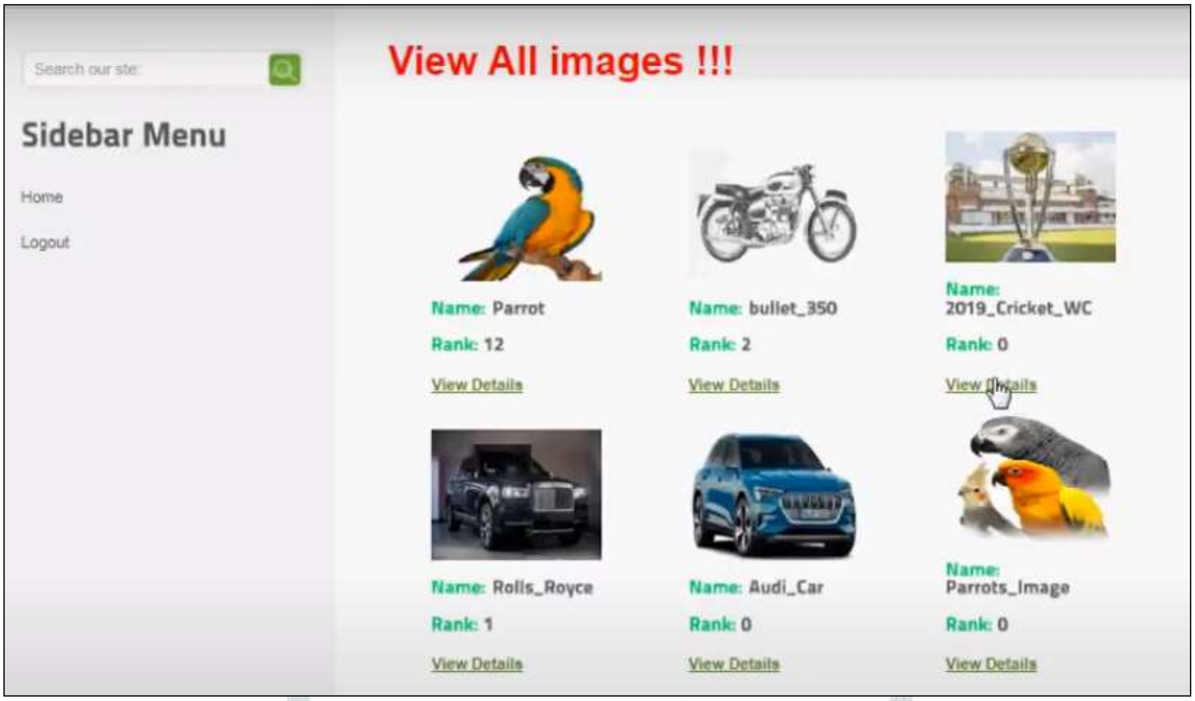
User Details :

User Professional	Username	Email	status	View
India	Harish	Harish.123@gmail.com	Authorized	more info..

User Details :

User Professional	Username	Email	status	View
India	Kumar	Kumar.123@gmail.com	Authorized	more info..

VIEW ALL IMAGE DETAILS



View All images !!!

Name: Parrot
Rank: 12
View Details

Name: bullet_350
Rank: 2
View Details

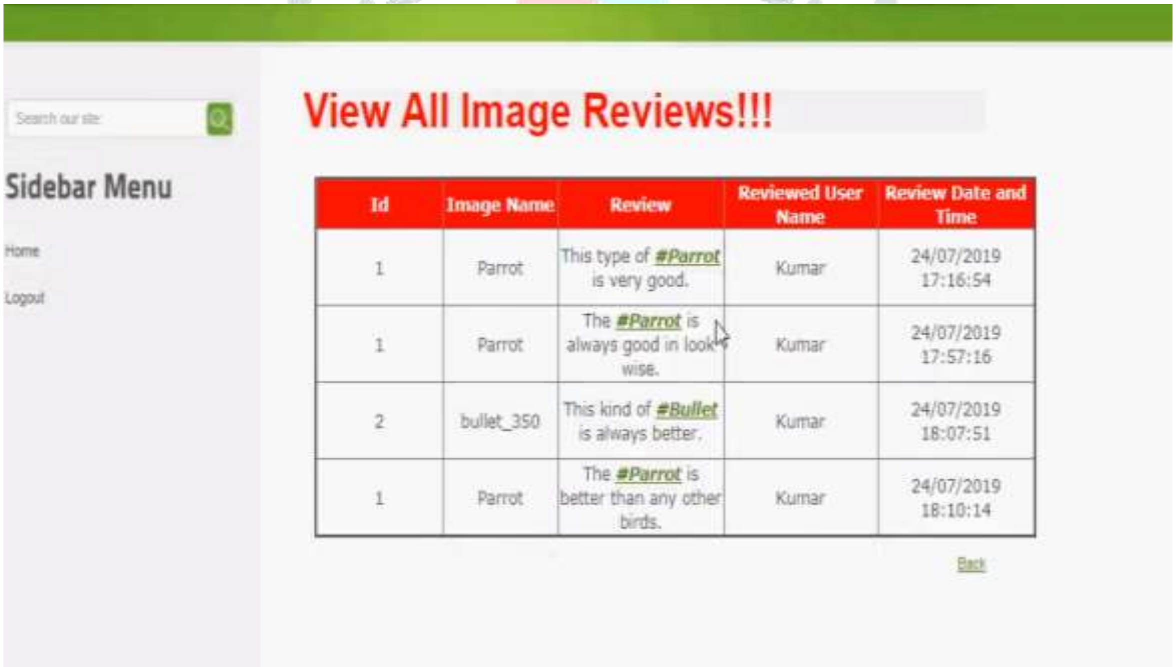
Name: 2019_Cricket_WC
Rank: 0
View Details

Name: Rolls_Royce
Rank: 1
View Details

Name: Audi_Car
Rank: 0
View Details

Name: Parrots_Image
Rank: 0
View Details

ADMIN VIEWS ALL IMAGE REVIEWS

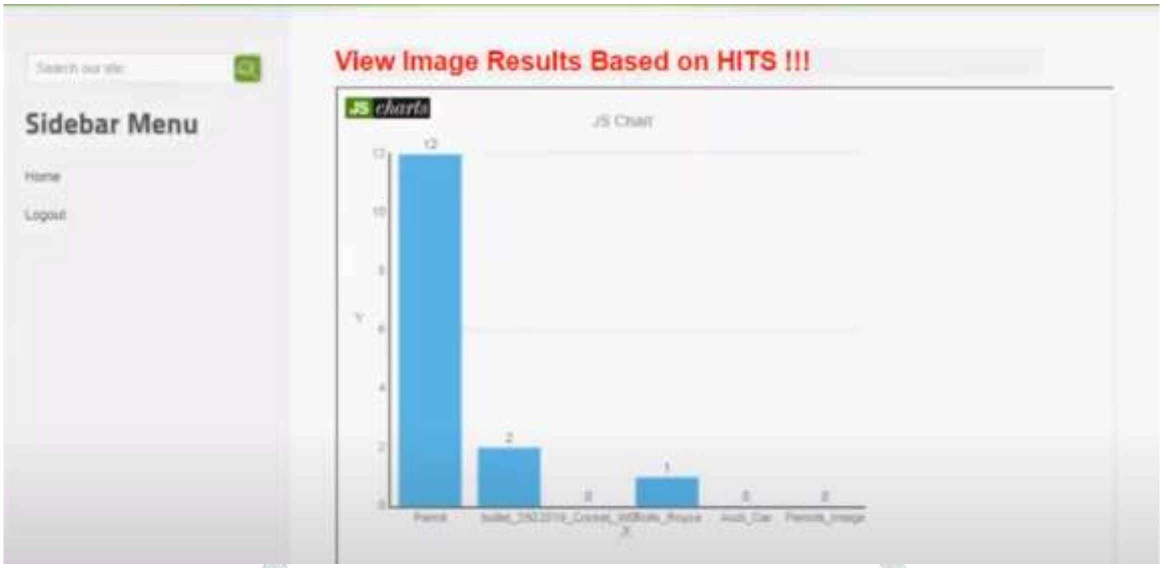


View All Image Reviews!!!

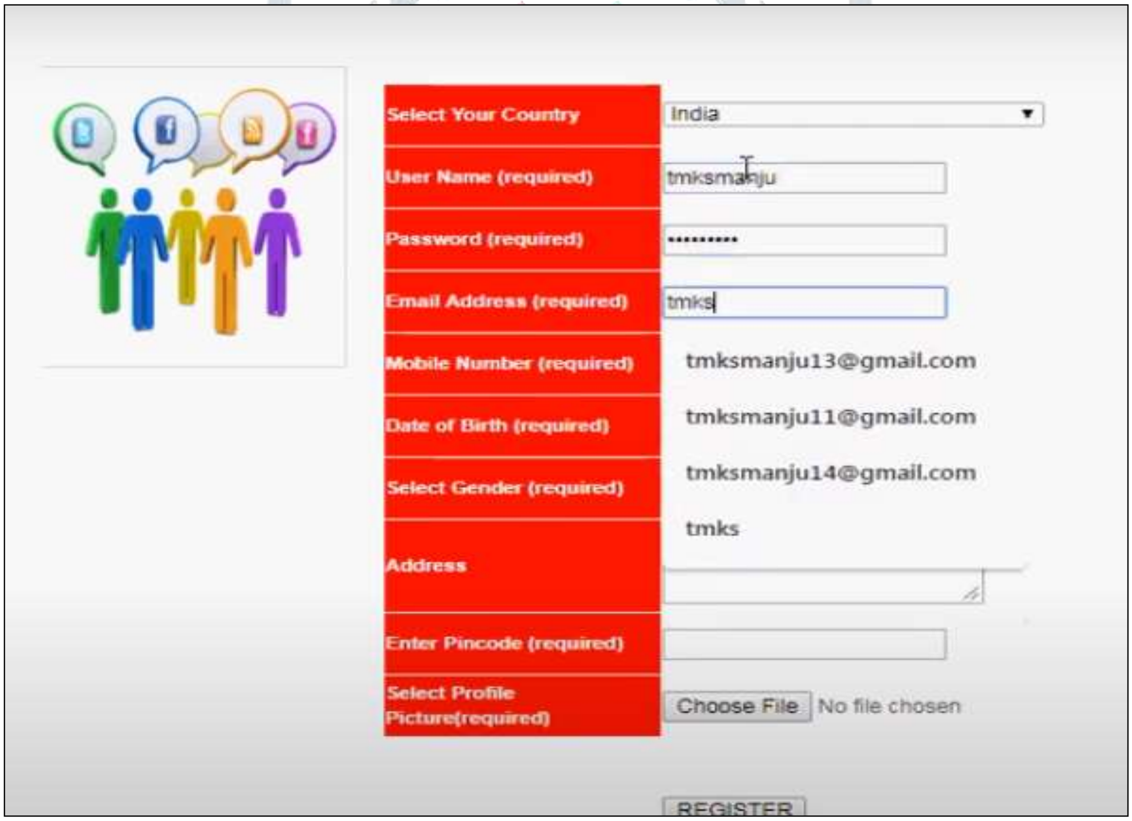
Id	Image Name	Review	Reviewed User Name	Review Date and Time
1	Parrot	This type of #Parrot is very good.	Kumar	24/07/2019 17:16:54
1	Parrot	The #Parrot is always good in look wise.	Kumar	24/07/2019 17:57:16
2	bullet_350	This kind of #Bullet is always better.	Kumar	24/07/2019 18:07:51
1	Parrot	The #Parrot is better than any other birds.	Kumar	24/07/2019 18:10:14

Back

VIEW IMAGE REVIEWS BASED ON HITS ALGORITHM

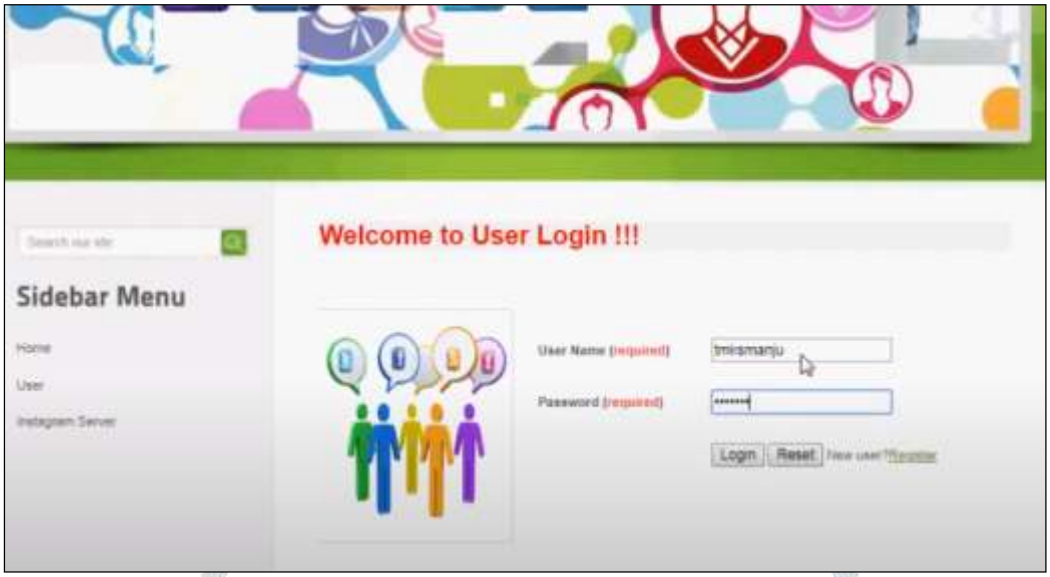


USER REGISTRATION PAGE



Select Your Country	India
User Name (required)	tmksmanju
Password (required)	••••••••••
Email Address (required)	tmks
Mobile Number (required)	tmksmanju13@gmail.com
Date of Birth (required)	tmksmanju11@gmail.com
Select Gender (required)	tmksmanju14@gmail.com
Address	tmks
Enter Pincode (required)	
Select Profile Picture(required)	Choose File No file chosen

USER LOGIN

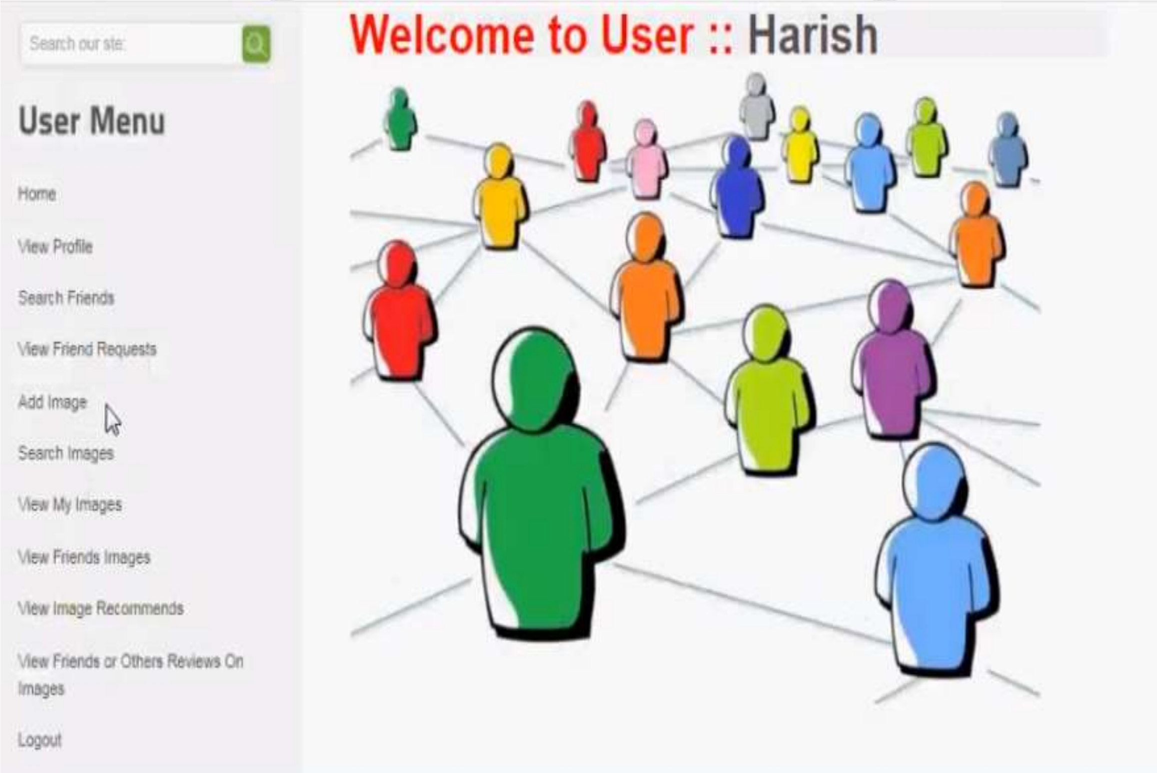


Welcome to User Login !!!

User Name (required) tmksmanju

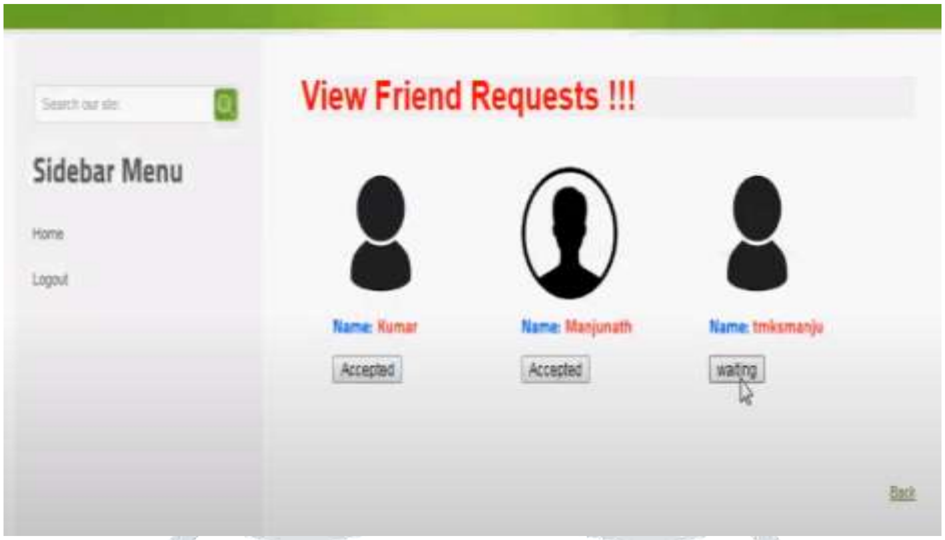
Password (required) *******

USER HOME PAGE



Welcome to User :: Harish

USER CAN VIEW NEW FRIEND REQUEST



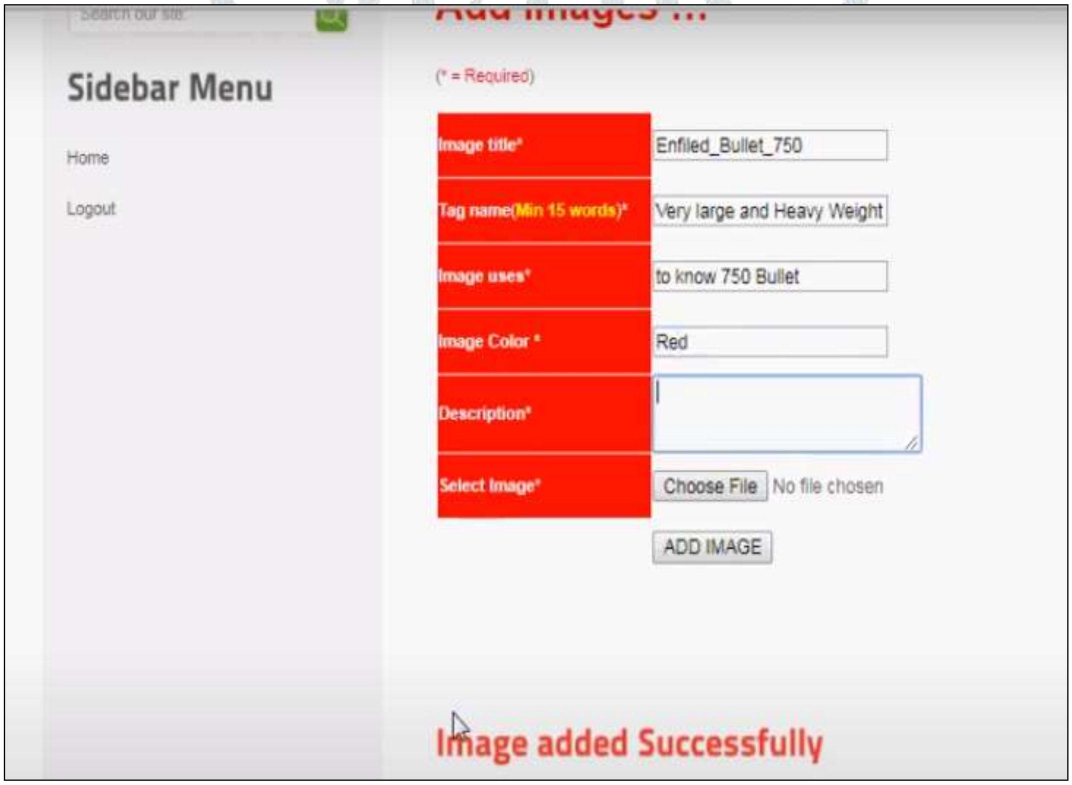
View Friend Requests !!!

Name: Kumar

Name: Manjunath

Name: tmksmanju

USER ADD NEW IMAGES



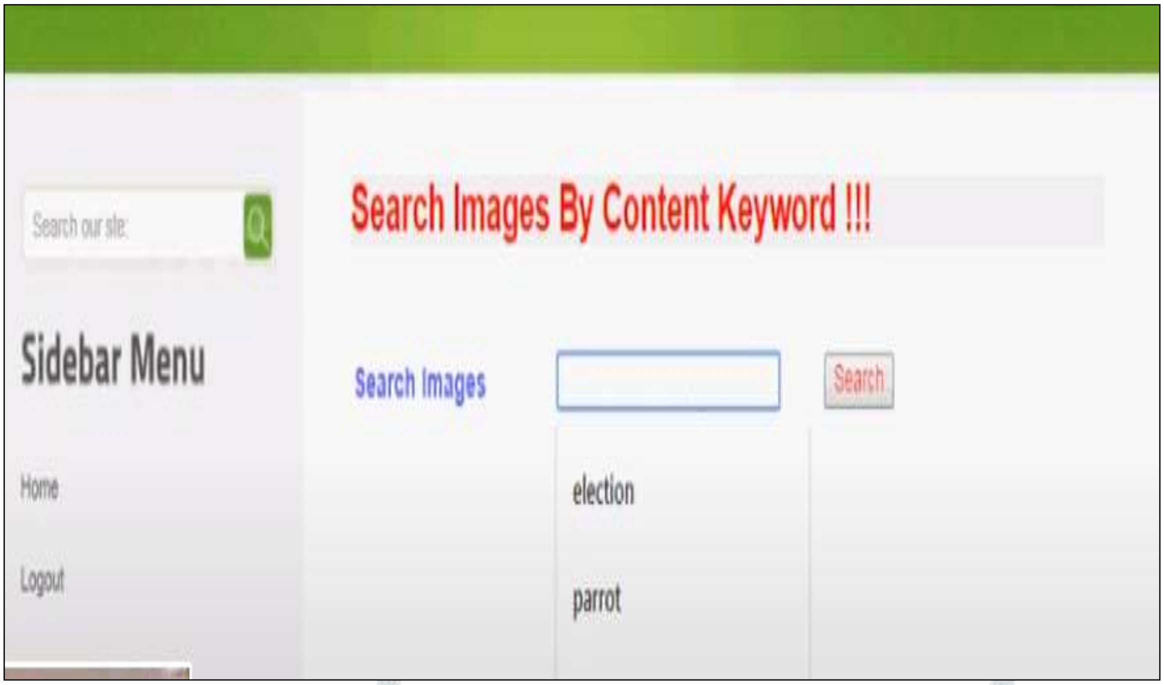
Add Images !!!

(* = Required)

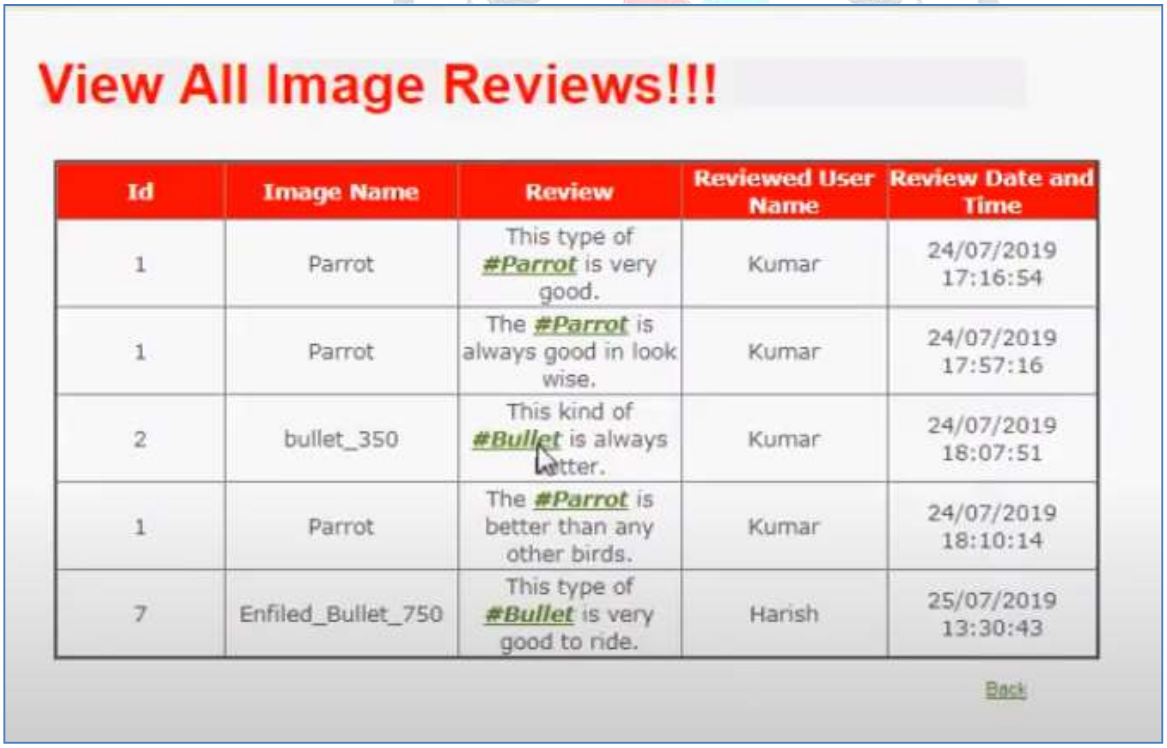
Image title*	Enfiled_Bullet_750
Tag name(Min 15 words)*	Very large and Heavy Weight
Image uses*	to know 750 Bullet
Image Color *	Red
Description*	
Select Image*	No file chosen

Image added Successfully

USER CAN SEARCH IMAGES BASED ON KEYWORD



Search Images By Content Keyword !!!



View All Image Reviews!!!

Id	Image Name	Review	Reviewed User Name	Review Date and Time
1	Parrot	This type of #Parrot is very good.	Kumar	24/07/2019 17:16:54
1	Parrot	The #Parrot is always good in look wise.	Kumar	24/07/2019 17:57:16
2	bullet_350	This kind of #Bullet is always better.	Kumar	24/07/2019 18:07:51
1	Parrot	The #Parrot is better than any other birds.	Kumar	24/07/2019 18:10:14
7	Enfiled_Bullet_750	This type of #Bullet is very good to ride.	Harish	25/07/2019 13:30:43

7. CONCLUSION

In the current work, we have presented an innovative methodology, based on the HITS algorithm and the principles of collective intelligence, for the identification of Instagram hashtags that describe the visual content of the images they are associated with. 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.

8. REFERENCES

[1] A. Argyrou, S. Giannoulakis and N. Tsapatsoulis,"Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation?" in Proc. 13th International Workshop on Semantic and Social Media Adaptation and Personalization, 2018, pp. 61-67.

[2] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas,"Crowdsourcing for multiple-choice ques-tion answering" in Proc. 28th. AAAI Conference on Artificial Intelligence, 2014, pp. 2946–2953.

[3] C. D. Cabrall, Z. Lu, M. Kyriakidis, L. Manca, C. Dijksterhuis, R. Happee, and J. de Winter, "Validity and reliability of naturalistic driving scene categorization judgments from crowdsourcing," Accident Analysis & Prevention, vol. 114, pp. 25–33, 2018.

[4] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, "A survey and analysis on automatic image annotation," Pattern Recognition, vol. 79, pp. 242–259, 2018.

[5] N. Craswell, "Mean Reciprocal Rank," in Encyclopedia of Database Systems, London : Springer, 2009, pp. 1703-1703.

[6] H. Cui, Q. Li, H. Li, and Z. Yan, "Healthcare fraud detection based on trustworthiness of doctors," in Proc. Trustcom/BigDataSE/I SPA, IEEE, 2016, pp. 74–81.

[7] A. R. Daer, R. Hoffman, and S. Goodman, "Rhetorical functions of hashtag forms across social media applications," in Proc. 32nd ACM Int. Conf. on the Design of Communication CD-ROM, ACM, 2014, p. 16.

[8] E. Ferrara, R. Interdonato, and A. Tagarelli, "Online popularity and topical interests through the lens of instagram," in Proc. 25th ACM Conf. on Hypertext and Social Media, ACM, 2014, pp. 24–34.

[9] 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. 02, p. 1750013, 2018.

[10] M. Gao, L. Chen, B. Li, Y. Li, W. Liu, and Y.-c. Xu, "Projectionbased link prediction in a bipartite network," Information Sciences, vol. 376, pp. 158–171, 2017.

[11] S. I. Gass and C. M. Harris, "Bipartite Graph," in Encyclopedia of operations research and management science, Boston: Springer, 2013, pp. 126.

[12] 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, 2016.

[13] S. Giannoulakis and N. Tsapatsoulis, “Defining and identifying stophashtags in instagram,” in Proc. INNS Conference on Big Data, Springer, 2016, pp. 304–313.

[14] S. Giannoulakis, N. Tsapatsoulis, and K. Ntalianis, “Identifying image tags from instagram hashtags using the HITS algorithm,” in Proc. 3rd Intl. Conf. on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/- DataCom/CyberSciTech), IEEE, 2017, pp. 89–94.

[15] 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, p. 12, 2018.

[16] M. Gupta, R. Li, Z. Yin, and J. Han. 2010. Survey on social tagging techniques. SIGKDD Explor. Newsl. 12, 1, November 2010, pp. 58-72.

[17] A. Hotho, R. Jaschke, C. Schmitz and G. Stumme, “FolkRank: ¨ A Ranking Algorithm for Folksonomies,” in Proc. of the 12th Workshop on Knowledge Discovery, Data Mining, and Machine Learning, 2006, pp. 111–114.