Who Will Like the Post? A Case Study of Predicting Likers on Flickr

Samin Mohammadi, Reza Farahbakhsh, Noel Crespi
Institut Mines-Telecom, Telecom Sud-Paris, CNRS UMR 5157, France.
{samin.mohammadi, reza.farahbakhsh, noel.crespi}@it-sudparis.eu

***Abstract*—Reacting to a published post on a social media is one of the main activities of users which can happen in different forms comprising to like the post, leave a comment or reshare it. Finding a way to predict the size of users future interactions and more interestingly identifying the users who are going to react to a post are the two important research topics which benefit different domains from efficient advertising campaign to enhanced content delivery systems. In this paper, we aim to predict the users who are going to react to a newly published post in future. Toward this aim, we implement a novel approach based on Point-wise Mutual Information (PMI) which derives users latent similarities from their interactions log and exploits them to predict future interacting users. The proposed method is evaluated using a large dataset of Flickr including 2.3M users and 11.2M published photos. The empirical findings support the idea of employing interactions log to detect future likers of posts by achieving noticeable prediction results for the tested dataset. Moreover, the analysis of the prediction task implies that likers prediction for the photos of publishers with a high number of followers and engagements is more accurate than the other publishers photos.**

Index Terms Online Social Networks, User similarity, Point-wise mutual information, Liker, Prediction

I. Introduction

A great portion of the fast-growing research activities on social media has been devoted to the analysis of the data, which is available in these networks and more specifically the analysis of information propagation, users' characteristics, and engagements prediction [1] [2]. Users are the main actors of social networks who publish posts as well as reacting to the published posts by other users in various forms such as like, share or leaving comments. Users reacting to the posts on social media, are called textslreactors in this study. Reactors play a substantial role in information propagation and popularity of a post [3][4].

The total number of engagement on a post shows the number of reactors, also known as the popularity number. Predicting this value and its involved reactors are two significant prediction tasks, which supply valuable information for many applications such as providing better solutions for content placement in networks, more efficient advertisement campaigns, and providing accurate recommendations. Among the existing efforts on these two prediction tasks the first one, predicting popularity size, has been inspected many times [5] [6]. However, identifying the users, reacting to the post has been neglected.

The key aim of this study is to identify future reactors of a post using of the prior information acquired from users' interaction log. Towards this aim, we have implemented a framework based on Point-wise Mutual Information (PMI) inspired by the $Word2vec$ language model [7]. Word2vec is a language model which derives word embeddings considering the co-occurrence of words in a *window* of vocabularies of size w. The proposed model in this study exploits different lists of users who have reacted to the published post via a *like* (marking the post as favorite) called *like sequences*, and computes the engagement probabilities of users on a newly published post. Since the reaction type in this study is specialized by *like*, we refer to reactors by *likers* term from now on. Using like sequences, we consider the co-occurrence of users in a window of size w to measure point-wise mutual information between users. Considering users' co-occurrences in a window helps to discover the latent relation between them representing their similar preferences and favorite aspects which are not directly comprehensible from their friendships or profiles. In our method, PMI values show the strength of users' latent similarities in terms of their favorite contents.

We build a graph of users and their interactions, where nodes represent the users, and edges reflect the engagement probability of users on each of the other's posts. In order to build users graph, we consider three different approaches indicating three types of users graph which differ in the type of links between users (directed or undirected) and in how the weight of these links is computed. The computed PMI value between two users is assigned to the weight of the edge between them. Given a new published post, we use the created graphs to find l users possessing the strongest links to the post's publisher representing the future likers of that post. These l users are the *l-nearest-neighbors* to the publisher, who are selected based on the PMI values between them and their neighbors, which are the most probable users who will like the post in future.

Besides the prediction of likers merely based on the publisher, we assume the availability of a prior-knowledge about k early likers of a post in addition to its publisher in order to take advantage of this knowledge and improve prediction results. In this case, we choose the l-nearest-neighbors from the neighbors of all k early likers. Prediction results are compared for different k numbers.

The main contributions of this study are:

1) We propose a novel approach to identify users who will react to the post by extracting users' latent similarity without using handcrafted features.

2) Although the likers of a post are not limited to its publisher's friends, comparing the prediction results when future likers are chosen from *all neighbors* versus from *only friends* shows that friends' interactions are more predictable than those of non-friends.
3) We found that taking advantage of the *window* idea [7] to compute PMI values helps to predict more accurately.
4) Our experiments reveal that future likers of a post are more dependent on the publisher of the post than early likers.
5) We identified number of followers and number of engagements of publishers as the most important properties that provide a better success rate in predicting future likers.

The rest of this paper is organized as follows: Section II summarizes the relevant previous studies. The proposed methodology is presented in Section III. The evaluation results of prediction and characterizing the successful publishers are discussed in Sections IV and V, respectively and Section VI concludes the study and points avenues for future research.

II. Related Work

The previous studies relevant to this paper can be summarized in two main parts: (i) popularity prediction on Online Social Networks (OSNs) and (ii) PMI applications on OSNs and co-occurrence computation.

A. Prediction of Popularity

Prediction of popularity is an interesting research topic which can be investigated about users [8] or contents [9]. From the content perspective, once a content is published on a social network, it attracts different amount of users interactions depending on its interestingness, topic, publisher's reputation, published time and etc [10] [11]. Meanwhile, some contents are succeeded to attract more user engagements and become popular [12]. Popularity of a content usually assesses by different cascading metrics such as number of likes, shares, views, etc.

Predicting the trend of popularity for a content (which can be a text, video, or image) and more importantly identifying the users who are going to react to that content are very valuable information for different entities such as service providers to rank the content better [13], to early discover of trending posts, to improve recommendations and even to improve their content delivery networks and user experiences [14]. This kind of prediction tasks are mainly based on the features of contents and early adapters. Depending on the social network's type, adapters can be interpreted as either likers, resharers, viewers, or so on. In [15], popularity of a content is predicted using the structural diversity of early adapters. In other studies, temporal features of early adapters are realized as the most predictive features among different features of content, user and network [5] [6] [16].

Looking the models that have been developed on different content popularity prediction tasks on OSNs shows that most of them focused on predicting the popularity size of contents in future. There are very rare researches on identifying the users who are going to react to the contents published on OSN in future. Although, interactors prediction on OSNs is somehow similar to well-studied rate prediction on recommender systems (RS), but there is a main different which makes RS models improper to apply directly on interactors prediction on OSNs. Rate prediction models on RS are mainly based on favorness, whereas OSNs models are primarily based on friendship. From the few studies in this area on OSNs, Petrovic *et al.* [17] tried to predict interactors using a machine learning method based on the passive-aggressive algorithm.

B. Pointwise Mutual Information on OSNs

Point-wise mutual information (PMI) is a measure to model the dependency of two instances of random variables used widely in information theory, Natural Language Processing (NLP), Recommender Systems (RS) and OSNs. NLP models use PMI to find the strength of association between words [18] [19] [20]. In [21], PMI is used to compute semantic similarity and relatedness of words where it achieves outperforming results. RS also take advantage of PMI as one of the measures which used to find users and items similarities [22] [23]. Kaminskas *et al.* used PMI between items to measure surprise in RS and compared its results with a content-based surprise measurement [24]. In [25], authors get profit of PMI between different recipes' ingredients and predict recipe ratings given by web users. Spertus *et al.* compared different similarity metrics including PMI to compute similarity of Orkut communities in order to find users' interesting communities and exploit them in a recommendation task [26].

Social networks applications also benefit from PMI and use it for two primarily objectives, word and consequently content similarity, and user similarity. Different problems have been studied on OSNs using words' PMI metric such as content sentiment analysis, topic detection, content classification and so on [27] [28], but our focus in this study will be on users similarity. Authors in [29] exploited PMI to measure the network similarity of users based on their mutual friends on social networks. Following the aim of this study, we use PMI between users to find their interaction similarities. Our proposed model is inspired by Word2vec language model [30] to compute users co-occurrences. Akin to Wodr2vec which extracts word-context pairs from sentences considering a *window* of size w, our model will also employ the idea of *window* and consider each user to be paired with w users before and after that user in the like streams. More detail will be provided in Section III-A

III. Prediction Methodology

People interactions on the published posts produce temporal lists of users, representing the order of their interactions in different timestamps starting just after the published time. The main goal of this study is to design a model that is able to predict the potential users who will interact with a new published post[1] having prior-knowledge of the post's publisher and its early interactors. Taking advantage of PMI and inspired

[1] we call them likers through this study.

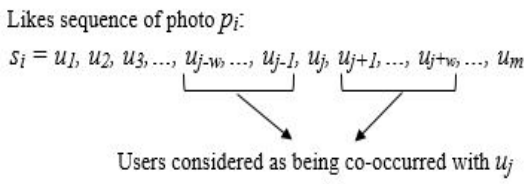

Fig. 1: Co-occurrences are computed for each user in the like sequences with her surrounded users, placed in a window of size w from two directions.

by word2vec model [7], we implement a novel model to extract users latent similarities and their associations from their interaction logs.

A. Users Similarity

Given a social network with a set of N users (U), engagement of users on a given post p_i will be shown as $s_i = \{u_j \mid u_j \in U, j = 1, 2, ..., m\}$ and called interaction sequences (s_i), where m is the number of interactors on the post p_i, and index j refers to the index of users in a temporal order. Therefore, as shown in Figure 1 each s_i is a subset of users who reacted to the post p_i with the publisher of the post in the first place of the sequence (u_1).
In a dataset of P published posts, interaction sequences over those posts will be presented as the collection S where $S = \{s_i \mid i = 1, 2, ..., P\}$. As mentioned, each s_i includes the interactors of the corresponding post, p_i.

PMI values are computed for each pair of users as follows:

$$PMI(u_i, u_j) = \log \frac{P(u_i, u_j)}{P(u_i).P(u_j)} \tag{1}$$

Where $P(u_i, u_j)$ is the probability that two users u_i and u_j have co-occurred in interaction sequences. $P(u_i)$ and $P(u_j)$ are the probabilities that u_i and u_j appeared in an s, respectively. To compute $PMI(u_i, u_j)$, $P(u_i, u_j)$ is the first requirement which needs the number of users co-occurrences. In order to show the impact of *window* concept inspired by Word2vec on computing PMI values, we measure $P(u_i, u_j)$ in two approaches:

i. *Publisher-liker adjacency*: Co-occurrence is defined as the number of times that u_i is the publisher of a post and u_j is the user who reacts to that post.
ii. *Window adjacency*: This approach uses a window inspired by word2vec, to compute the PMI values between user pairs. In this approach, we consider the co-occurrence of users in a window of size w, shown in Figure 1. This means that w users before and w users after u_i in the like sequence are considered as the users who are co-occurred with u_i.

Computed PMI values are assigned to the weight of the edges in the aforementioned user graphs. As already stated, the interaction graph of users is defined by considering users as its nodes. The edge between each pair of nodes is defined when one of the users reacts to the post published by the other. $PMI(u_i, u_j)$ is assigned to the weight of the edge between u_i and u_j in the interaction graph. The results are presented from the output of the following three approaches which are considered to build the activity graph:

- *Directed Publisher-Liker (DPL)*: The edges of this graph are directed and $PMI(u_i, u_j)$ is assigned to the edge which goes from node u_i to node u_j, if u_j reacts to u_i's post. PMI values are computed by aforementioned publisher-liker adjacency.
- *Undirected Publisher-Liker (UPL)*: The edges are undirected and the weight of the edge between u_i and u_j is the sum of the weights of the two directed edges between these two nodes in the previous approach (DPL).
- *Undirected Window (UW)*: This approach exploits *window adjacency* to compute the weights of the edges. Since the window adjacency considers different subsets of users from interaction sequences in which their relationship is not necessarily publisher-liker, the graph cannot be a directed one.

Two DPL and UPL approaches which use conventional definition of PMI are considered as the baseline methods to compare with UW approach where it uses new definition of PMI between two users under *window adjacency*.

B. Prediction Model

Here we describe in detail how our proposed method will identify the likers of a published post based on the users' latent similarities. Although PMI is widely used in prediction tasks on RS and NLP models, to the best of our knowledge, there is no previous study addressing the prediction of future engaging users using PMI and without handcrafted features.

Due to the successful studies on predicting the popularity of posts by exploiting the information of early interactors [5][15], we will also take into account the information of k early interactors of each post as a prior-knowledge and predict upcoming likers based on those earlier ones.

Given k early likers, we spot these k nodes on interaction graph, find the neighbors of each node, and make a collection of k nodes' neighbors. For each node in the collection, we compute the average weight of the edges between that node and k early likers. To identify future likers, we first sort the nodes available in the collection based on their already computed average weights and choose l top nodes with the highest weights referred by l *Nearest Neighbors* (l-NN). As the weights of edges are PMI values, the strongest edges imply the highest values on PMI. We select l-NN in two manners, choosing them from *all neighbors* of k early likers like what described above as shown in Equation 2, and choosing them from *only the friends* of early likers according to Equation 3.

$$l\text{-}NN(k) = l\text{-}MAX(\frac{1}{k}\sum_{i=1}^{k} PMI(u_i, :)) \tag{2}$$

$$l\text{-}NN(k) = l\text{-}MAX(\frac{1}{k}\sum_{i=1}^{k} (PMI(u_i, :) * Friends(u_i, :))) \tag{3}$$

Where l is the number of chosen neighbors, k is the number of early likers as input, PMI is the matrix of PMI values, $PMI(u_i, :)$ is a row of PMI matrix indicating the PMI values between u_i and other users, and $Friends$ is the

TABLE I: The Flickr Dataset Characteristic

Attribute	Value
#Photos	11.2M
#Users	2.3M
#Photos with $\geq$ 30 favorites	128K
Avg(#favorites) of $\geq$ 30 favorites	61
Median(#favorite) of $\geq$ 30 favorites	45

binary friendship matrix[2] and $*$ operation is the element-wise multiplication of two PMI and Friends matrices. Summation sign in both formulas applies an element-wise summation over the PMI matrix's rows belong to the k early likers. The average of this summation, which it is also element-wise average, is the input of l-MAX function as a vector. This function selects l indices from the input vector having the highest average PMI values as l future likers. In Equation 2, future likers are chosen from all neighbors of early likers but in Equation 3, they are chosen only from the friends of early likers where it is achieved by multiplying PMI matrix by the binary friendship matrix. Two l-NN equations will be used to choose future likers based on early likers where the connection between users are defined according to the three approaches, DPL, UPL, and UW.

Next, we evaluate our proposed model by using a large Flickr dataset and present the outcomes of the prediction based on the different presented approaches.

IV. Evaluation and Results

This section evaluates the proposed prediction method and presents the dataset information used in the evaluation as well as the results obtained from the experiments.

A. Dataset Description

To evaluate the proposed model of likers prediction, we used a Flickr dataset [31] including more than 11M photos and the activity history of 2.3M users for 100 days. User reactions to the photos in this dataset are indicated by marking them as favorites. In this study, we will refer this action by *like*, and the interacted users by *likers*. Table I shows the characteristics of the dataset and the values of its different attributes.

Since our method is based on the photos' like sequences, we consider those photos that have at least 30 likes to have enough length to apply the aforementioned idea of the *window*. Applying this filtering leaves the dataset to include 128k photos where each photo has the minimum number of 30 likes. In addition, to produce the reliable users' co-occurrence probabilities, a minimum frequency of likers is required. To fulfill this requirement, we pick only the users who have appeared at least 50 times in the dataset called *active users*. The thresholds for the number of likes and number of user's repetition are adapted from previous studies [32] [33]. The dataset is divided into two parts, namely train and test datasets, with 70% and 30% volume of the dataset, respectively. The train dataset is used to compute the PMI matrix and test dataset is exploited to predict the future likers.

[2] $Friends(u_i, u_j)$ is 1 if u_i follows u_j otherwise it is 0.

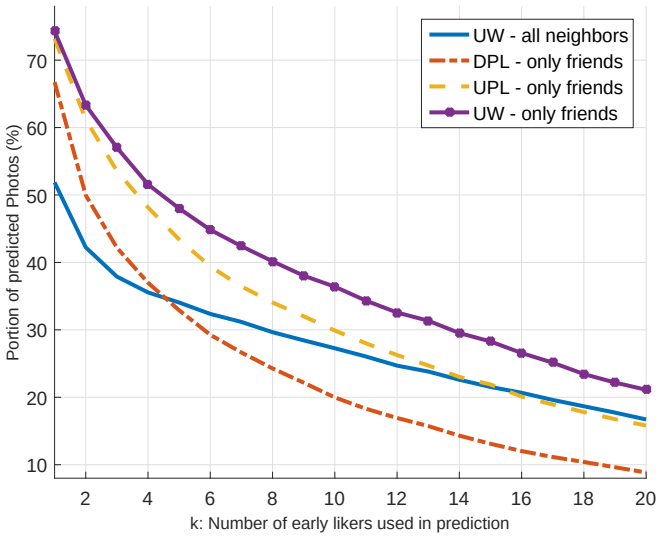

Fig. 2: The portion of photos with at least one correct prediction in their future likers for different k numbers of early likers. Choosing likers from friendships improves the prediction results.

B. Future Likers Prediction

The principal goal of the present study is to predict the group of users that have more probability to react in near future to a given post, where the prediction model will use only users' engagement log. To this end, we first have to choose the window size in the prediction model. We examined different values of w, but due to the space limit, we present the result for our model with $w = 10$. PMI between users is computed from like sequences available in the train dataset, through the neighborhood of size w using the Equation 1. To predict future likers, the number of early likers (k) is set to vary from 1 to 20 in the two previously described Equations 2 and 3. By assuming to be aware of k early likers, we find l top users who are most expected to like a given post as the future likers of that post. Selected l users have the maximum amounts of average PMI values with early likers, representing the closest and similar users to the early likers. In order to set the value of l, we need to know the potential number of likers that will be predicted for each photo. This number comes from the number of active users in each like sequence. Because non-active users are already eliminated from the like sequences due to their repetition less than 50 times in the dataset. Considering that this number of active users is different for each like sequence, we fixed the maximum number of likers to predict (l) to 20, which is the average number of active users in the like sequences of the train dataset. The prediction phase is conducted over the test dataset.

We applied l-NN function using three approaches mentioned in section III-A and chose future likers. Prediction result is represented in two aspects, photos precision and likers precision.

1) Photo Precision: Photos precision indicates the portion of photos which at least one of their future likers out of 20 ($l = 20$) has been predicted correctly (called *predicted-photos*). Figure 2 shows the photo precision of different

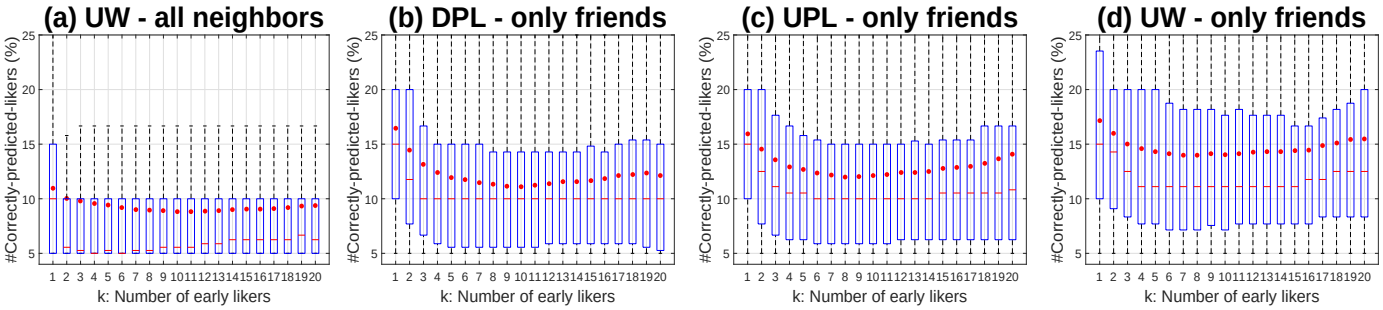

Fig. 3: Distribution of #correctly-predicted-likers of photos with different values of early likers (k). (Red dots show the mean values).

approaches along the Y-axis for the different number of early likers (k) along the X-axis as prior-knowledge. As it shows, we performed prediction of likers using *DPL*, *UPL*, and *UW* in two categories, first choosing likers from all neighbors, and second from only friends. Since choosing likers from all neighbors shows significantly lower accuracy than choosing them from only friends, we circumvent to present the results of it in all three approaches. However, we present the result of *UW* from this category as the best representation of this group only to display its low accuracy. It is somehow expected that looking for future likers among friends provides more accurate predictions. First, because most of the users on any social network mark a post as liked due to their friendship with the publisher of that post, without considering the content of the post. Second, choosing future interactors from the entire users without restricting the search space, especially in such big datasets will not be intelligent and applicable. We also examined the random selection of likers as a baseline method to compare with our proposed approaches. But due to its very inaccurate result, we avoid presenting it.

As Figure 2 depicts, among four examined approaches, *UW* and UPL when they choose likers from only friends, can predict likers for the higher number of photos. In addition, it shows that by increasing the prior-knowledge about early likers (k along the X-axis) and subsequently choosing future likers based on them, the portion of predicted photos has been substantially declined. The highest number of correctly predicted photos is when the value of k equals 1, which is the case when we choose the nearest neighbors to the first user in the like sequence of a post who is the publisher of that post. It can be interpreted that the future likers are practically dependent upon the publisher. In other words, being aware of more early likers than publisher not only will not enhance the prediction precision but also will inject noisy data which leads to an inaccurate selection of likers.

2) Likers Precision: Likers precision is defined for each photo separately and indicates the portion of l predicted likers that are predicted correctly. Figure 2 shows only the quantity of photos which at least one of their future likers is predicted correctly using different approaches, without representing the quality of prediction. To identify the quality of prediction which indicates likers precision, we inspect precisely the number of likers which are predicted correctly for each photo (#correctly-predicted-likers). Figure 3 presents the distribution of these numbers along the Y-axis (in percentage) for different k values. According to this plot, *UW* by choosing likers from all neighbors has the lowest mean value (presented by red points) of #correctly-predicted-likers where it is almost around 10%. On the other hand, *UW* with choosing likers from friends shows the best results such that first, the mean of #correctly-predicted-likers remains almost around 15% for different k numbers (against to *UPL* and *DPL* which drops) with the highest value at $k = 1$ and second, the distributions in each value of k show higher numbers of predicted likers in *UW - only friends* than other approaches. *UPL* and *DPL* have almost similar distributions of predicted likers as well as similar mean values in different numbers of k.

As we observed in both Figures 2 and 3, the number of predicted photos (photo precision) and number of correctly predicted likers (likers precision) have their highest values in $k = 1$. It practically signifies that unlike the popularity size prediction problem [6], the prediction of future likers depends more on the publisher of a post than other early likers. Due to this point, in the following section, we will focus on the results of $k = 1$ where the photo precisions in Figure 2 are 51%, 66%, 73%, and 74% for *UW - all neighbors*, *DPL*, *UPL*, and *UW - only friends*, respectively. The purpose of this focus is to choose the best prediction approach among the presented ones in the elaborated presentation of #correctly-predicted-likers distributed in Figure 3 when $k = 1$.

C. Publishers as Predictors

As mentioned earlier, the concentration of this section is on presenting the results of the prediction on $k = 1$ which leaves the prediction problem to find future likers based on only publisher. To elaborate the results obtained from different approaches, we compute the precision of prediction for each photo (p) called $Precision_p$ as follows:

$$Precision_p = \frac{\#correctly - predicted - likers_p}{\#likers - to - predict_p} \quad (4)$$

Where #correctly-predicted-likers$_p$ is the number of likers of the photo p who are predicted correctly, and #likers-to-predict$_p$ is the number of photo p's likers.To provide simpler representation, we group $Precision_p$ values into ranges. Figure 4 displays the distribution of photos' precisions ($Precision_p$) computed from the results of four prediction

approaches. In this figure, the first bar in the range of 5-10%, which belongs to *UW - all neighbors* approach, shows that this approach can predict only 5 to 10 percent of likers correctly for 23% out of 51% *predicted-photos*, 10 to 15 percent correct prediction for 18% and so on.

Comparing different approaches reveals that *UW - all neighbors* has the majority of its correctly predicted photos in the range of 5-10%. It means that only 5 to 10 percent of likers are predictable for almost half of the predicted-photos (23% out of 51%) using *UW - all neighbors* approach. Therefore, this approach not only has the lowest percentage of predicted-photos but also is not able to predict more than a few percentages of likers. Contrary to *UW - all neighbors*, the other three approaches perform better and the likers of the majority of photos are predicted by 10-15% and 15-20% precision using those three approaches. It implies that when likers selection is restricted to choose them only from friends instead of from all neighbors, the precision of the results is substantially enhanced. The reasons behind this phenomenon are previously discussed in Section IV-B as well.

From the three better performing approaches, *UW - only friends* outperforms *DPL - only friends* and *UPL - only friends* by resulting a higher number of photos with high $precision_p$ in likers prediction. As Figure 4 shows in the precision ranges higher than 20%, the number of predicted-photos by *UW - only friends* beats the others. Accordingly, it substantiates the success of this method in predicting high number of likers. In summary, we found that prediction of future likers considering their relation with publisher provides a better result than with other k early likers. Restricting the prediction to choose likers only from friends instead of selecting them from all neighbors elevates the quantity of number of predicted-photos by more than 20% (from 51% in *UW - all neighbors* to 74% in *UW - only friends*), and the precision of likers prediction from low to high ranges.

Finally, *UW - only friends* succeeds to predict the higher amount of photos with higher precision of likers in compare to the other three approaches. As stated previously, this method exploits the co-occurrence of users in a window of size w, which makes it able to derive the latent similarity between users even if they have not interacted directly on the posts of each other. Consequently, considering a window to compute the co-occurrences of users helps *UW - only friends* to improve the precision of likers in the prediction task.

V. Publishers Analysis

As we observed in section IV-B likers precision is different for each photos. In order to identify why some photos have more correctly predicted likers than others, we study the properties of their publishers. Looking at the prediction result shows that the *UW - only friends* approach produces the best outcomes (although *UPL - only friends* was very close). Thus we study the result of this approach to discover the common features of those publishers that likers of their photos are predicted more accurately. To this purpose, we investigate publishers properties studying their relationships, activities, and engagements. Four metrics are considered for each publisher: #followers, #followings, #activities (published photos) and #engagements (number of times that a publisher reacted to the photos of other publishers).

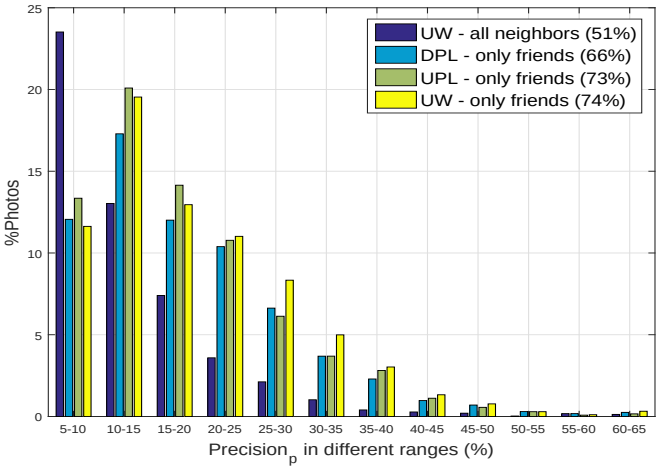

Fig. 4: Distribution of predicted photos (Y-axis) over the different ranges of $precision_p$ which is computed per photo separately in $k = 1$. (the portion of each bar is from the percentages shown in the legend)

Earlier we defined two *#likers-to-predict*$_p$ and *#correctly-predicted-likers*$_p$ metrics for individual photos which are the number of users in p's like sequence and the true predicted likers of p, respectively. Now we will define the same metrics for publishers. Since each publisher has different number of photos in the dataset, we compute the average of these values for the photos of each publisher and associate them to the corresponding publisher as the average values of potential #likers-to-predict and #correctly-predicted-likers of that publisher. On the other side, to show how many photos of each publisher have at least one correct prediction of their likers, $Predict - frac_{pl}$ represents the percentage of the following fraction:

$$Predict - frac_{pl} = \frac{\#predicted - photos_{pl}}{\#published - photos_{pl}} \quad (5)$$

Where #published-photos$_{pl}$ is the number of photos published by the publisher (pl) and #predicted-photos$_{pl}$ is the number of her photos with at least one correct predicted liker.

Figure 5 compares publishers in terms of their $predict - frac_{pl}$ shown by color, the number of predicted photos shown by the size of the circles, the average number of predicted likers in the Y-axis, and the average number of likers to predict along the X-axis. From the perspective of the quantity of predicted photos, the most successful predictions belong to the publishers with the higher values of $predict - frac_{pl}$ represented by blue (and darker) colors and the higher #predicted-photo presented by larger circles in Figure 5. In addition, from the perspective of prediction quality, the most successful predictions are associated with the publishers whose average number of correctly-predicted-likers are high. We call them the high-predictable publishers, located on top of the plot along the Y-axis. We used a heuristic to find a reasonable number of high-predictable publishers. We intuitively filtered publishers

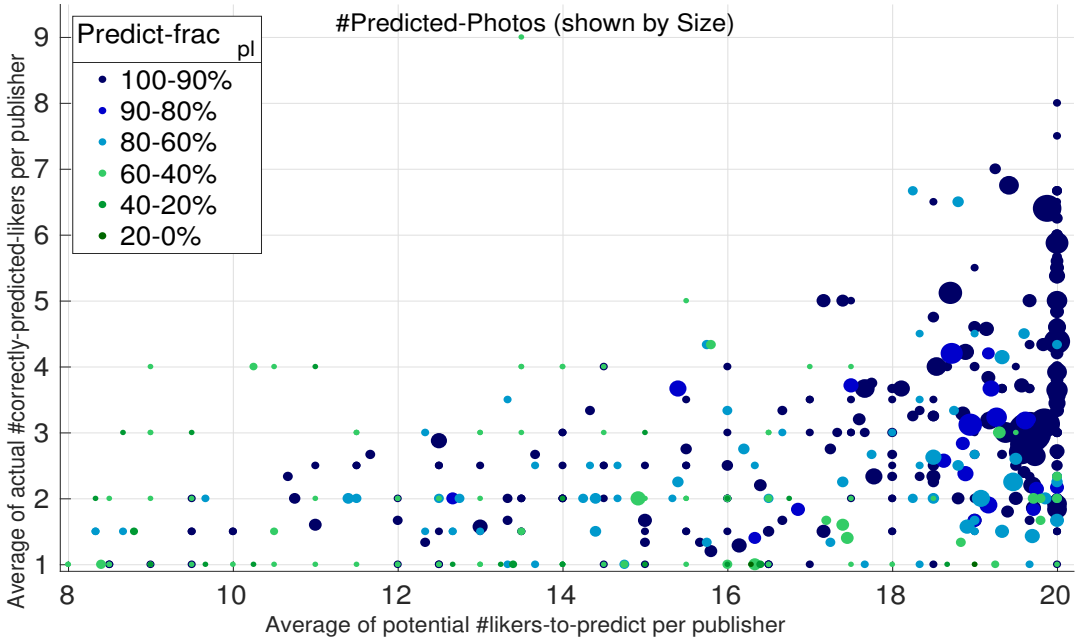

Fig. 5: Four-dimensional representation of publishers consisting avg. potential #Likers-to-Predict (X-axis), Avg. #correclty-predicted-likers (Y-axis), actual #predicted-photos (shown by the size of circles), and predict-frac (shown by color). The bigger size of circles shows the higher number of predicated photos and vice versa.

by selecting those which own #predicted-photos of more than 20, or have $predict - frac_{pl}$ value greater than 80% or those whose the average number of correctly-predicted-likers is more than 5. Among the selected publishers, the ones who met three applied filtering conditions are grouped as the highly predictable publishers with 22% of selected population, and the others who met only two or one of the filtering conditions are grouped as lowly predictable publishers with 78% of the whole filtered publishers.

To determine the characteristics of these two groups and to identify the influential factors in the success of highly predictable publishers, we compare the four previously mentioned metrics of the publishers in those two groups in Figure 6. The value of each metric is the average value in this diagram. It shows that high-predictable publishers have significantly higher values for their number of followers and engagements than the low-predictable ones. These values are almost twice bigger for high-predictable publishers. #followings of high-predictable publishers is almost 50% greater than the value of the same metric for the low-predictable publishers. However, the average amounts of the published photos (activities) by those two groups are almost equal, which indicates that a user's activity-amount regarding publishing posts is not a referable metric to determine the predictability of her photos' likers.

These observations reveal the substantial effect of a publishers' high number of followers and high number of engagements on achieving a successful prediction of her content's likers, employing only user's interaction history. This means that the future reactions to a post published by a publisher with high #engagements and high #followers are more predictable. In addition, since we predict likers by exploiting the PMI

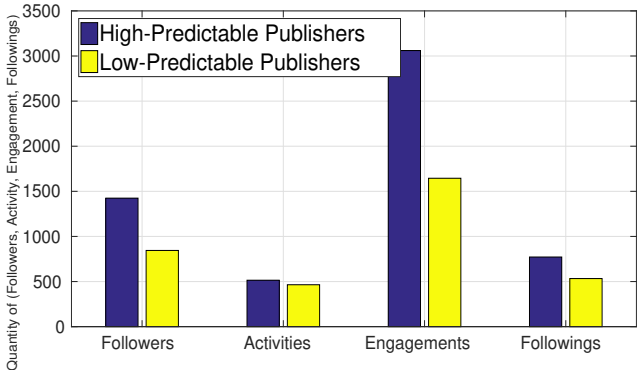

Fig. 6: Comparison of high-predictable and low-predictable publishers in terms of their average values of #followers, #engagements, #activities (published-photos), and #followings.

closeness in *UW-only friends*, we can infer that using the latent similarity of users derived from their engagement history can produce more reliable results for those users with a high number of followers and engagements to use in this prediction task.

High correctness of prediction for the posts of publishers with high #engagements implies that engaging a user in the posts published by other users helps to reveal their common preferences as well as helps to make other users more predictable in reacting to her future posts. On the other side, publishers with more followers increase the probability of accurate prediction results because their contents will probably get a high number of likes. The results of this section illustrate that exploiting activity sequences can effectively extract trustworthy latent similarities between active users to employ

in predicting future likers especially for the publishers with high number of followers and engagements.

VI. Conclusion

This study sheds light on the interesting topic of predicting the users who are most likely to react to the posts published in social media. A novel model based on PMI and inspired by word2vec was implemented to extract users' latent similarity. The similarity of users is exploited to predict the future likers of a post based on the information of the post's publisher as well as its early likers. Our findings disclose that considering users adjacency under a window of neighborhood reveals users hidden similarities and leads to more precise PMI values. As well as, we found that predicting future likers of a post is considerably correlated to the publisher of that post than other early likers. We studied in details the output of prediction model from photos precision and likers precision perspectives. Evaluation of experiments over a large Flickr dataset confirmed the ability of the proposed method to identify future likers of Flickr's posts, especially those published by super interactive publishers. Although the study has reached the worthy results, it is limited by the lack of homogeneous data from other social networks to generalize the results for larger number of social network platforms. The proposed prediction approach can help advertising campaigns, recommender systems, and content placement controllers by providing prior-knowledge of future engaging users. Further research could include improving the outcomes of the proposed method by augmenting users' information to the content of posts as well as fine-tuning this technique to extract users' latent relations and preferences. Some of these goals could be realized by applying this method to distinct datasets.

References

[1] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Record*, vol. 42, no. 2, pp. 17–28, 2013.

[2] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing social networks*. SAGE Publications Limited, 2013.

[3] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *arXiv preprint arXiv:1405.1790*, 2014.

[4] E. Dubois and D. Gaffney, "The multiple facets of influence: identifying political influentials and opinion leaders on twitter," *American Behavioral Scientist*, vol. 58, no. 10, pp. 1260–1277, 2014.

[5] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 925–936.

[6] B. Shulman, A. Sharma, and D. Cosley, "Predictability of popularity: Gaps between prediction and understanding," 2016.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[8] S. Mohammadi, R. Farahbakhsh, and N. Crespi, "Popularity evolution of professional users on facebook," 2017.

[9] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *Journal of Internet Services and Applications*, vol. 5, 2014.

[10] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose, "User engagement in online news: Under the scope of sentiment, interest, affect, and gaze," *Journal of the Association for Information Science and Technology*, vol. 65, no. 10, pp. 1988–2005, 2014.

[11] A. Susarla, J.-H. Oh, and Y. Tan, "Social networks and the diffusion of user-generated content: Evidence from youtube," *Information Systems Research*, vol. 23, no. 1, pp. 23–41, 2012.

[12] S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: Photos with faces attract more likes and comments on instagram," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 965–974.

[13] P. Gong and H. Wu, "A cache partition policy of ccn based on content popularity," *International Journal of Advanced Science and Technology*, vol. 92, pp. 9–16, 2016.

[14] Z. Wang, W. Zhu, M. Chen, L. Sun, and S. Yang, "Cpcdn: Content delivery powered by context and user intelligence," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 92–103, 2015.

[15] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng, "Popularity prediction in microblogging network: a case study on sina weibo," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 177–178.

[16] S. Elsharkawy, G. Hassan, T. Nabhan, and M. Roushdy, "Towards feature selection for cascade growth prediction on twitter," in *Proceedings of the 10th International Conference on Informatics and Systems*. ACM, 2016, pp. 166–172.

[17] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter." in *ICWSM*, 2011.

[18] P. Turney, "Mining the web for synonyms: Pmi-ir versus lsa on toefl," *Machine Learning: ECML 2001*, pp. 491–502, 2001.

[19] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.

[20] A. Islam and D. Inkpen, "Second order co-occurrence pmi for determining the semantic similarity of words," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2006, pp. 1033–1038.

[21] G. Recchia and M. N. Jones, "More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis," *Behavior research methods*, vol. 41, no. 3, pp. 647–656, 2009.

[22] T. Martin, "community2vec: Vector representations of online communities encode semantic relationships," in *Proceedings of the Second Workshop on NLP and Computational Social Science*, 2017, pp. 27–31.

[23] D. Liang, J. Altosaar, L. Charlin, and D. M. Blei, "Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence," in *Proceedings of the 10th ACM conference on recommender systems*. ACM, 2016, pp. 59–66.

[24] M. Kaminskas and D. Bridge, "Measuring surprise in recommender systems," in *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the 8th ACM Conference on Recommender Systems)*, 2014.

[25] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012, pp. 298–307.

[26] E. Spertus, M. Sahami, and O. Buyukkokten, "Evaluating similarity measures: a large-scale study in the orkut social network," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 678–684.

[27] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 151–160.

[28] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Facta: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559–2560, 2008.

[29] C. G. Akcora, B. Carminati, and E. Ferrari, "Network and profile based measures for user similarities on social networks," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. IEEE, 2011, pp. 292–298.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013.

[31] M. Cha, A. Mislove, and K. P. Gummadi, "A Measurement-driven Analysis of Information Propagation in the Flickr Social Network," in *In Proceedings of the 18th International World Wide Web Conference (WWW'09)*, Madrid, Spain, April 2009.

[32] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[33] O. Barkan and N. Koenigstein, "Item2vec: Neural item embedding for collaborative filtering," *arXiv preprint arXiv:1603.04259*, 2016.