International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 3157-3164

User Influence on Twitter Hashtags Evolution: A Case Study from Career Opportunities Groups

Layal Abu Daher[1], Rached Zantout[2], Islam Elkabani[3]

[1]*Mathematics & Computer Science Department, Beirut Arab University, Beirut, Lebanon*

[2]*Associate Professor, College of Engineering, Rafik Hariri University, Mechref, Lebanon*

[3]*Assistant Professor, Mathematics & Computer Science Department, Beirut Arab University, Beirut, Lebanon*

Abstract

Hashtags on Twitter are used to make groups, initiate conversational subjects, share life events and post personal experiences. Due to the enormous growth of social media networks, researchers conducted many studies on the evolution of groups on such networks to predict how they retain their members, attract new ones and grow over time. In this study, we focus on the factors causing people to participate in Twitter hashtags, which in turn affects the evolution of these hashtags. In order to study this evolution, a dataset of some Twitter hashtags related to career opportunities was collected for a period of two months between February and March 2016. The collected dataset allowed us to study the reciprocal effect of the users' topological features and their activity levels. In this paper, three new measures are introduced (two influence measures and one topological measure). Those measures, in addition to the measures available in the literature, are used to spot the measures that can be used for influencing a user to attract other users to a certain hashtag.

Keywords: Evolution; Influence; Measures; Ranking

INTRODUCTION

Social media is the trendiest field of research nowadays. These social networks represent a social structure of people forming social relationships. These social networks consist of nodes and edges where the relationship between them is either directed or undirected depending on the type of the social network graph. Facebook, for example, is an undirected network graph because two users are friends on Facebook if one accepts the friend request of the second. In this paper, we concentrate on the factors affecting users to participate in hashtags on Twitter, one of the most commonly used and investigated online social network [1]. This study is conducted to focus on the measures affecting the evolution of some Twitter hashtags and whether the activity of users or the content of the post are affecting the evolution of such hashtags.

Twitter is an online news and social networking platform where users post and interact using tweets, which are messages of up to 280 characters. The relationship created between users does not need to be reciprocal. According to the study conducted in [2], only 22% of connections between users on Twitter are reciprocated. Thus, an incoming directed edge to a node from another node shows that the latter is following the first causing the representation between Twitter users to be directed. A user can retweet a tweet posted by another user by copying that tweet and sharing it with his followers. The most popular means of interaction on Twitter is "hashtags", a collection of characters preceded by the (#) symbol, which is used to create online communities from different users. Users create such hashtags on Twitter seeking entertainment, expressing political points of view, criticizing or complaining from economic situations and even seeking job opportunities through these hashtags. For example, #Job was one of the top trendy hashtags in the period between February and March 2016. Twitter users posted millions of tweets on #Job and other related hashtags. Some of these hashtags survived, others faded when users stopped posting.

In this paper, we conducted an analytical study on a dataset collected from some Twitter hashtags to study the factors affecting their evolution. Two topological measures were adopted, namely, Betweenness centrality and closeness centrality and three activity measures, namely, general activity, topical signal and signal strength. A new topological measure is introduced in this paper, namely Jaccard weighted in-degrees seeking for a better topological measure indicator. These three topological measures along with the three activity measures are used to study the influence of users' topology and activity level on the evolution of Twitter hashtags under study. In order to set a standard to compare with for identifying the users' influence based on the previously described measures, we introduced two novel influence measures, namely Count of Posting Followers and Users Being Retweeted..

The paper is organized as follows: in Section II, related work is discussed; in Section III, the approach is discussed focusing on hashtags selection and data representation in addition to measures identification and ranking; in Section IV, the results are presented and discussed and finally, conclusions and future work are presented in section V

RELATED WORK

Online Social Networks (OSNs) have forced themselves to be a part of our daily regular tasks, a phenomenon that lead

researchers to study online social networks in all its aspects. A wide range of studies is directed towards evolution, prediction of users' participation, evaluation of influential nodes and graphing the results of users on such OSNs.

The authors in [3] explored whether a topic will be interesting for users through building association rules to predict the participation of users on public groups on Facebook. Association Rule Learning was adopted to discover the relationship between users interacting on posts on Facebook groups. The prediction revealed a high level of accuracy concerning user participation on Facebook groups.

During 2016, an interesting study [4] examined the changes exhibited on Twitter graph in addition to the behavior of users since the first comprehensive study of Twitter done in [2] in 2010. The study revealed that Twitter popularity and usage increased 10-fold and the increase in reciprocal edges between nodes showed that 12.5% of 2009 Twitter users have left Twitter by the time the study of 2016 was done. The authors pointed out the decrease of network connectivity and the movement of edges to popular users on the network. Severe changes to influential users on Twitter has been also noticed where non-popular users in the past are considered the most popular and influential by the time this study was conducted.

Another study done in [5] investigated user influence dynamics across topics and time. It measured user influence in Twitter Social Network by comparing influence measures which are: in-degree, retweets and mentions. The results of this study showed that most influential users hold significant influence over a variety of topics on Twitter. This means that users in the network can be influential if employed in different topics, and also influential if they focus on a single topic using smart, insightful and creative posts that are spotted as valuable posts to other users on the network. The outcomes of the study showed that influential users might be of added value to the marketing, the political and the emotional sectors if users were active on them, and ordinary users with even single post, might be influential enough to conceive other users with a certain point of view.

In a study done in [6], the authors used a dataset of millions of tweets and thousands of hashtags. They focused on content based prediction of the spread of ideas in microblogging communities and captured the acceptance of a hashtag by the count of its appearance in a time interval, and as a result, they predicted its frequency after some time. The results of this study revealed three main factors to accept an idea in a hashtag: the content of an idea, the context within an idea and the social graph. The advantage of this study is because it is the first to focus on the content features in addition to the framework presented, which in turn efficiently modeled the exposure and acceptance of an idea in Twitter hashtags.

The study conducted in [7] shows that the process by which people come together, call for new members, and change over time is a central subject of study in social science such as political actions and religious movements. Thus, analyzing social networks leads to understanding the structure and dynamics of social groups.

Concerning the area of group evolution and the area of influential users prediction, various studies were conducted. The authors in [8] introduced a new measure of group proximity in OSNs. They estimated the link cardinality between groups. This helped in understanding the evolution of these groups by using friendship links across groups and common membership.

Another study done by [9] inspected group evolution through studying the prediction of group stability in OSNs. Their aim was to predict if a group on OSNs will remain stable or shrink over time. The analysis was done on a dataset collected from the "World of Warcraft" multiplayer online game where users interact.

In the same area of study, the authors in [10] conducted experimental studies on four evolving social networks in which ten different classifiers were used. They used the GED method used in [11] to discover group evolution where inclusion, as a measure, was the most important component. This measure allowed evaluating the inclusion of one group in another depending on the group quantity (what portion of members from the first group is in the second group) and quality (what contribution of important members from the first group is in the second group). Such a measure shows a balance between the groups containing many of less important members and the groups of only few but key members. The authors could extract the sequence of group sizes and events between time frames through following the GED method steps.

Moreover, the study in [12] analyzed the disappearance of a group of nodes in a social network. This was important to avoid the loss of information that could follow the disappearance of a node and to identify a substitute to its disappearance whenever it is critical in terms of its centrality measures. In other words, a potential substitute of the disappearing node along with the new links is found between individuals that share common characteristics such as values, education and beliefs. Such an approach adds enough links to maintain the quality of the information flow within the network as it was before a node deletion. The consequences of the steps are: success and maintaining the network connected after the disappearance of a set of nodes, reasonable execution times after a node deletion, and constant quality of the information flow.

Focusing on the prediction of community evolution, in [13] the process of group prediction was analyzed by conducting experimental studies on three different datasets: Facebook, DBLP and Salon24-Polish blogosphere. Four main bases were covered in the study: the first phase included data collection and splitting data into time frames. The second phase consisted of classifying the social networks for each period and identifying the social community. The third phase concentrated on identifying the changes or events in a group and detecting the chains preceding the recent state of the group. The last phase was characterized by building the predictive model. This was done by learning the classifier and validating it. In addition two algorithms were used: Stable Group Changes Identification (SGCI) and Group Evolution Discovery (GED). The experimental studies on the three different databases revealed that as the length of the evolution chain increases, the accuracy of prediction is improved. This means that, more comprehensive history with more information about the group is detected by longer evolution chains. The best results for SGCI method using either Random Forest or AdaBoost classifier were recorded for

Salon24 dataset with evolution chains of length 6 and longer. As to GED method, the best predictive ability was also recognised for Salon24 dataset with evolution chains of length 7 and using the same classifiers used in SGCI method. As to influential users' prediction in online social networks, various studies were conducted on different online social networks based on comment mining as in [14] and predicting information cascade as in [15]. The important factor in different online social networks is the user's role and thus, influential users' identification is of great interest in the field of social network analysis.

As to the centrality of a user on online social network graph, it determines the importance of that user on the network; how and to whom this user is connected. Centrality measures used in the literature reviews are many with closeness and betweenness centrality measures being the most popular. Closeness calculates the minimum sum of the shortest paths from a node to all other nodes in a network [16]. Betweenness calculates the number of shortest paths passing through a vertex [17].

APPROACH

In the following sections, the selection of hashtags under study is presented, in addition to the collection of dataset. Moreover, the representation of our dataset will also be denoted in order to visualize the network graph, in addition to the identification of the measures that will be used in our analytical study.

*A. **Hashtags Selection***

In order to select the hashtags for our analytical study, "hashtagify.me" website *[18]* is visited. This website consists of the top trendy hashtags during the period covering February and March 2016. This website does not only provide a list of the top trendy hashtags, but it also provides a list of correlated hashtags to each hashtag in the list of the trendy hashtags. These correlated hashtags represent the hashtags that are concurrently occurring with the associated hashtag in the same post. One of the top trendy hashtags in the above mentioned period was #Job. For #Job, "hashtagify.me" website provided the percentage of correlation of other ten correlated hashtags as represented in *Fig. 1*.

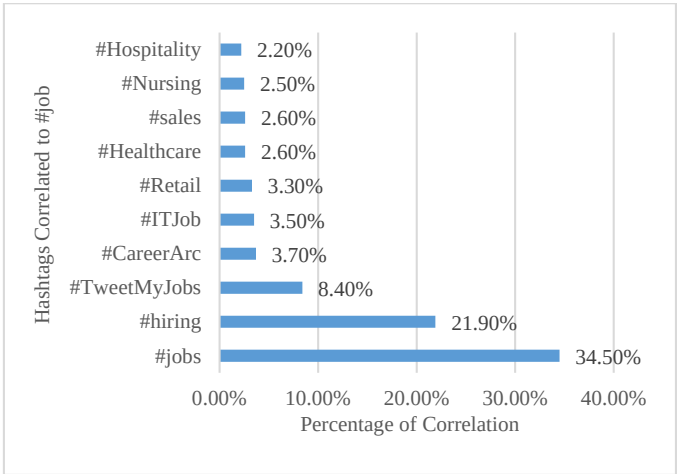

Figure 1: The % of the 10 correlated hashtags to #Job

The top hashtags that have the highest percentage of correlation with #Job are #Jobs and #Hiring. This is why they are chosen (with #Job) to build the corpus for this analytical study. The other remaining eight hashtags are not considered in our analytical study due to the low percentage of correlation and the inadequate number of Tweets and Retweets as represented in Table I. The three chosen hashtags are also the most general ones. This means that these hashtags are not restricted to any field in the job market, however, they are general enough to contain job seekers from different fields resulting in a diversity of users posting generally about careers and jobs opportunities.

Table I. Number of Tweets and Retweets in the 11 #s

Hashtags under study	Tweets	Retweets
#job	37791	2601
#jobs	33312	2385
#hiring	19383	1297
#TweetMyJobs	2899	54
#CareerArc	2936	35
#ITJob	1254	194
#Retail	7356	430
#Healthcare	2414	556
#sales	2731	224
#Nursing	2728	242
#Hospitality	2894	76

B. Data Collection

In order to collect the dataset from the selected hashtags, a crawler designed in [19] is used. This crawler downloads Tweets and Retweets on specific hashtags on Twitter. The downloaded data contains basic information about the users posting on the hashtags such as: original tweet text, retweet text, number of followers of the posting users. We also implemented another crawler to collect the list of followers of the posting users. The collected data is parsed and made available in SQL database for preprocessing. The collected dataset of the hashtags under study constitutes of 21199 unique users with 115698 tweets and 8094 retweets as shown in Table I. 11.05 % (2344 out of 21199) of the users posting on these hashtags has private profiles where we could not crawl their profiles, thus they were ignored in the preprocessing phase.

C. Data Representation

The collected dataset represents a directed graph where the nodes are the distinct posting users and the links are the relationships between the posting users. Each link between two nodes has three attributes: the source, which is the initial node; the target, which is the target node; and the weight of the link. In our study, we considered that there is a directed edge from user A to user B if user A is found in the list of followers of user

B. As to the weight of edges between the target and the source nodes, we considered the Jaccard's similarity coefficient presented in [20] as the weight of an edge. The Jaccard's Similarity between two nodes A and B is calculated using equation 1.

$$J(A,B) = \frac{\text{Followers(A)} \cap \text{Followers(B)}}{\text{Followers(A)} \cup \text{Followers(B)}} \quad (1)$$

D. Identifying Measures

In this study, three *Activity Measures* and two centrality measures are used from the literature. Two new *Influence Measures* are introduced to complement the *Activity Measures* and the *Topological Measure*s. Additionally, a new *Topological Measure* is also introduced to determine the user's level of social connectedness to users posting on the same hashtags.

1) Influence Measures

We propose two influence measures that we based our comparison on. The first measure is the *Count of Posting Followers measure (CPF)* that finds the total number of followers of user A who are posting on the same hashtag where user A is posting as shown in equation 2.

$$CPF = \sum_{i \in F(A)} x_i \quad (2)$$

where

F(A) is the set of followers of User A

$x_i = 0$ if there are no posting followers of User A

$x_i = 1$ if there is one or more posting follower of User A

The second proposed measure is the *Users Being Retweeted (UBR)* that sums the Retweets of the tweets of user A as depicted in equation 3.

$$UBR(A) = \sum_{t \in T(A)} RT(t) \quad (3)$$

where

T(A) denotes the set of tweets by user A

RT(t) denotes the number of retweets of a tweet t

2) Activity Measures

Active users are the ones who contribute with real participation on the online social network through either posting or sharing. There are numerous twitter users who are very active readers, however their activity will be neither spotted nor counted on the network without real participation of posting.

The number of tweets of user A is the simplest activity measure, where it counts the number of tweets that a user posts on a certain hashtag. Another simple activity measure is the number of retweets done by user A. Thus, a reasonable activity measure is the sum of all the visible actions of a user and is referred to as *General Activity* as stated in [21] and illustrated in equation 4.

$$\text{General Activity (A)} = T + RT \quad (4)$$

where

T is the number of tweets of User A

RT is the number of Retweets of user A

The *Topical Signal* as introduced in [22], is another activity measure used in our work which is the general activity of a certain active user divided by the total number of tweets in a specific hashtag as shown in equation 5.

$$TS(A) = \frac{(T+RT)_{|specific\ Hashtag}}{Total\ Number\ of\ Tweets} \quad (5)$$

A third activity measure used in our study and also introduced in [22] is *Signal Strength* which determines how strong is the *Topical Signal* and it is as shown in equation 6, the ratio of the original tweets by the summation of original tweets and retweets of a user.

$$SS(A) = \frac{T}{T+RT} \quad (6)$$

3) *Topological Measures*

Two types of *Topological Measures* are used in this paper, *Centrality Measures* and *Jaccard's Weighted In-degree* measure *(JWI)*. In our study, we used the *Betweenness* and *Closeness Centrality* measures to study the relationship between the centrality of the user in a social network and his influence on the evolution of a certain hashtag.

JWI is a new *Topological Measure* introduced in this paper as a better measure for the social connectedness of a user in Twitter. *JWI* measure of a user is calculated by summing the Jaccard's similarity weights of all its incoming edges.

$$JWI(A) = \sum_{i=1}^{n} w_i\,(A) \quad (7)$$

where

w_i is the weight of link i ending on node A

E. Ranking Based on the Identified Measures

A social network between the nodes of users posting on same hashtags is created in such a way that there exist a directed edge between two users if one of the users is found in the list of followers of the second. For this social network, we calculate the *Influence Measures*, namely *CPF* and *UBR* illustrated in section 1. Based on these measures, ordered (descending) lists are created. Similar lists are generated for the *Activity Measures* explained in section 2. As to the *Topological Measures* mentioned in section 3, calculation of *JWI* measure is performed to prepare similar descending lists and Gephi software [23] is used to calculate the centrality measures. Finally, comparison is done between the top ranked users identified from the *CPF Influence Measure* and the top users identified from the *Activity Measures*. Moreover, the same comparison is done between the top ranked users identified from the *CPF Influence Measure* and the top users identified from the *Topological Measures*. In addition, comparison is done between the top ranked users

identified from the *UBR Influence Measure* and the top users identified from the *Activity Measures* and between the top ranked users identified from the *UBR Influence Measure* and the top users identified from the *Topological Measures*. In this section, we are going to present the most interesting results from the three hashtags #Job, #Jobs and #Hiring which were chosen as our corpus for this analytical study.

RESULTS AND DISCUSSION

We focused on studying relationships between the levels of *Activity Measures*, *Centrality Measures* and *JWI* on one hand and the levels of the two *Influence Measures*, *CPF* and *UBR* for different percentages of top users (25%, 50% and 75%). Our analysis focused on the 25% of top users and reinforced by the other percentages of top users.

A. Results

The results of the *CPF* intersection with the level of the three *Activity Measures*, namely *General Activity*, *Topical Signal* and *Signal Strength* for the three selected hashtags are presented in Table II. The *CPF* intersection with *Activity Measures* did not exceed 22% in #Job, 16% in #Jobs, and 36% in #Hiring as shown in the column labelled *25% intersection* in Table II. This means that the followers of a user are posting on the same hashtag but not necessarily due to the influence of the user's activity. Comparing the intersection results across different percentages of top users (i.e. 25%, 50%, and 75%), a logical observation can be made that it increases linearly.

Table II. Similarity between CPF and Activity Measures

***CPF* Intersection Results**						
#Job Results						
Activity Measures	25% ∩ 1636 Users		50% ∩ 3271 Users		75% ∩ 4907 Users	
	No.	%	No.	%	No.	%
General Activity	353	22	1571	48	3554	72
Topical Signal	353	22	1571	48	3554	72
Signal Strength	1	0.06	667	20	3649	74
#Jobs Results						
Activity Measures	25% ∩ 1154 Users		50% ∩ 2308 Users		75% ∩ 3461 Users	
	No.	%	No.	%	No.	%
General Activity	181	16	932	40	2463	71
Topical Signal	181	16	932	40	2463	71
Signal Strength	1	0.08	670	29	2676	77
#Hiring Results						
Activity Measures	25% ∩ 863 Users		50% ∩ 1727 Users		75% ∩ 2590 Users	
	No.	%	No.	%	No.	%
General Activity	309	36	925	54	1761	68
Topical Signal	309	36	925	54	1761	68
Signal Strength	68	8	739	43	2018	78

The results of the *CPF* intersection with the level of the three *Topological Measures*, namely *Betweenness Centrality*, *Closeness Centrality* and *JWI* for the three selected hashtags are shown in Table III. The *CPF* intersection with the *Betweenness* and the *Closeness Centrality* measures was between 48% and 77% in #Job, however it reached 95% in *JWI*. As to #Jobs, *CPF* intersection with the *Betweenness* and the *Closeness* Centrality measures was between 52% and 85% but the *JWI* reached 91%. The *CPF* intersection with the three *Topological Measures* was between 50% and 83% but the *JWI* reached 83% in #Hiring. This indicates that the more socially connected users are more likely to affect their followers to post on the same hashtags. Comparing the intersection results across different percentages of top users (i.e. 25%, 50%, and 75%), a logical observation can be made that it increases linearly in most of the cases.

Table IV shows the results of the *UBR* intersection with the level of the three *Activity Measures*, namely *General Activity*, *Topical Signal* and *Signal Strength* for the three selected hashtags. The *UBR* intersection with *General Activity* and *Topical Signal* were between 40% and 45% in #Job and #Jobs, whereas the *Signal strength* reached 86%. However, in #Hiring, the three *Activity Measures* reached 55% as shown in the column labelled *25% intersection* in Table IV. These results indicate that in #Job, the users' original tweets have more influence than the users' activity. Comparing the intersection results across different percentages of top users (i.e. 25%, 50%, and 75%), a logical observation can be made that it increases linearly in most of the cases.

Table III. Similarity between CPF and Topological Measures

***CPF* Intersection Results**						
#Job Results						
Topological Measures	25% ∩ 1636 Users		50% ∩ 3271 Users		75% ∩ 4907 Users	
	No.	%	No.	%	No.	%
Betweenness Centrality	782	48	2796	85	4338	88
Closeness Centrality	1264	77	2849	87	4731	96
JWI	1559	95	2933	90	4322	88
#Jobs Results						
Topological Measures	25% ∩ 1154 Users		50% ∩ 2308 Users		75% ∩ 3461 Users	
	No.	%	No.	%	No.	%
Betweenness Centrality	597	52	1893	82	2960	86
Closeness Centrality	976	85	2009	87	3272	95
JWI	1048	91	2015	87	2972	86
#Hiring Results						
Topological Measures	25% ∩ 863 Users		50% ∩ 1727 Users		75% ∩ 2590 Users	
	No.	%	No.	%	No.	%
Betweenness Centrality	434	50	1325	77	2451	95
Closeness Centrality	537	62	1434	83	2522	97
JWI	712	83	1620	94	2531	98

Table IV. Similarity between UBR and Activity Measures

***UBR* Intersection Results**						
#Job Results						
Activity Measures	25% ∩ 1636 Users		50% ∩ 3271 Users		75% ∩ 4907 Users	
	No.	%	No.	%	No.	%
General Activity	660	40	1881	58	4044	82
Topical Signal	660	40	1881	58	4044	82
Signal Strength	1412	86	2848	87	4411	90
#Jobs Results						
Activity Measures	25% ∩ 1154 Users		50% ∩ 2308 Users		75% ∩ 3461 Users	
	No.	%	No.	%	No.	%
General Activity	516	45	1616	70	2953	85
Topical Signal	516	45	1616	70	2953	85
Signal Strength	606	53	1740	75	2844	82
#Hiring Results						
Activity Measures	25% ∩ 863 Users		50% ∩ 1727 Users		75% ∩ 2590 Users	
	No.	%	No.	%	No.	%
General Activity	468	54	1139	66	2173	84
Topical Signal	468	54	1139	66	2173	84
Signal Strength	477	55	1325	77	2302	89

Table V. Similarity between UBR and Topological Measures

***UBR* Intersection Results**						
#Job Results						
Topological Measures	25% ∩ 1636 Users		50% ∩ 3271 Users		75% ∩ 4907 Users	
	No.	%	No.	%	No.	%
Betweenness Centrality	277	17	880	27	3502	71
Closeness Centrality	98	6	684	21	3420	70
JWI	0	0	650	20	3453	70
#Jobs Results						
Topological Measures	25% ∩ 1154 Users		50% ∩ 2308 Users		75% ∩ 3461 Users	
	No.	%	No.	%	No.	%
Betweenness Centrality	244	21	781	34	2595	75
Closeness Centrality	142	12	697	30	2415	70
JWI	107	9	737	32	2531	73
#Hiring Results						
Topological Measures	25% ∩ 863 Users		50% ∩ 1727 Users		75% ∩ 2590 Users	
	No.	%	No.	%	No.	%
Betweenness Centrality	198	23	676	39	1826	71
Closeness Centrality	178	21	664	38	1810	70
JWI	109	13	602	35	1818	70

Table V presents the results of the *UBR* intersection with the level of the three *Topological Measures*, namely *Betweenness Centrality*, *Closeness Centrality* and *JWI* for the three selected hashtags. The *UBR* intersection with *Topological Measures* did not exceed 23% in #Job, #Jobs and #Hiring. This indicates that the users who are more retweeted are not necessarily the ones who are socially connected. Comparing the intersection results across different percentages of top users (i.e. 25%, 50%, and 75%), a logical observation can be made that it increases linearly in most of the cases.

B. Discussion

The results shown in Tables II to IV revealed many interesting findings. First of all, the followers of a user posting on the same hashtag are not necessarily influenced by the user's activity level. This is obvious from *Fig. 2* where *General Activity* in the three hashtags did not exceed 36%.

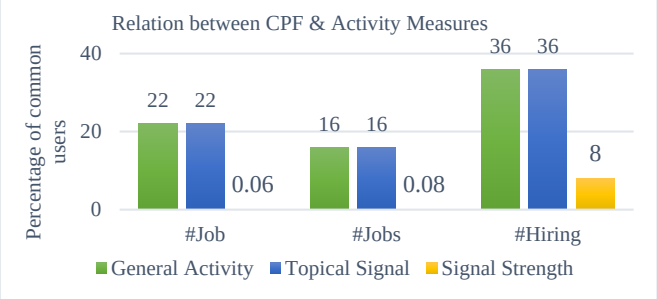

Figure 2. Activity Measures & CPF for Top 25% Ranked Users

Moreover, as presented in *Fig. 3*, among the three *Topological Measures*, the *JWI* measure gave more confidence regarding the influence of a user on his/her followers in the same hashtag. This was valid in all cases compared to *Betweenness* and *Closeness Centrality* measures. In other words, the level of the social connectedness of the user represented by *JWI* is a better indicator for the influence of the user on his/her followers than the other two *Centrality* measures adopted in this work. This proves that the *JWI* measure, which was newly introduced in this work, insures that as the social connectedness involving the quality and the number of connections one has with other people in a social circle increases, the number of followers posting on same hashtags also increases.

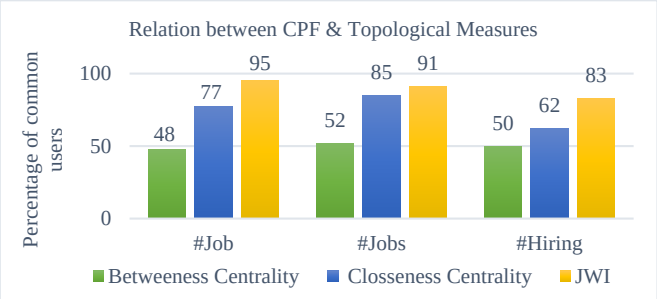

Figure 3. Topological Measures and CPF for the top 25% Ranked Users

Furthermore, in the three hashtags under study, and as depicted in *Fig. 4*, the level of social connectedness, measured by *JWI*, proved that it is a better indicator for the influence of a user on his/her followers to post on same hashtags, whereas it is not considered a good indicator for identifying users being retweeted. The results show that the most socially connected users were not retweeted but were among the users having the most posting followers.

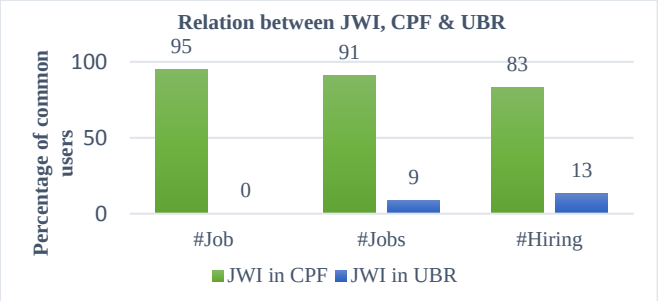

Figure 4. JWI, CPF & UBR for Top 25% Ranked Users

As to the results of intersection between *Signal Strength, CPF and UBR* for the three hashtags under study, we can conclude from *Fig. 5* that users who are posting original tweets are being retweeted more; however, they are not affecting their followers to post. This indicates that retweeting is more influenced by the content of the tweet, which is seeking for job vacancies in the market, more than being influenced by the level of activity of the user.

The vast majority of the intersections increased linearly with the percentage of top users. However, some outliers existed because there were many users having the same rank in the top 50% and 75%. This is why we focused on the top 25% in the analysis of the similarity comparison results.

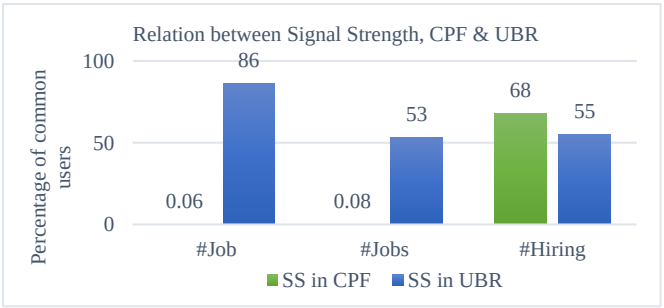

Figure 5. Signal Strength, CPF & UBR for the top 25% Ranked Users

CONCLUSION AND FUTUR WORK

In this paper, we investigate the factors affecting the participation of users in some selected Twitter hashtags. We introduced three measures which were proven to be better at measuring the influence and activity of a user in a hashtag. In order to draw conclusions about the behavior of users in Twitter, three hashtags were selected as our corpus. The information about users posting on those hashtags as well as tweets and retweets were used. Interesting findings are revealed where the level of activity is found not to affect the count of posting followers. Retweeting is more influenced by the content of the tweet than by the level of activity of the user. As to the new *JWI* measure, it is proven to be a better indicator for the influence of a user on his followers to post on the same hashtags. In the future, we intend to study the behavior of users in different online social networks in order to predict their participation in social networks communities. Moreover, identifying influential users using association rule learning will be a very interesting topic in this field.

REFERENCES

[1] Z. Tufekci, "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," in *In ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014.

[2] C. Lee, H. Kwak, H. Park and S. Moon, "What is Twitter, a social network or a news media?," in *WWW '10 Proceedings of the 19th international conference on World wide web*, Raleigh, North Carolina, USA, 2010.

[3] F. Erlandsson, A. Borg, H. Johnson and P. Bródka, "Predicting User Participation in Social Media," in *NetSci-X 2016 Proceedings of the 12th International Conference and School on Advances in Network Science*, Wroclaw, Poland, 2016.

[4] H. Efstathiades, D. Antoniades and G. Pallis, "Online social network evolution: Revisiting the Twitter graph," in *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016.

[5] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *In 4th International AAAI Conference on Weblogs and Social Media*, Washington, DC, 2010.

[6] O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining*, Seattle, Washington, USA, 2012.

[7] L. Backstrom, D. Huttenlocher and J. Kleinberg, "Group Formation in Large Social Networks: Membership, Growth, and Evolution," in *KDD '06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA, 2006.

[8] B. Saha and L. Getoor, "Group Proximity Measure for Recommending Groups in Online Social Networks," in *The 2nd SNA-KDD Workshop '08 (SNA-KDD'08)*, Las Vegas, Nevada , USA., 2008.

[9] A. Patil, J. Liu and J. Gao, "Predicting group stability in online social networks," in *Proceedings of the 22nd*

International Conference on World Wide Web (WWW'13), Rio de Janeiro, Brazil, 2013.

[10] P. Bródka, P. Kazienko and B. Kołoszczyk, "Predicting Group Evolution in the Social Network," in *International Conference on Social Informatics*, Berlin, Heidelberg, 2012.

[11] P. Bro´dka, S. Saganowski and P. Kazienko, "GED: the method for group evolution discovery in social networks," *Social Network Analysis and Mining,* vol. 3, no. 1, pp. 1-14, 2013.

[12] I. Sarr, R. Missaoui and R. Lalande, "Group disappearance in social networks with communities," *Social Network Analysis and Mining,* vol. 3, no. 3, p. 651–665, 2013.

[13] S. Saganowski, B. Gliwa, P. Bródka, A. Zygmunt, P. Kazienko and J. Koźlak, "Predicting Community Evolution in Social Networks," *Entropy,* vol. 17, no. 5, p. 3053–3096, 2015.

[14] S. Jamali and H. Rangwala, "Digging Digg: Comment Mining, Popularity Prediction, and So-cial Network Analysis," in *Proceedings of the 2009 International Conference on Web Infor-mation Systems and Mining*, Washington, DC, USA, 2009.

[15] M. A. Nur Hakim and M. L. Khodra, "Predicting information cascade on Twitter using support vector regression," in *In Proceedings of the 2014 International Conference on Data and Software Engi-neering (ICODSE)*, Hyderabad, India, 2014.

[16] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *Journal of the American Society for Information Science and Technology ,* vol. 60, no. 1, pp. 2107-2118, 2009.

[17] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks,* vol. 1, no. 3, pp. 215-239, 1979.

[18] "Hashtagify: Find, Analyse, Amplify," CyBranding Ltd., 2011. [Online]. Available: https://hashtagify.me/explorer/. [Accessed 1 March 2015].

[19] M. Hawksey, "The musing of Martin Hawksey (EdTech Explorer)," WordPress and Stargazer, 2011. [Online]. Available: https://mashe.hawksey.info/. [Accessed 10 November 2015].

[20] S. Niwattanakul, J. Singthongchai, E. Naenudorn and S. Wanapu, "Using of Jaccard coefficient for keywords similarity," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, China, 2013.

[21] F. Riquelme and P. González-Cantergiani, "Measuring user influence on Twitter: a survey," *Information Processing & Management,* vol. 52, no. 5, pp. 949-975, 2016.

[22] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, China, 2011.

[23] M. Bastian, S. Heymann and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Network," in *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009*, San Jose, California, USA, 2009.