2015 International Conference on Information Technology Systems and Innovation (ICITSI)
Bandung - Bali, November 16 - 19, 2015
ISBN: 978-1-4673-6664-9

TF-IDF Method in Ranking Keywords of Instagram Users' Image Captions

Bernardus Ari Kuncoro
Master of Information Technology
Bina Nusantara University
Jakarta, Indonesia
Email: b.kuncoro [at] binus.ac.id

Bambang Heru Iswanto
Department of Physics
Jakarta State University
Jakarta, Indonesia
Email: bhi [at] unj.ac.id

***Abstract*—Instagram is one of the popular social media applications used by a wide range of people around the world. The significant growth of active Instagram users affects the size of Instagram data. The more number of users, the larger and more various Instagram data is posted. In line with its popularity, in recent years many researchers begin to study and analyze it for various purposes, such as detecting event photos based on location, clustering the photo content, advertising strategies based on user types, and so on. As of now there are three types of data available in Instagram which are text, image, and video. In this paper we propose Term-Frequency and Inverse Document Frequency (TF-IDF) method to rank keywords of top twenty most followed Instagram users based on image captions of Instagram. The objective of this research is to automatically know the main idea of Instagram users based on 50 recent image captions posted. In our experiments, TF-IDF has been successfully implemented to reveal a set of keywords with its ranking. The highest ranking of keyword is indeed the main topic of a user, indicated by the value of TF-IDF. The result of study indicates that TF-IDF method is very useful to find and rank the keywords of Instagram users image captions. In the future research, the ranking keywords are needed in solving classification and clustering tasks as feature extractions.**

Keywords*—*Instagram; text mining; Term-Frequency and Inverse Document Frequency, social media

I. Introduction

Instagram is one of the popular social media platforms that provides users a quick way to capture and share their life moments with followers through a series of filter-manipulated photo and video. It is more popular amongst a younger demographic. Over 35% of people using Instagram are between ages 18-29 years [1]. Since establishment in October 2010 until this paper was written, the growth of active Instagram users has significantly increased. According to an updated data by the official Instagram account in September 2015 [2], Instagram has been registered by 400 million users which is 25% higher than the number of registered users in December 2014. Another interesting fact is that the average of photos being uploaded by users per day is more than 80 million photos.

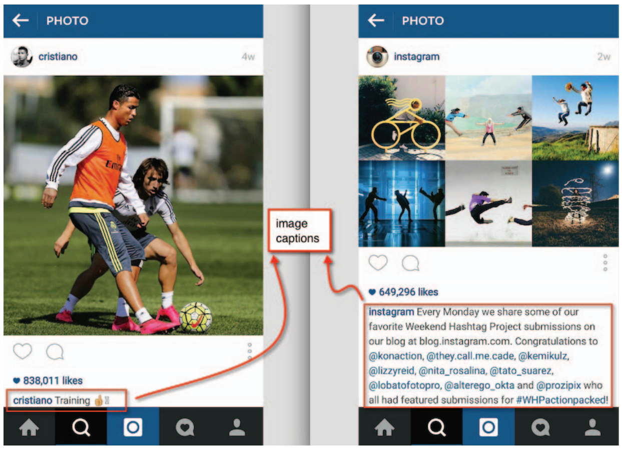

Fig. 1. Example of Cristiano's and Instagram's Posts with Image Captions

Despite the fact of Instagram popularity, the number of researches focused in Instagram is very low. In 2014, Hu, Manikonda, and Kambhampati [3] wrote that their work is believed to be the first study to conduct a deep analysis of photo content and user activities and types on Instagram. In their study, computer vision and identification by clustering were successfully applied thus eight popular categories of photos and five distinct types of Instagram users were revealed. A dissertation related to Instagram was reported by McCune. He investigated peoples motivations of using Instagram through a survey study of 23 Instagram users [4]. In 2013, Silva, Vaz de Melo, Almeida, Salles, and Loureiro have applied visualization and cultural analytics on Instagram photos from different cities in the world to trace their social and cultural differences [5].

Instagram has three types of data which are text, image, and video. To narrow down the idea of this study, only text data was used. The text data used in this study was the image caption that represents the description of the image. As illustrated in Fig. 1, the image caption is located under the image that was posted by the user.

The research question of this study is *"How to find keywords* *and the rankings of Instagram account based on the image caption data posted?"*. To answer this question, text mining (TF-IDF) method is used. The output of this study is the keywords with the ranking value. The higher the ranking, the more relevant the keyword with the captions that users posted. The significances of this study are as follows. First, the ranking keywords of username image captions can be used as features of advanced research such as clustering, classification, and profiling of Instagram username. Second, this study adds the diversity of Instagram data research with a different approach which is text mining. Third, the method can be used to expedite researchers in retrieving significant words of the users, as this can be done automatically rather than a manual retrieval, by keeping an eye on the captions posted by the users.

II. Dataset

The dataset was crawled using API of Instagram. First, the top 20 most followed Instagram usernames were collected. The list is based on http://socialblade.com/instagram/top/100/followers accessed on October 7, 2015 [6] and it can be seen in Table I. Second, in order to know the most updated keywords of the users, only 50 of the most recent image captions were used. Each username is assumed as one document that contains a bag of words, hence there are 20 documents in total.

The following are characteristics of Instagram image caption data. Please be noted that these can be changed in the future without prior notice due to Instagram updates.

1) Image caption character limit: the limit for captions on the photo and subsequent comments caps is 2200 characters each. User is also allowed not to write a caption at all.
2) Hashtag limit: The limit of hashtag is 30 hashtag per caption.
3) Symbol characters: Some of the users uses the symbol characters provided in the smartphone keyboard.
4) Writing technique: As Instagram has 2200 characters limit, spelling and cyber slang in the image caption is not often used by users compared to Tweets in Twitter.
5) Availability: The amount of data available is extremely large. According to the Instagram official release in September 2015, there are 80 million photos uploaded daily. The Instagram API facilitates the collection of image captions as well as the URL Link of image.
6) Topics: Instagram users post photos and videos in a wide variety of topics. Previous research observed that there are eight main photo categories which are friends, food, gadget, captioned photo, pet, activity, selfie, and fashion posted in Instagram [3].
7) Weekend Onpeak: Users tend to post the photos and videos during weekends and at the end of the day. [7]

TABLE I
Top 20 Instagram Profiles

Rank	Username	Media	Followers	Following
1	instagram	2,509	103,226,690	182
2	taylorswift	732	49,451,242	77
3	kimkardashian	3,167	48,014,416	96
4	beyonce	1,172	47,173,577	0
5	selenagomez	1,028	45,858,936	173
6	arianagrande	1,869	44,598,791	952
7	justinbieber	2,508	40,228,982	73
8	kendalljenner	2,343	38,055,799	170
9	kyliejenner	3,338	38,075,231	186
10	nickiminaj	3,387	35,185,711	382
11	khloekardashian	2,935	33,091,863	149
12	natgeo	8,432	32,835,979	94
13	neymarjr	3,018	32,555,977	1,023
14	cristiano	602	31,865,306	198
15	mileycyrus	4,280	29,539,569	384
16	katyperry	366	28,891,826	217
17	therock	1,343	28,778,259	64
18	jlo	1,185	27,824,613	966
19	badgalriri	3,267	26,976,059	1,166
20	kourtneykardash	2,021	25,875,453	72

III. Methodology

The methodology of this study is illustrated in Fig. 2. Basically, there are three moduls used. They are retrieval, preprocessing, and ranking moduls. In retrieval modul, each of the usernames was used to request the recent 50 image captions via Instagram API. The output of this retrieval is a group of text files. Since the number of username used is 20, hence the output of this modul is 20 text files. This methodology is inspired by Kumar and Sebastian research in 2012 [8].

The next modul is the preprocessing modul. It is needed to pass the important words and filter irrelevant words and characters in each document. The first preprocessing modul step is removal of HTML and symbol characters. It is important because, the users commonly write symbol characters that has less significant meaning or a non keyword symbol. The second step of preprocessing modul is punctuation, #tag, @tag, and stopwords removal. The main goal of this step is to retrive essential words and to eliminate words that has less significance towards the documents such as "the", "is", "are", "an", "of", "to", etc. It is also useful to reduce indexing file size, improving efficiency and effectiveness. The third step is standardizing words. For example the user sometimes writes 'go hooooome', thus the output of this step is 'go home'. Last step on preprocessing modul is URLs removal. It is clear that the URL link is not significant to be used to reveal the keywords.

Upon finishing preprocessing the data, ranking process is then applied. The first step of this modul is tokenization. Its objective in this case is to break the text up into words or other meaningful elements called tokens. Then each tokens, or commonly refered to as terms are used to form vector space model.

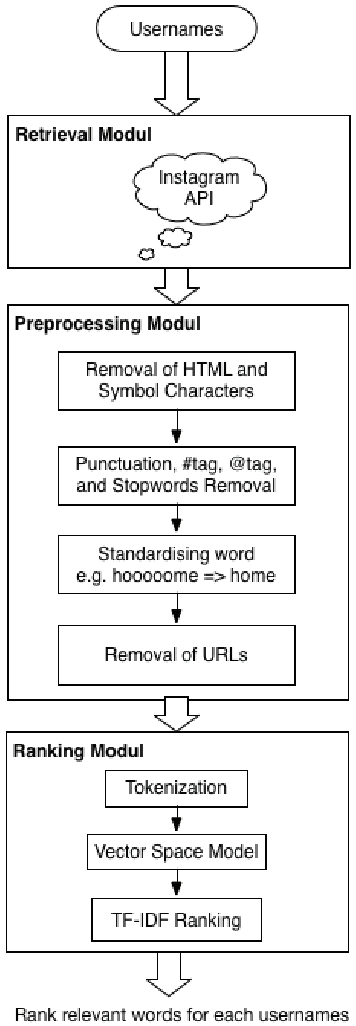

Fig. 2. Methodology

TF-IDF stands for "Term Frequency, Inverse Document Frequency". It is a way to score the importance of words (or "terms") in a document based on how frequently they appear across multiple documents. Besides that, it is the most common weighting method used to describe documents in the Vector Space Model (VSM), particularly in Information Retrieval problems. TF-IDF is a relatively old method proposed by Salton and Buckley in 1988 [9]. Despite its age, it is simple and effective, making it a popular starting point compared to the more recent algorithms. To know more about the TF-IDF, here are the descriptions of TF and IDF.

1) TF is a measure of how many times the terms t present in each file document d. The formula of TF in mathematical symbol is as follows:

$$\mathrm{TF}(t,d) = \sum_{x \in d} \mathrm{fr}(x,t) \tag{1}$$

where the $\mathrm{fr}(x,t)$ is a simple function defined as

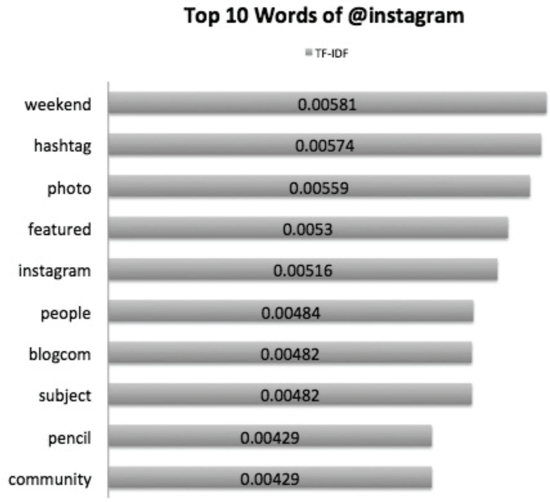

Fig. 3. Top 10 Words of @instagram Account

$$\mathrm{fr}(x,t) = \begin{cases} 1, & \text{if } x = t \\ 0, & \text{otherwise} \end{cases}$$

Hence, $\mathrm{TF}(t,d)$ returns how many times the term t is present in the document d.

2) IDF is defined with the following formula:

$$\mathrm{IDF}(t) = \log \frac{|D|}{1 + |\{d : t \in d\}|} \tag{2}$$

where $|\{d : t \in d\}|$ is the number of documents where the term t appears, when the term-frequency function satisfies $\mathrm{TF}(t,d) \neq 0$, were only adding 1 into the formula to avoid zero-division.

3) TF-IDF formula is defined as follows:

$$\text{TF-IDF}(t) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t) \tag{3}$$

The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general, thus the more appears in a document, the more a word is estimated to be significant in that document.

IV. Result and Discussion

The proposed method was applied to the top twenty most followed Instagram usernames as input. The number of ranking keywords can be varied and in this study was limited up to 10 ranks. Thus the result are 20 username items with 10 ranking keywords. Three samples of the results that represents the top 10 words and TF-IDF value of each Instagram users are illustrated with a bar chart in Fig 3, 4, and 5. The first bar is the highest ranking words or the most relevant word of a specific user. According to those figures, the most relevant words for each @instagram, @taylorswift, and @cristiano users are weekend, toronto, and drive respectively.

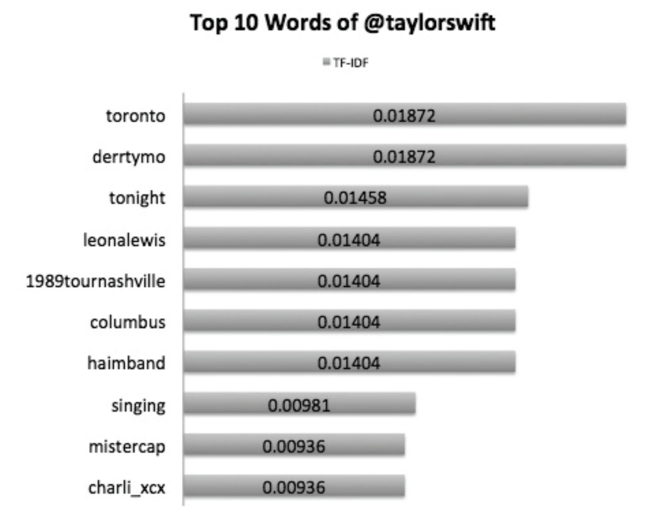

Fig. 4. Top 10 Words of @taylorswift Instagram Account

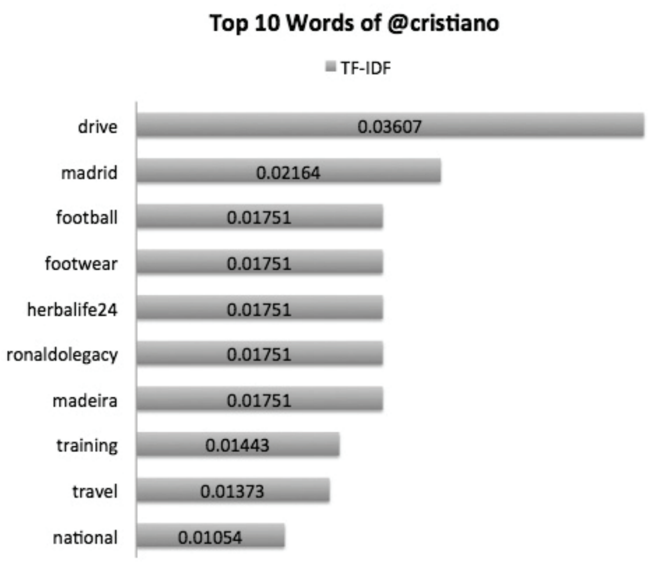

Fig. 5. Top 10 Words of @cristiano Instagram Account

Going more deeply to the highest ranked keyword in each username, it turns out that they have different reasons why it becomes the highest. The term 'weekend' in @instagram account becomes the highest keyword, because during the time data was crawled, @instagram held Weekend Hashtag Project. The username @taylorswift whose has term 'toronto' as her highest rank keyword, because she has just shared several photos about her concert in Toronto, Canada. While the term 'drive' in @cristiano becomes the highest rank keyword, because he is currently endorsing his new sport drink product and named CR7Drive.

Figure 6 and 7 illustrates the result of keyword ranking for the remaining 17 usernames. They are arranged from the higest rank to the lowest. For example, @arianagrande highest rank keyword is 'focus' followed by 'babes', 'andrea', until the least rank term: 'tulsa'. Other than that, some of the terms in the result are not easily understood due to it being slang terms (e.g. yall, j4, poo, etc.), usernames of other Instagram users (e.g. ronyalwin), numbers (mostly date), and non-English language. This needs to be improved in future works.

arianagrande	TF-IDF
focus	0.0296
babes	0.02458
andrea	0.01639
yall	0.01285
lexie1225	0.01229
frankiejgrande	0.01229
pups	0.01146
chicago	0.01146
ultabeauty	0.00819
tulsa	0.00819

therock	TF-IDF
hardest	0.008
brutus	0.007
puppies	0.007
room	0.006
hobbs	0.006
surgery	0.006
workers	0.005
appreciate	0.004
service	0.004
walker	0.004

justinbieber	TF-IDF
floydmaywear	0.061
rorykramer	0.061
6weeks	0.04575
lol	0.03198
diplo	0.0305
camcorderapp	0.0305
champ	0.02781
knowing	0.02513
via	0.02392
talkshit	0.01525

badgalriri	TF-IDF
brazil	0.05247
rihanna	0.02939
r100	0.02939
chile	0.01615
stancemuse	0.0147
santiagode	0.0147
flashback	0.01211
stance	0.01211
london	0.01027
yall	0.01025

katyperry	TF-IDF
ronyalwin	0.03567
whitneymuseun	0.01189
prismatic	0.01189
icons	0.00793
92515	0.00793
flute	0.00793
credits	0.00793
burning	0.00793
picchu	0.00793
ray	0.00716

beyonce	TF-IDF
2030	0.10234
globalcitizens	0.10234
globalcitizeno	0.10234
dime	0.08432
poverty	0.08432
effort	0.08432
involved	0.07153
spending	0.06161
without	0.06161
action	0.05351

jlo	TF-IDF
vegas	0.064
83115	0.043
flux	0.043
papermagazine	0.043
olivier_rousteing	0.035
idol	0.032
iheartradio	0.027
luxein	0.027
balmainparis	0.026
luxe	0.016

kendalljenner	TF-IDF
voguejapan	0.03892
sims	0.03542
luigiandiango	0.03542
vogueparis	0.03542
luigimurenu	0.03542
esteelauder	0.02362
j4	0.02362
voguemagazine	0.02362
david	0.02133
inside	0.01422

Fig. 6. Result of Keyword Ranking for Remaining Usernames - part 1

khloekardashian	TF-IDF
jenatkinhair	0.01126
khlo	0.01126
bio	0.01072
ktu9f	0.01025
gunnarfitness	0.01025
download	0.00844
etienneortega	0.00844
streams	0.00844
yeezy	0.00844
apps	0.00844

kyliejenner	TF-IDF
kylie	0.02641
jennercom	0.02112
styledbyhrush	0.0174
links	0.01584
lizzie	0.01056
iammorethan	0.01056
bullying	0.01056
prevention	0.01056
right	0.00954
livelokai	0.0087

neymarjr	TF-IDF
deus	0.04023
nos	0.02981
proteja	0.02785
abenoe	0.02785
felicidades	0.02476
voc	0.01947
parabns	0.01785
hoje	0.01547
irmo	0.01547
meu	0.01238

kimkardashian	TF-IDF
westcom	0.01402
kardashian	0.01225
kim	0.01225
tutorial	0.01155
bodysuits	0.01051
juergen	0.01051
teller	0.01051
sorbetmag	0.01051
glam	0.0098
makeupbyariel	0.00701

mileycyrus	TF-IDF
mileyonsnl	0.10965
nbcsnl	0.05482
andherdeadpetz	0.04934
oct3	0.04386
weregoingontour	0.03289
poo	0.02999
lesdogg	0.01645
gecko	0.01096
goingontour	0.01096
jimmyfallon	0.01096

nickiminaj	TF-IDF
givenchy	0.01483
milan	0.01166
balenciaga	0.01061
winning	0.01061
spoonfulofsass	0.01061
riccardotisci17	0.00874
chanel	0.00874
actress	0.00874
bag	0.00852
barbie	0.00742

kourtneykardash	TF-IDF
millersinmontana	0.03838
bye	0.01535
bridge	0.01535
obsessed	0.01535
hike	0.01265
yellowstone	0.01265
currently	0.01073
pretty	0.00924
robe	0.00924
babies	0.00924

natgeo	TF-IDF
photosociety	0.01714
photo	0.011
joelsartore	0.01082
national	0.01032
park	0.00977
lake	0.00902
creative	0.00883
assignment	0.00817
photograph	0.00669
story	0.00613

selenagomez	TF-IDF
same	0.01177
nicolasghesquiere	0.01071
revival	0.01071
byjakebailey	0.01071
bio	0.00933
favorites	0.00882
109	0.00714
flaremag	0.00714
louisvuitton	0.00714
annabels	0.00714

Fig. 7. Result of Keyword Ranking for Remaining Usernames - part 2

V. Conclusion

A set of keywords with its ranking have been successfully revealed from image captions of the top 20 most followed Instagram users. The use of the proposed method in which TF-IDF is implemented is very simple and effective in revealing the keywords and its ranking from a certain user. The results show that the highest ranking of keyword is indeed the main topic of a user, indicated by the value of TF-IDF. The higher the TF-IDF value, the more relevant that keyword is to the specific Instagram username. However, this work still needs to be improved in terms of understanding slang words and non-English language, adding feature of keywords based on annotation images, and so on.

References

[1] J. Golbeck, *Introduction to Social Media Investigation: A Hands-On Approach*, 1st ed. Massachusetts: Syngress, 2015.

[2] Instagram, "Instagram 400,000,000," 2015. [Online]. Available: https://instagram.com/p/78n-7MBQU8/

[3] Y. Hu, L. Manikonda, and S. Kambhampati, "What we Instagram : a first analysis of Instagram photo content and user types," *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media*, pp. 595–598, 2014.

[4] Z. Mccune and J. Thompson, "Consumer Production in Social Media Networks : A Case Study of the Instagram iPhone App," Ph.D. dissertation, University of Cambridge, 2011.

[5] T. H. Silva, P. O. S. V. D. Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "A picture of instagram is worth more than a thousand words: Workload characterization and application," *Proceedings - IEEE International Conference on Distributed Computing in Sensor Systems, DCoSS 2013*, no. i, pp. 123–132, 2013.

[6] Socialblade, "Top 100 Instagram Users by Followers," 2015. [Online]. Available: http://socialblade.com/instagram/top/100/followers

[7] C. S. Araujo, L. P. D. Correa, A. P. C. D. Silva, R. O. Prates, and W. Meira, "It is Not Just a Picture: Revealing Some User Practices in Instagram," *2014 9th Latin American Web Congress*, no. May, pp. 19–23, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7000167

[8] A. Kumar and T. M. Sebastian, "Sentiment Analysis on Twitter," *International Journal of Computer Science Issues*, vol. 9, no. 4, pp. 372–378, 2012.

[9] G. Salton and C. Buckley, "Term-weighted approaches to automatic text retrieval." *In Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.