Date of publication SEP-30, 2021, date of current version JUN-04, 2021.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.19.3.2287

Hashtags Influence Maximization: Choosing the most Influential Hashtags on Instagram

KRISTO RADION PURBA
School of Computer Science and Engineering, Taylor's University, Malaysia
(e-mail: kristoradionpurba@sd.taylors.edu.my)

Corresponding author: Kristo Radion Purba (e-mail: kristoradionpurba@sd.taylors.edu.my).

ABSTRACT This study aims to find influential hashtags using Influence Maximization (IM). The IM approach was implemented using hashtags network collected from Instagram. This study can help business or ordinary users to choose the most engaging hashtags for posting, as opposed to selecting influencers, which was widely studied using the IM approach. The network was build based on the hashtags co-appearance frequency. Three IM algorithms, i.e. SSA, DSSA, and IMM, were simulated under the IC and LT models. The algorithms were compared against TopUsage, which is the top hashtags based on the usage count. The IM algorithms have a similar performance with TopUsage in influence spread, which shows that IM can adapt to the hashtags network. However, the IM algorithms produced better hashtags based on the UER (unique engagement rate) metric. The best UER performance was shown by DSSA under the LT model, where it outperformed TopUsage by 17.23%. In the hashtags categorization scenario, DSSA-LT outperformed the UER of TopUsage by up to 6.87%. This categorization is more useful in a practical scenario, to find only relevant hashtags for posting. The hashtags generated by DSSA-LT are about 30-35% different from TopUsage.

KEYWORDS Influence maximization; information diffusion; data analysis; social network; Instagram.

I. INTRODUCTION

THE increasing use of social networks has generated large amounts of data [1], including user data, posts, and networks. Social media has become a research interest [2], such as influence maximization (IM) [3] as a widely studied problem [4]. As opposed to identifying influential users, this research aims to identify influential hashtags using IM. Several social studies have proven the effectiveness of hashtags in increasing likes [5], [6]. In an initial observation, some highly used hashtags didn't proportionally generate many likes. These *low-performing* hashtags were overused by a small group of people to increase exposure. Thus, choosing hashtags with a high usage count is not always the best option.

Influence maximization (or information diffusion) has a practical benefit for brand marketers. Despite the emergence of studies on IM, most of them used users as the influencers. Identification of influential hashtags has been previously studied [7]; however, it used hashtag's virality (usage count) as the main metric. As previously stated, a high usage count doesn't directly correlate to likes. Other studies on hashtags [8], [9], [10] lacked hashtags network analysis and more focused on users' popularity.

The IM approach in this study was benchmarked theoretically using influence spread, and realistically using the average engagement rate. The following questions were addressed in this research, i.e. (Q1) What are the signs of spammy hashtags? (Q2) What is the performance of the produced hashtags set using the IM technique if compared to the hashtags set based on usage count? (Q3) What is the performance of the produced hashtags set using IM in the hashtag's categorization scenario?

The contributions of this research are as follows: (1) The creation of hashtags network based on relatedness using

Instagram data, (2) implementation of IM technique on hashtags, (3) using verification using engagement rate data. This research can help business or ordinary users to increase exposure. The rest of this paper is organized as follows, i.e., related works, research methodology, data analysis, experimental results, and conclusion.

II. RELATED WORKS

The effectiveness of hashtag has been widely researched in a social context. This includes its function in social networks [11], [12] or as a tool to increase likes [5], [6]. A Twitter hashtags graph was developed in [13] to analyze the semantic relatedness between hashtags.

Previous studies on identifying influential hashtags used the PageRank algorithm [14] and independent cascade diffusion model [7] to obtain information from the hashtags graph. However, both studies used hashtags usage count as the ground truth. While they produced a high level of accuracy, Twitter's social environment can be different from Instagram. On the latter, the engagement rate is a critical measure of success [15].

Various studies have proven the effectiveness of graph-based techniques, such as in a market prediction, the spread of opinion or rumour [16], and fake users identification [17], [18]. Fake users were found to have many similar mutual connections [17], [18]. Similar to fake users or spam, our initial observation revealed that specific hashtags are often used for spamming, which is only used by certain groups of people. Using a graph-based approach, we aim to identify more engaging hashtags.

III. METHODOLOGY

This section presents the engagement metrics, the data collection process, the creation of the hashtags network, and the baselines for the IM implementation.

A. ENGAGEMENT METRICS

In this study, two popularity metrics were used, namely Unique Engagement Rate (UER) and average Engagement Rate (ER). As commonly used in other studies, ER is defined as *likes* + *comments* divided by *followers* [19]. In this research, ER can be calculated for each post or each hashtag. ER is formulated as follows:

$$ER(post) = \frac{likes(post) + comments(post)}{followers(post)} \quad (1)$$

$$ER(hashtag) = \frac{\sum_{i=1}^{npost} ER(i)}{npost(hashtag)} \quad (2)$$

where:

- *likes*, *comments*, *followers* = The metric values
- *ER(post)* is the ER of a post, while ER(hashtag) is the ER of a hashtag
- *ER(hashtag)* is the average ER of all posts that contains the hashtag
- *npost(hashtag)* = The number of posts that includes the hashtag (in the dataset)

Based on the same idea as ER, we established UER, defined as the *total unique likers* divided by *total unique followers* from all posts containing the hashtag. The *unique liker* represents the size of the engaging audience of a hashtag, while the *unique follower* represents the size of potential users. While including comments was also an option, we couldn't collect them reliably. UER is formulated as follows:

$$UER(hashtag) = \frac{unique\ likers(hashtag)}{unique\ followers(hashtag)}, \quad (3)$$

where:

- *unique likers (hashtag)* = The unique number of likers from all posts containing the hashtag
- *unique followers (hashtag)* = The unique number of followers from all users that upload a post containing the hashtag

Like ER, a low number of UER indicates a less interesting hashtag, which may contain posts uploaded by fake or spammy users. However, UER emphasizes getting the true audience size of a hashtag. For example, the UER of #malaysia is quite low at 1.76%, even though it has a high ER of 7.69%. A hashtag with low UER but high ER indicates that the likes come from a small group of people. Both ER and UER can be calculated for one or multiple hashtags by using average.

Users with low followers are known to produce high ER [19], which can produce unfairness for average ER calculation. In our dataset, the average ER of posts from users with 0-99 followers = 31.01%; 100-199 followers = 16.24% 200-299 followers = 13.65%. Overall average ER of all posts from users with >= 100 followers is 10.167%. Thus, several filters were applied for the calculation of ER and UER, i.e.

- Posts from users with followers < 100 were excluded.
- Each post's ER is capped at 20.33%, twice the overall average ER. The ER limit is applied due to some unusual values, such as 3,299% (59,397 likes/comments, 1,800 followers).

B. DATA COLLECTION

The dataset[1] was collected from April to May 2020 from the followers of 24 universities (Instagram accounts) in Malaysia. This localized network is useful in creating a large number of connections between hashtags. The data from each user were collected using the Instagram API and third-

[1] The dataset used in this research is available on https://www.kaggle.com/krpurba/im-hashtag-network

party websites. Besides, we collected post *likers* and users' *followers* for benchmarking purposes. The summary of dataset sizes is shown in Table 1.

Table 1. Raw dataset sizes

Dataset	Data Count
Users	70,409
Posts (that have hashtags)	383,281
Hashtags (graph's nodes)	72,592
Hashtags connections (graph's edges)	2,116,718
Post's likers	47,689,496
User's followers	295,088,184

C. HASHTAGS NETWORK/GRAPH

The original independent cascade (IC) and linear threshold (LT) models used an inverse proportion of the number of followees as the edge weights [3]. This research used the weights based on the frequency of each two hashtags appear together in a post. Initially, this hashtags network was an undirected graph, which was then converted to a bidirectional graph by using the following edge weight formula:

$$w(s,t) = \frac{co - appearance\ frequency\ (s,t)}{total\ appearance\ frequency\ (s)}, \qquad (4)$$

where:

- *s* = source hashtag, *t* = target hashtag.
- *co-appearance frequency* = the number of posts in which both hashtags appear together.
- *total appearance frequency* = the number of posts in which a hashtag is used together with any other hashtags.

According to the edge weight formula, if only one hashtag is used in a post, it will not produce an edge to the graph. In a perfect scenario where #hashtag1 and #hashtag2 appear together in *N* posts but never appear together with any other hashtags, the *w(s,t)* of #hashtag1 to #hashtag2 and vice versa equal to 1.0. The *w(s,t)*=1.0 suggests a strong connection, where the audience of #hashtag1 will most likely visit #hashtag2 as well (and vice versa). Meanwhile, *w(s,t)*=0.5 suggests a moderate connection, where the audience of #hashtag1 has a 50% chance to visit #hashtag2 (and vice versa) because both hashtags have connections to other hashtags as well. Lastly, *w(s,t)*=0 suggests that both hashtags have never appeared together.

D. IM BASELINES

The IM approach was implemented using the hashtags network and validated using two types of benchmarks. Influence spread and runtime [3] were used as the *synthetic benchmarks*, whereas hashtag's ER and UER were used as the *engagement benchmarks*. In each experiment, the state-of-the-art IM algorithms, namely SSA (Stop-and-Stare) [20], DSSA algorithm (Dynamic SSA) [20], and IMM (Influence Maximization via Martingale) [21] algorithms were simulated under IC (Independent Cascade) and LT (Linear Threshold) [3] models.

The IM algorithms were compared against the top-ranked hashtags by usage count (labeled *TopUsage* in the upcoming sections). For example, for *k*=3 (where *k* is the number of seeds), the SSA algorithm under the IC model produced the following seeds, i.e., #malaysia, #love, #throwback. For *k=3*, based on *TopUsage*, the top three hashtags are #malaysia, #love, #kualalumpur (refer to Table 2). Note that *TopUsage* doesn't involve any computation from IM algorithms. For the synthetic benchmarks, these hashtags were simulated under IC or LT model. As for the engagement benchmarks, the ER and UER of these hashtags were calculated.

For the fairness of the engagement benchmarks, it is important to note that the hashtags graph does not provide any hints of engagement rate since the graph was built using hashtags relation. Thus, the benchmark will prove the effectiveness of the hashtags network in producing more engaging hashtags.

IV. DATA ANALYSIS

The most popular hashtags are shown in Table 2. Some hashtags have low UER even with the high usage count, such as #malaysia, #kualalumpur, #repost, #sayajual. Upon inspection, these hashtags were used to promote micro brands or for spamming (overusing hashtags without a proper context). Another suspicious thing is that these hashtags were overused by a small number of people. The presented facts show that promoting an authentic brand using spammy hashtags is bad for visibility since a post can be quickly buried among spam posts.

As seen in Table 2, the indication of spammy hashtags is the low UER. Similarly, a previous study [22] also suggested that a low engagement rate indicates fake users, along with other factors such as a high number of posts and followees. Figure 1 shows the relation between UER and POS and FLG, where POS is the average number of posts of the *unique users*, and FLG is the average number of followees of *unique users*. *Unique users* are users that uploaded a post containing the hashtag.

Table 2. Most Popular Hashtags (by Use Count)

Hashtag	Use count (posts)	Use count (users)	Related hashtags*	UER %	ER %
malaysia	18,264	4,141	169,259	1.76	7.69
love	8,710	2,516	129,904	3.27	8.84
kualalumpur	7,456	1,685	70,906	1.30	6.58
photography	7,423	1,667	91,430	3.41	10.55
instagood	6,140	1,213	74,821	2.61	9.76
throwback	6,071	3,036	78,264	2.21	10.74
photooftheday	5,711	1,151	72,123	2.53	9.32
travel	5,544	1,598	87,592	2.71	9.59
repost	5,223	1,657	32,511	1.79	5.65
ootd	5,142	1,558	54,127	3.98	9.65
instagram	4,453	1,019	56,244	2.76	9.42
art	4,276	1,156	66,254	2.29	10.16
sayajual	4,190	635	16,609	0.81	3.35
fashion	3,733	893	50,745	1.77	8.03
nature	3,519	1,325	76,693	2.53	10.52

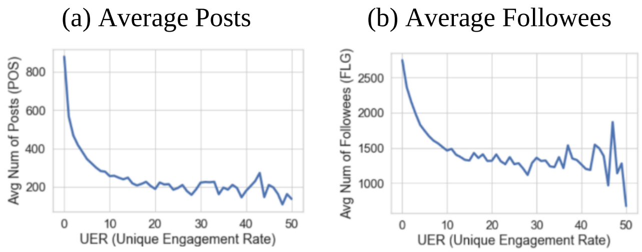
(a) Average Posts (b) Average Followees

Figure 1. Other Indicators of Spammy Hashtags, i.e., the Number of Posts and Followees

As seen in Figure 1, hashtags with low UER have higher POS and FLG, which indicates that they are usually posted by fake or spammy users. Note that there is a limited number of data with high UER (such as 2,254 data for UER>=30%), which caused a higher plot oscillation. Another user's metadata that can be included in this analysis is the number of followers. However, the number could be unfair since it can go up to millions, whereas FLR is just between 0 to 7,500 due to Instagram limitation.

V. EXPERIMENTAL RESULT

This section presents the setup, synthetic benchmarks, and engagement benchmarks for the Influence Maximization using the hashtags network.

A. SYNTHETIC BENCHMARK

The influence spread benchmark under IC and LT models are presented in Figure 2. The IMM algorithm struggled during *k*=1 to 20. This was because we modified the edge weights for IC and LT models, in which IMM might not adapt very well. The SSA and DSSA algorithms performed similarly with *TopUsage*, which means they can adapt to the hashtags network. However, since the influence spread is theoretical and not user-aware, it can't be used to show the effectiveness of our approach.

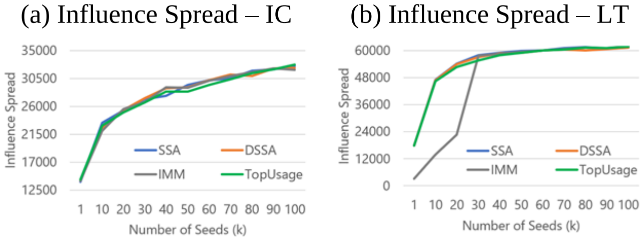
(a) Influence Spread – IC (b) Influence Spread – LT

Figure 2. Influence Spread under IC and LT models

The runtime is shown in Figure 3. SSA and DSSA algorithms have a significantly faster runtime. A fast runtime is essential since it implies the speed of finding the best hashtags in a practical scenario.

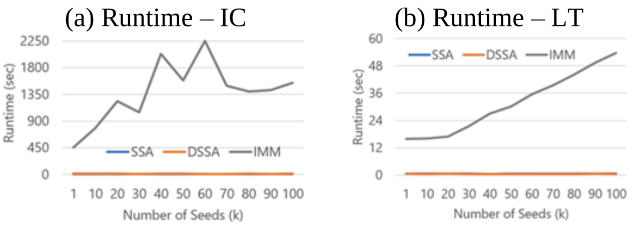
(a) Runtime – IC (b) Runtime – LT

Figure 3. Runtime under IC and LT models

B. ENGAGEMENT BENCHMARK

The hashtags ER benchmark is shown in Figure 4. The ER benchmark shows the effectiveness of the hashtags IM approach. Based on the average difference of ER from *k*=1 to *k*=100, the IM algorithms performed better than *TopUsage*, as shown in Table 3. SSA has the best performance, which outperformed *TopUsage* by 2.76% under IC model and 2.59% under LT model. Note that the *average difference* is simply formulated as follows:

$$dif(IM,TU) = \frac{100}{kmax}\, x \sum_{k=1}^{kmax} \frac{M(IM,k)-M(TU,k)}{M(TU,k)}, \qquad (5)$$

where:

- *IM* = The IM algorithm, *TU* = *TopUsage*
- *k* = number of seeds, *kmax* = maximum *k* (=100)
- *M* = the chosen metric (ER, HER, or influence spread)

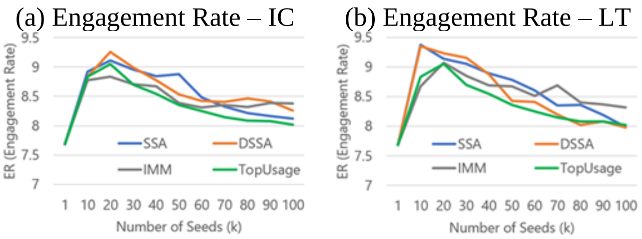
(a) Engagement Rate – IC (b) Engagement Rate – LT

Figure 4. Engagement Rate (ER) of Hashtags under IC and LT models

Table 3. Average ER Difference between the IM Algorithms and TopUsage (k=1 to 100)

Diff. Model	SSA	DSSA	IMM
IC	2.76%	2.59%	2.10%
LT	2.59%	2.31%	2.42%

As previously mentioned, UER is a fairer metric for engagement since it punishes the overuse of hashtags. The hashtags UER benchmark is shown in Figure 5. As seen in the graph, the UER of the IM algorithms becomes significantly different from *TopUsage* as the number of seeds increase.

The average difference of UER between the algorithms and *TopUsage* is shown in Table 4. The best performance is produced by DSSA under LT model, with a 17.23% average UER difference with *TopUsage*. The performance gap between DSSA under IC and LT model suggests that the LT model is more useful in a practical scenario.

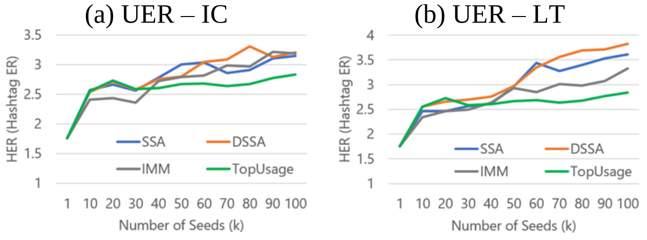

Figure 5. Unique ER (UER) of Hashtags under IC and LT models

Table 4. Average UER Difference between the IM Algorithms and TopUsage (k=1 to 100)

Diff. Model	SSA	DSSA	IMM
IC	7.17%	7.83%	3.90%
LT	12.07%	17.23%	6.15%

C. PRODUCED HASHTAGS

The outcome of the IM implementation on the hashtags network is a set of *k* hashtags, depending on the chosen *k* (number of seeds). As proven in Section 5.2, hashtags with a high usage count (*TopUsage*) don't always have the highest engagement. The best set of hashtags, based on the UER benchmark was produced by the DSSA algorithm.

The difference of the hashtag outputs can be seen in Figure 6, where *TopUsage* is used as the reference, and the graph shows the percentage of hashtags overlap with *TopUsage*. For simplicity, only the result of DSSA is shown, as the best performer in the UER benchmark. On average, from *k*=1 to 100, DSSA-IC and DSSA-LT have 71.72% and 70.32% overlap with *TopUsage*, respectively. This finding suggests that DSSA has around 29% different output with *TopUsage*, which implies the effect of choosing hashtags using *TopUsage* vs. DSSA in a practical scenario.

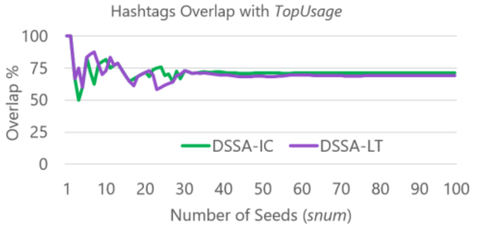

Figure 6. Hashtags Overlap with *TopUsage*

D. HASHTAGS CATEGORIZATION

The previous sections were focused on finding the best hashtags globally. In a practical scenario, users need to use hashtags based on a topic. For example, a photograph can be posted with #photography as the main hashtag. To expand the list of hashtags, users need to find the related hashtags. This section shows the advantage of the IM algorithms over *Top Usage* in choosing hashtags by topic.

To get the *hashtag rank* from the IM algorithms, we executed SSA and DSSA under both IC and LT models with *k*=(the total number of hashtags). The hashtags were ranked according to the order of appearance as seeds. For example, #malaysia, #love, #photography are the top three hashtags according to DSSA under LT model. As for the *TopUsage*, the *hashtag rank* is based on the *usage count*.

We selected hashtags with a usage count of at least 50 (to exclude hashtags with very minimal related hashtags count), which will be called the *main hashtag*. There are 6,543 *main hashtags*. From each *main hashtag*, *N_related* most related hashtags will be chosen based on *hashtag rank* and *closeness*.

The *closeness* of each two hashtags is based on the highest *co-appearance frequency* (Section 3.3). To combine the *closeness* and the *hashtag rank*, a set of related hashtags are first selected based on *closeness* with the *main hashtag*, then be ordered based on *hashtag rank*, and the top *N_related* hashtags are the output. Note that we chose *N_related*=11 and *N_related*=30, since currently, Instagram allows a maximum of 30 hashtags in a post, while using 11 hashtags is the best for engagement rate [19].

The performance metric for this section is the average UER of all related hashtags of the main hashtags. The average UER results for each algorithm/model are presented in Table 5, and the average hashtags overlap with *TopUsage* is shown in Table 6. Note that *N_sample* works in conjunction with *N_related*. For example, *N_sample*=200 and *N_related*=30 mean that a set of 200 related hashtags are initially chosen based on *closeness*. These hashtags are then sorted by *hashtag rank*, and the top 30 hashtags are taken.

Table 5. Average UER in the Hashtags Categorization Setup

N_ related	*N_ sample*	*Top Usage*	SSA -IC	SSA -LT	DSSA -IC	DSSA -LT
11	100	5.34	5.35	5.41	5.69	5.67
	200	5.6	5.58	5.58	5.83	5.97
	300	5.83	5.82	5.83	5.95	6.11
	400	5.96	5.95	5.97	6.05	6.26
	500	6.03	6.03	6.05	6.11	6.32
30	100	5.45	5.43	5.39	5.47	5.5
	200	5.75	5.76	5.8	5.94	5.95
	300	5.75	5.75	5.81	6.08	6.14
	400	5.82	5.82	5.86	6.11	6.22
	500	5.89	5.88	5.92	6.13	6.25

Table 6. Average Hashtags Overlap with *TopUsage*

N_ related	*N_ sample*	SSA -IC	SSA -LT	DSSA -IC	DSSA -LT
11	100	97.11	88.25	68.63	60.15
	200	96.45	89.11	74.86	62.45
	300	96.59	89.4	78.03	64.43
	400	97.1	89.56	79.35	66.36
	500	97.35	89.47	79.47	67.21
30	100	97.69	88.58	70.32	62.6
	200	97.34	87.97	70.4	61.97
	300	97.22	88.38	71.54	63.26
	400	97.14	88.76	73.44	64.77
	500	97.11	89.19	74.83	65.53

As seen in Table 5, the performance of the SSA algorithm is mostly similar to *TopUsage*. Meanwhile, DSSA produced the best performance, with DSSA under LT model as the best performer. The DSSA-LT outperformed *TopUsage* by 6.61% in (*N_related*=11, *N_sample*=200), and 6.87% in (*N_related*=30, *N_sample*=400). This performance can be explained in Table 6, where it has the most different decision if compared to others.

VI. CONCLUSION

This study aims to identify influential hashtags, which potentially generate a higher engagement rate and less used for spamming. The Unique Engagement Rate (UER) is used as the primary metric to benchmark the IM approach, which is the total unique likers divided by total unique followers of the users that used a hashtag. Some indications of spammy hashtags are low UER value, a high average number of followees and posts. The IM approach has a better tendency to avoid such hashtags, which can be seen in the UER benchmark.

The relation between hashtags in the hashtags network was based on how often two hashtags appear together. In terms of the influence spread, the IM algorithms have similar performance with *TopUsage* (top hashtags based on the usage count), which indicates that the IM approach can adapt to the hashtags network. In terms of engagement benchmark, DSSA-LT has the best performance, which outperformed *TopUsage* by 17.23%.

In a practical scenario, Instagram users need to find a set of hashtags which is related to the post. In the Hashtags Categorization (Section 5.4), each IM algorithm/model identifies the set of hashtags related to several main hashtags. DSSA-LT outperformed *TopUsage* by up to 6.87% in the hashtags categorization.

This study has successfully proved the effectiveness of IM using a hashtags network. The outcome of the IM approach is a global set of hashtags or a categorized set of hashtags that can be used for an Instagram post for a better engagement rate. This will help business or ordinary users to increase exposure. In future work, IM technique can also be implemented in various other scenarios, such as finding fake users, finding posts similarity, or for a recommendation engine.

References

[1] U. Can and B. Alatas, "Big Social Network Data and Sustainable," *Sustainability*, vol. 9, no. 11, p. 2027, 2017. https://doi.org/10.3390/su9112027.

[2] I. Himelboim, "Social Network Analysis (Social Media)," in *The International Encyclopedia of Communication Research Methods*, John Wiley & Sons, Inc., 2017. https://doi.org/10.1002/9781118901731.iecrm0236.

[3] D. Kempe, J. Kleinberg and É. Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2003, pp. 137-146. https://doi.org/10.1145/956750.956769.

[4] Y. Li, J. Fan, Y. Wang and K.-L. Tan, "Influence Maximization on Social Graphs: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1852 – 1872, 2018. https://doi.org/10.1109/TKDE.2018.2807843.

[5] N. L. Khalid, S. Y. Jayasainan and N. Hassim, "Social media influencers – shaping consumption culture among Malaysian youth," in *SHS Web of Conferences Vol. 53*, 2018. https://doi.org/10.1051/shsconf/20185302008.

[6] G. D. Saxton and R. Waters, "What do stakeholders like on Facebook? Examining public reactions to nonprofit organizations' informational, promotional, and community-building messages," *Journal of Public Relations Research*, vol. 26, no. 3, pp. 280-299, 2014. https://doi.org/10.1080/1062726X.2014.908721.

[7] Y. Kim and J. Seo, "Detection of Rapidly Spreading Hashtags via Social Networks," *IEEE Access*, vol. 8, pp. 39847-39860, 2020. https://doi.org/10.1109/ACCESS.2020.2976126.

[8] B. Bashari and E. Fazl-Ersi, "Influential post identification on Instagram through caption and hashtag analysis," *Measurement and Control*, vol. 53, no. 3-4, p. 409–415, 2020. https://doi.org/10.1177/0020294019877489.

[9] C. J. Qian, J. D. Tang, M. A. Penza and C. M. Ferri, "Instagram Popularity Prediction via Neural Networks and Regression Analysis," in *IEEE Transactions on Multimedia 19.11*, pp. 2561-2570, 2017. https://doi.org/10.1109/TMM.2017.2695439.

[10] F. Gelli, T. Uricchio, M. Bertini, A. D. Bimbo and S.-F. Chang, "Image Popularity Prediction in Social Media Using Sentiment and Context Features," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 907-910, 2015. https://doi.org/10.1145/2733373.2806361.

[11] P. A. Rauschnabel, P. Sheldon and E. Herzfeldt, "What motivates users to hashtag on social media?," *Psychology & Marketing*, vol. 36, no. 5, pp. 473-488, 2019. https://doi.org/10.1002/mar.21191.

[12] A. Laucuka, "Communicative Functions of Hashtags," *Economics and Culture*, vol. 15, no. 1, pp. 56-62, 2018. https://doi.org/10.2478/jec-2018-0006.

[13] P. Ferragina, F. Piccinno and R. Santoro, "On Analyzing Hashtags in Twitter," in *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pp. 110-119, 2015.

[14] C. Zhang, Z. Gao and X. Liu, "How Others Affect Your Twitter #hashtag Adoption? Examination of Communitybased and Context-based Information Diffusion in Twitter," in *IConference 2016 Proceedings*, 2016.

[15] M. Code, "Instagram, Social Media, and the "Like": Exploring Virtual Identity's Role in 21st Century Students' New Socialization Experience," Faculty of Education, Brock University, St. Catharines, Ontario, 2015.

[16] M. Li, X. Wang, K. Gao and S. Zhang, "A Survey on Information Diffusion in Online Social," *Information*, vol. 8, no. 4, p. 118, 2017. https://doi.org/10.3390/info8040118.

[17] F. Masood, G. Ammad, A. Almogren, A. Abbas, H. A. Khattak, I. U. Din, M. Guizani and M. Zuair, "Spammer Detection and Fake User Identification on Social Networks," *IEEE Access*, vol. 1, no. 1-14, p. 7, 2019. https://doi.org/10.1109/ACCESS.2019.2918196.

[18] M. Mohammadrezaei, M. E. Shiri and A. M. Rahmani, "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms," *Security and Communication Networks*, 2018. https://doi.org/10.1155/2018/5923156.

[19] K. R. Purba, D. Asirvatham and R. K. Murugesan, "Analysis and Prediction of Instagram Users Popularity using Regression Techniques based on Metadata, Media and Hashtags Analysis," *Engineering Letters*, vol. 28, no. 3, pp. 812-819, 2020. https://doi.org/10.31219/osf.io/uezyk.

[20] H. T. Nguyen, M. T. Thai and T. N. Dinh, "Stop-and-Stare: Optimal Sampling Algorithms for Viral Marketing in Billion-scale Networks," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 695-710. https://doi.org/10.1145/2882903.2915207.

[21] Y. Tang, Y. Shi and X. Xiao, "Influence Maximization in Near-Linear

Time: A Martingale Approach," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, Melbourne, ACM, 2015, pp. 1539-1554. https://doi.org/10.1145/2723372.2723734.

[22] K. R. Purba, D. Asirvatham and R. K. Murugesan, "Classification of instagram fake users using supervised machine learning algorithms," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2763-2772, 2020. https://doi.org/10.11591/ijece.v10i3.pp2763-2772.

__KRISTO RADION PURBA__ is currently a computer science PhD student at Taylor's University Malaysia, starting from 2018. His research interests are in artificial intelligence, machine learning, and social network influence maximization. Prior to joining Taylor', he was an informatics lecturer at Petra Christian University, Indonesia for 4 years (2014-2018), and also a contracted programmer at EHS (Environment, Health and Safety) department at PT. HM. Sampoerna, Tbk, Indonesia (2013-2017). He is also active mobile apps, games, websites developer since 2008 until now.