Identifying Influential Users on Instagram Through Visual Content Analysis

WAFAA HASAN ALWAN, EHSAN FAZL-ERSI, AND ABEDIN VAHEDIAN
Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran
Corresponding author: Ehsan Fazl-Ersi (fazlersi@um.ac.ir)

ABSTRACT In recent years, social networks have attracted many users’ interests. People express their daily experiences and emotions through social networks and become aware of others’ interests and thoughts. In these networks, influential users play an important role in broadcasting information to their communities. As a result, discovering such users in social networks has become very important, in particular, for marketing purposes. Various solutions already exist for identifying influential users, most of which are structure-based, in that they investigate the influence of a user according to his location in the network graph. This article proposes a novel approach for identifying influential users on Instagram, by examining User Generated Contents (UGC). More specifically, the proposed method combines various types of features extracted from the images posted on Instagram to determine whether or not a user is influential, without requiring any information about the structure of the network. An extensive set of experiments are performed to validate the effectiveness of the proposed method in identifying influential users.

INDEX TERMS Online social networks, influencers, content analysis, feature extraction, SVM, combined classifiers.

I. INTRODUCTION
The study of social-based recommender systems has attracted significant attention since the development of Web 2.0. While primary systems often ignored social interactions among users, the more recent approaches try to incorporate social network information in order to improve the quality of the recommendations and make them more personalized. According to 42 Essential social media statistics for 2020:

There are 3.484 billion active social media users. People spend 2 hours and 23 minutes daily on social media browsing and messaging. Additionally, 98.55% of people use at least four social media channels daily. Top social media networks are Facebook (95%), Twitter (84%), Instagram (74%), LinkedIn (62%), and YouTube (61%). When these people encounter multiple potentially confusing recommendations, they tend to turn to influential people, whether they are a part of the supply chain (e.g., retailers or manufacturers) or are value-added influencers, such as industry analysts or professional advisers. In general, influencers in a social network are individuals whose impacts are broadened through the network. This impact is usually not comprehensive and influencers often focus on limited topics to advertise.

Being able to discover or identify the influencers on social networks could be of great value since influencers can play a significant role in the success of various social, political and viral marketing campaigns as well as the entertainment events. The key to the influential users’ identification problem is finding ways to measure the spreading capability of influencers in social media by discovering the main characteristics of their contents.

In this article, we attempt to identify influential users on Instagram, a photo and video-sharing social networking service with more than a billion registered users as of May 2019. Instagram has over one billion monthly active users and 500 million daily Story users. It is the second most downloaded free app in the Apple App Store. It is the second most popular platform among teens after Snapchat, because more than half of Instagram users are in the 18-34 year old age bracket. The top five countries in terms of Instagram usage are USA, India, Brazil, Indonesia, and Russia. 77% of people use Instagram for viewing photos, 51% for watching videos, 45% for sharing content, 23% for networking, and 11% for finding/shopping for products, and video content is well-loved on Instagram. In 2019, there was an 80% increase in time spent on watching Instagram videos. How-to tutorials are the most watched videos on Instagram (80%). Based on the large number of users on Instagram and also the various relations that exist among them, providing an effective
method to identify influential users has been gradually considered as an essential factor.

The most significant contribution of the proposed method is that unlike the vast majority of the existing solutions, it only relies on the analysis of the User Generated Content (UGC) and not the extent of user interactions (e.g. number of posts, number of followers, number of ‘likes’ and so on), or the structure of their connectivity. More specifically, the focus is on extracting various low- and high-level visual features from images and then combining the informative ones towards creating a classifier that can distinguish between images posted by influencers and those posted by ordinary users.

This article is organized as follows: in Section 2, the relevant literature concerning the topic of influence evaluation in social networks is reviewed. The proposed method and its various components are described in Section 3. In Section 4, the conducted experiments are presented and their results are discussed. Finally, Section 5 concludes the paper with a brief overview of the proposed method and highlights a few directions for future work.

II. RELATED WORKS
There are several reasons why it is important to identify influencers in social networks, and the main reason for this identification is marketing. Due to the large number of social network users, many companies have changed their styles of advertising in these social networks [1]. Also, forwarding the contents of advertisements by users to their friends and family can be more effective than ads provided by companies. The confidence that users feel towards their friends in comparison to these companies is one of the reasons for this claim. Accordingly, most commercial companies have changed their tendencies towards a method known as viral marketing [2]. In this method, marketing begins with a small number of users as the primary group, and the identification of influential users as the primary source can be more efficient in the spreading process [2], [3].

In general, the available methods for identifying influential people can be divided into three main categories: 1) Structure-based methods, which compute the influence of a user according to his location in a given network’s graph; 2) User-interaction based methods, which also incorporate the interactions among users; and 3) Content-based methods where the influence of a user is computed according to the content of the user’s published posts in a given social network. Below, each category is elaborated in more details.

A. STRUCTURE-BASED METHODS
In these methods, the influence of a node is determined by the topology of the social network, often represented as a graph. The graph properties, which are used to describe the influence of a node, are mostly referred to as centrality measures – the term Centrality is one of the most important terms used in social network analysis for identifying influential users and are grouped into three types: local, global, and semi-local. In local approaches (e.g., [4]), the influence of a node is determined by the node itself and its neighbors. One of the most famous local measures is degree centrality (DC) in which the node with a greater number of neighbors receives a higher score. Local methods have low computational complexity, but due to the lack of attention to the node’s structural location in the network, they also have a low accuracy. In global methods (e.g., [5], [6]), all network nodes participate in computing the influence of a node. Betweenness, closeness and k-shell centrality are some of the most famous global methods. The weakness of these methods, however, is that the larger the graph is, the higher the computational complexity becomes, making them unusable in large networks. The semi-local methods (e.g., [7]–[11]) attempt to provide a trade-off between the accuracy and complexity of local and global methods. In these methods and in order to compute the influence of a node, both its direct and indirect neighbors for different levels are taken into account. Neighborhood centrality detects the influence of complex network nodes by combining degree centrality and k-shell, to obtain local features and global information of the node. Local centrality (LC) considers both the nearest and the second nearest neighbors of a node. Hybrid degree centrality (HC) uses either DC or LC depending on the given spreading probability. Local structural centrality (LSC), considers the number of the nearest and second nearest neighbors in addition to the topological connections among neighbors. All these methods are examples of semi-local centrality measures that compute the final influence of a given node by considering the influence of its neighbors.

B. USER-INTERACTION-BASED METHODS
Interaction is at the core of all social networks. Many influence studies are based on users’ interactions with others through mechanisms such as like, share, mention and comment. In a seminal study, [12] compared three different measures on Twitter, namely the number of followers, number of retweets and number of mentions. They observed that the most followed users do not necessarily score high on numbers of retweets and mentions. The number of followers of a user is indicative of her popularity and not necessarily related to her influence. In some works, the influence on Twitter is measured by techniques such as PageRank [5] and short URL click [13].

Another group of researchers focus on combining interactive and structure-based techniques. Examples of such attempts are the works of [11]–[15]. In [16] researchers ranked the influence of users by the number of their followers, PageRank and by retweets. They found the first two rankings to be similar, but rankings based on retweets differ from the previous two, indicating a gap in influence induced from the number of followers and that from the popularity of one’s tweets. In [17] the researchers suggested a user influence ranking (UIRank) algorithm to identify the influential users through the combination of two measures: influence of tweets, which is measured by counting the number of retweets and comments on the user’s tweets, and applying graph theory...
C. CONTENT-BASED METHODS

Simple interaction-based approaches often fail to reliably and accurately predict the extent of a user’s influence, particularly with the rise of social bots and fake accounts that automatically gather followers and generate messages. The more complicated approaches, while potentially able to provide better accuracy, require extensive computational resources to compute more complicated features. The content-based methods, on the other hand, attempt to use yet another source of information, ignored by the previous two groups of solutions, which is the content of the users’ posts (or User Generated Content). These methods propose various ways of integrating this information into previous methods in order to provide a better evaluation of users’ influence. For instance, In [20], the authors take a supervised learning approach with human annotators to extract the characteristic attributes of influential tweets and users. They process the information on Twitter, a social network, and derive from it a set of attributes describing users. Each attribute belongs to (I) an activity related to their followers, (II) qualities computed by analyzing their content i.e. the quality of the tweets, or (III) centrality which incorporates the position of users in the graph. Also in [21] the authors use Flesh- Kincaid Grade Level, which measures the quality of tweets, together with a combination of the user’s position in the social network and the polarity of their opinions. Another group suggested that the influence of users can be interpreted as the “authority of a web page”, therefore they proposed the TwitterRank algorithm, which is topic sensitive and uses Latent Dirichlet Allocation (LDA) to identify the topics that twitter users are interested in based on the tweets they publish [22]. [23] uses a novel Follow ship LDA (FLDA) model that properly captures the content-related and content-independent reasons why a user follows another in a microblog network. [24] uses hashtags to identify influential users in news-related communities. In another work, [25] proposed a method for identifying influential posts on Instagram by means of text analysis. While Instagram is a social media mainly based on photos, they showed that the photos’ accompanying textual data can play a significant role in separating contents made

and posted by influencers from those created by ordinary (non-influential) users.

III. THE PROPOSED METHOD

The two concepts of identifying influencers in social networks and detecting influential posts are relatively interconnected because influencers intend to have more influential posts. So, identifying influencers can be investigated by detecting influential posts. As shown in Figure 1, influential posts contain features and characteristics that make them effectively impact the followers. One of these features is the presence of an object that reflects the image subject’s goal and title, occupying the center of the image, so that it attracts the attention of all viewers. An example of this can be seen in the image of coast, or fruit. Another important feature is the color differentiation of the target objects and their distinctiveness from the background of the image, such as the ring image. Another feature is focusing the camera on the target object so that it occupies most of the image, as can be seen in the shoes’ ad image. There are many features such as high resolution objects and blurriness in the background of images, which distinguish influential and non-influential posts. We propose a comprehensive method to identify influential posts using supervised machine learning, by extracting local and global features from posts’ images. Figure 2 demonstrates the general ideas of this article.

The proposed method can be divided into three parts: 1) Data Collection, 2) Feature Extraction and Influential Post Identification, which focus on various visual representation of the image and classification of the image into influential and non-influential sets, and finally, 3) Classifier Development and Influential User Identification which fuse different classifying methods in order to develop a final classifier to find both influential users and influential posts.

A. DATA COLLECTION

Given that there is no available dataset to categorize users into influential and non-influential sets, one of the first objectives of this research was to create a valid dataset for performing experimental results. To this aim, we collected

FIGURE 1. Some examples of Instagram posts.
the information of Instagram users who are listed on www.influence.iconosquare.com, and have been verified as influential users in various fields. Then, given that influential users usually do not follow each other on social networks, we chose a set of non-influential individuals among the users who follow one of the influential users and also have less than 200 links with others. The data for each user was collected using Instagram’s API and included bio, number of followers, number of followings and posts; the posts themselves contained media ID, textual content and image URLs. The collected dataset covers contents on various subjects such as Tourism, Fashion, Food, etc.

The dataset consists of two main parts. The first part includes posts on general topics, containing 67,249 posts from 1,608 users, with 31,462 posts by 259 influential users and 35,787 posts by 1,349 non-influential users, and is further divided into two sets to allow us to measure the effects of variety in the dataset on the performance of our proposed method. The first subset, which we refer to as DB_1 contains 1,049 users, 168 influential users and 881 non-influential ones, and has 18,853 posts in total, 8,862 posts belonging to influential users and 9,991 posts belonging to non-influential users. The second subset, referred to as DB_2 includes 559 users and 48,396 posts. The users’ data contains 91 influential users with 22,600 posts and 468 non-influential users with 25,796 posts.

The second part of the dataset, referred to as DB_3, is dedicated to posts on a specific topic, namely fashion. As mentioned before, influential users generally focus on a limited range of topics, or at least are more influential on a specific subject. To take this into account, we choose fashion because of its popularity and used a hashtag-based theme extraction method to find users that operate in this field. We use this data to measure the performance of our proposed method in detecting the influencers in the field of fashion. This part of the database includes 243 influential and 468 non-influential users and 52,609 posts, 26,813 posts belonging to influential users and 25,796 posts to non-influential users. Table 1 shows the details and division for each part of the dataset.

### B. FEATURE EXTRACTION

The image of an influential post contains objects that show the importance of the subject and also have an impact on other users. The various methods used for analyzing and extracting the major global, local and visual features for each user’s posts as explained in the following sub-sections.

1) GLOBAL FEATURES

We used the images of all users’ posts to extract global features, which address the holistic features of the subject in the post. For this purpose, we used GIST, which refers to meaningful information that observers can identify at a glance from the scene. The GIST descriptors concentrate on the shape of an image and the correlation between the outline of surfaces and their properties. The representation of an image in the GIST is defined as a spatial envelope from the pre-processed input image. The pre-processed image is the input image that is first converted to gray-scale and then split into a grid with equal scales. Finally, the output of each cellular grid is calculated using a series of Gabor filters [26]. The input image is convolved with thirty two Gabor filters at four scales and eight orientations resulting in thirty two feature maps of a size equivalent to that of the input image. Each feature map is divided into sixteen regions and the feature values are

| Data Set | number of Users | number of Posts |
|----------|----------------|----------------|
|          | train | test | total | train | test | total |
| DB_1     | 525   | 524  | 1049  | 9429  | 9424 | 18853 |
| DB_2     | 280   | 279  | 559   | 28697 | 19699 | 48396 |
| DB_3     | 364   | 365  | 729   | 32904 | 20200 | 53104 |
averaged within each region of interest. The averaged feature values from all the sixteen regions contained in thirty two feature maps are concatenated to produce a GIST descriptor containing 512 features (16 regions \* 32 feature maps). In this manner, the GIST descriptor provides the gradient information of an image [27], [28].

2) LOCAL FEATURES
Influential posts seem to contain objects and entities that attract the attention of other users, so we aim to find influential posts by employing local features to detect significant objects in an image. The best local feature extraction methods in image classification systems are bag-of-features (BOW) [29], [30] and spatial pyramid matching (SPM) [30]. The BOW method represents an image as a histogram of its local features. It fully ignores spatial translations of features and disregards the information about the spatial layout of features, making it incapable of capturing shapes or locating an object. One of the most powerful extensions of the BOW is the SPM method, which partitions the image into increasingly finer spatial subregions and computes histograms of local features from each sub-region. Typically, 3 subregions (=0, 1, 2) are used. The SPM approach, based on BOW, contains four steps. In the first step, feature points are extracted from the image using SIFT, this gives the “Descriptor” category. Then, a codebook using K-means with M entries is applied to quantize each descriptor to generate the “Code” category, where each descriptor is converted into an RM code. Next in the “SPM” category, multiple codes from inside each sub-region are pooled together by averaging, and are normalized, to form a histogram. Finally, the histograms from all sub-regions are concatenated by applying a coefficient to create the final representation of the post for classification. The coefficient is larger for higher levels and smaller for lower levels to increase the importance of histograms that have more spatial information. The equation (1) demonstrates the length of the feature vector based on the number of levels (L) and the number of channels (M). For example, for 2 levels and 200 channels, the resulting vector has a size of 4200.

\[
\sum_{i=0}^{L} 4^i = M \frac{1}{3} 4^{l+1} - 1
\]  

(1)

3) PRE-TRAINED CNNS FOR FEATURE EXTRACTION
The latest researches show that the Deep CNN models are most beneficial at expanding feature extraction and representing images. Deep CNN’s convolutional layer, also called feature extraction layer, provides features that are robust to various transformations, such as translation, rotation and scaling [31]. Transfer Learning refers to the ability to migrate knowledge among tasks by making use of their similarities; to apply the knowledge learned from an existing environment or data to a new environment or data [32], [33]. The transfer learning of CNNs has shown wonderful potential in the field of image recognition and classification. Generally, transfer learning can be applied in two different ways: employing a pre-trained network as a feature extractor, or fine-tuning a pre-trained model. The pre-trained network can be used as a feature extractor by removing the last classifier layer from the network. The result is a fixed size feature extractor which can be employed on various tasks. The fine-tuning strategy not only re-trains the network to adapt to the target dataset, but also includes adjusting the parameters of the network by proceeding the back-propagation process [34]–[36]. In general, the earlier layers of CNNs learn low-level features that can be applied to most visual tasks, and the latter layers learn high-level features that are more useful for specific tasks [37], [38]. Therefore, keeping the earlier layers fixed and fine-tuning the last few layers is a useful strategy for transfer learning. In this section, the performance of two pre-trained CNNs are investigated for feature extraction, namely ResNet152 [39] and AlexNet. As shown in [34] and [35], features extracted globally from AlexNet perform well in image classification tasks, while ResNet often provides superior performance in encoding local attributes in images. Therefore, in our experiments, AlexNet and ResNet152 are used to extract 4096-dimensional global feature representations, and 1000-dimensional local feature representations from images, respectively.

4) VISUAL FEATURES
In observing influential and non-influential posts it can be seen that the influential posts, in addition to possessing impact objects, have several important visual features such as blurring in the background, high resolution objects and distinct color distribution. We measured three of these visual features from the images, in order to obtain a better classifier for influential posts’ identification. The following subsections cover the thorough description of these features.

a: BLURRING MEASURE
Image blurriness is one of the distortions that is expressed as a loss of detail and a decline of the edge sharpness in the content space. There are many factors that may cause blurriness, for example, defocus, camera shake and motion [40]. There are two techniques for measuring blurriness, namely subjective and objective. The subjective method uses human observers while the objective method is based on several metrics of an image, such as full-reference, reduced-reference and no-reference [41]. Image Quality Measure is used to investigate the added value of the visual attention aspects. Therefore, we compute the IQM with the no-reference image, in the frequency domain, using discrete Fourier transform in all regions of the image. The visual feature vector of length 64 is used for predicting the amount of the blurriness [42].

b: COLOR HISTOGRAM
The color histogram represents the number of pixels that have colors in each of a fixed list of color ranges. It may also be represented and displayed as a smooth function defined over the color space that approximates the pixel counts. The color histogram can be used for various color spaces such as RGB or HSV. We used the color histogram in RGB space with
TABLE 2. Accuracy of different feature extraction methods on all subsets.

| Dataset | Number of Posts | Accuracy Methods |
|---------|----------------|-----------------|
|         | Inf | Non | Inf | Non | Holistic features | Local features | Visual features |
|         |     |     |     |     | Gist | Alex | Resnet | SPM | Blur | Color | Entropy |
| Train   | 4426 | 5003 | 4436 | 4988 | 57.87% | 64.93% | 65.85% | 65.72% | 56.88% | 60.48% | 51% |
| Test    | 7068 | 12631 | 7569 | 12631 | 56.11% | 64.65% | 67.16% | 64.47% | 55.28% | 59.72% | 53% |
| Length of feature Vector | 512 | 4096 | 1000 | 4200 | 64 | 512 | 1 |

512 dimensions in order to distinguish between influential and non-influential posts [43].

\[ c: \text{ENTROPY} \]

The entropy can be interpreted as a way of measuring the order or amount of information. In some studies entropy is a standard for measuring the visual dimensions of an image [44] or a method for collecting information for some system parameters. An additional interpretation of entropy is as a way to measure information within an image. Using the equation 2, one can examine which images have more visual information.

\[ H(x) = -\sum_{i=1}^{N} p(x_i) \log (x_i) \] (2)

In the above equation, \( p(x_i) \) is the probability of occurrence of one symbol, here the pixels. In this study, entropy serves as a visual feature that predicts the influence of the posts based on information contained in users’ posts. Each value of the entropy measure represents a vector of one visual style’s features for each user’s post [45].

\[ \text{C. CLASSIFICATION OF POSTS} \]

We used separate Support Vector Machines (SVM) for classifying various features extracted from the users’ posts into influential and non-influential sets. The investigation of classifiers demonstrates that there is no single classifier that can predict influential posts accurately. Each classifier learned from specific features of images is capable of separating particular kinds of influential and non-influential posts. Therefore, we need to combine these classifiers to reach a single robust classifier that can classify influential and non-influential posts accurately. Following [46], we used a combined classifier based on fixed rules such as sum, product, and max in decision level, rank level and feature level for each classifier, so that the scores of all classifiers could be merged to increase the accuracy of the classification and reduce the error ratio of the final classifier. The result of the combined classifiers is explained later.

\[ \text{D. IDENTIFYING INFLUENTIAL USERS} \]

After extraction of global, local, and visual features from images of users’ posts, they can be divided into two categories, influential and non-influential, based on the score from the classifier. While the influence of a user can be determined in various ways, such as investigating the topology of the network, we are trying to find influential users by detecting influential posts. The data for each user in the dataset contains a certain number of posts and for each post a number which shows the influence score of that post. We present a new statistical analysis approach to identify influential users based on the scores of their posts. The five parameters that we applied to create the features were maximum, minimum, median, mean and standard deviation, so each user is represented with a feature vector of size 5. We used SVM on the users’ feature vectors to classify them into influential and non-influential sets. We also examine two methods based on the number of influential posts and mean scores of posts, to identify influential users. The results of the proposed methods are explained thoroughly in the next section.
W. H. Alwan et al.: Identifying Influential Users on Instagram Through Visual Content Analysis

**IV. RESULTS**

Achieved results can be divided into two categories, 1) results from experiments on detecting influential posts using various feature extraction methods and SVM, and 2) detecting influential users based on a statistical analysis of each user’s post scores, number of influential posts, and mean scores of posts. Table 2 to Table 7 represent the results for each category. The Instagram dataset collected for this research contains two main parts. The first part, consisting of two subsets, contains posts by users on general topics, while the second part is collected based on a specific topic. These subsets that we refer to as DB_1, DB_2, and DB_3 respectively, are shown in Table 1. We tried to achieve more accurate results for identifying influential users on Instagram for a specific topic based on feature extraction, and therefore collected the influential and non-influential users’ posts on fashion. SVM was used to create a hyperplane in a multidimensional space so that it could separate the dataset into two classes of influential and non-influential posts. Table 2 represents the accuracy, which we refer to in brief as Acc, of different methods for identifying influential posts by extracting their features over the different subsets.

The results in Table 2 demonstrate that the best accuracy in the various feature extraction methods is obtained from transfer learning of the ResNet152 network, which is due to the extraction of very precise local features from images in posts’ contents. The next best accuracy belongs to Spatial Pyramid Matching (SPM), which is a local feature extraction method, and then there is the transfer learning of AlexNet, which is a holistic feature extraction method. Finally, accuracies obtained using GIST, blurriness and color histogram are in the same rank. The ROC curves of different feature extraction methods over DB_1 are shown in Figure 3. The second part of the influential post detection experiments is focused on combining the scores obtained from the feature extraction methods with the statistical scores to determine the influential posts. Table 4 and Table 5 represent the accuracy of different methods for detecting influential posts on DB_2 and DB_3 respectively.

### TABLE 2. Accuracy of combinational methods for detecting influential posts.

| Methods combined | Acc.DB_1 | Acc.DB_2 | Acc.DB_3 |
|------------------|----------|----------|----------|
| Rank _level      |          |          |          |
| SUM              | 69.52%   | 75.10%   | 71.80%   |
| PROD             | 69.59%   | 75.65%   | 72.41%   |
| MAX              | 69.42%   | 75.71%   | 72.84%   |

### TABLE 4. Accuracy of the mean scores and the influential posts’ ratio methods on DB_2.

| Methods  | T_Posts | T_Mean Scores |
|----------|---------|---------------|
| Acc      | Threshold | Acc | Threshold |
| ResNet   | 75.00%   | 0.30   | 75.36%   | 0.354   |
| Alex     | 74.64%   | 0.30   | 73.57%   | 0.3810  |
| SPM      | 95.36%   | 0.30   | 91.45%   | 0.324   |
| Gist     | 68.21%   | 0.40   | 68.21%   | 0.453   |
| Blur     | 61.07%   | 0.60   | 63.21%   | 0.535   |
| Color_hist| 74.64%  | 0.45   | 77.14%   | 0.480   |

### TABLE 5. Accuracy of the mean scores and the influential posts’ ratio methods on DB_3.

| Methods         | T_Posts | T_Mean Scores |
|-----------------|---------|---------------|
| Acc             | Threshold | Acc | Threshold |
| ResNet          | 88.08%   | 0.46   | 87.36%   | 0.46    |
| Alex            | 85.39%   | 0.45   | 85.11%   | 0.46    |
| SPM             | 82.58%   | 0.56   | 82.87%   | 0.51    |
| Gist            | 82.02%   | 0.57   | 83.33%   | 0.532   |
| Blur            | 74.72%   | 0.732  | 76.13%   | 0.59    |
| Color_hist      | 78.37%   | 0.620  | 78.93%   | 0.55    |
TABLE 6. Accuracy of using the statistical method for identifying influential users on DB_2 and DB_3.

| Methods          | 2nd Dataset |               | 3rd Dataset |               |
|------------------|-------------|---------------|-------------|---------------|
|                  | Number of Users |               |             | Number of Users |               |
|                  |              | Train | Test |                | Train | Test |
|                  | Inf | Non Inf | Inf | Non Inf | Inf | Non Inf | Inf | Non Inf |
| 45               | 234 | 46     | 234 |          | 121 | 234     | 122 | 234     |
| Accuracy DB_2    | Acc_all | Acc_Inf | Acc_Non Inf |            | Acc_all | Acc_Inf | Acc_Non Inf |
| Gist             | 65.24% | 64.52%  | 65.38% |          | 79.85% | 79.01%  | 80.13% |
| Alex             | 80.75% | 80.64%  | 80.77% |          | 84.38% | 83.95%  | 84.61% |
| ResNet           | 82.35% | 83.87%  | 82.05% |          | 86.08% | 86.42%  | 85.90% |
| SPM              | 73.26% | 74.19%  | 73.08% |          | 83.54% | 83.95%  | 83.33% |
| Visual features  |         |          |       |           |         |          |       |
| Blur             | 58.29% | 58.06%  | 58.33% |          | 77.21% | 76.54%  | 77.56% |
| Color            | 60.42% | 58.06%  | 60.89% |          | 81.01% | 80.25%  | 81.41% |

TABLE 7. Accuracy of the combined classifier, using the statistical method for identifying influential users on DB_3.

| Methods combined on 3rd Database | Accuracy on 3rd Dataset |
|----------------------------------|------------------------|
| Features level                   | 86.50%                 |
| Rank level                       | 85.23%                 |
| Decision level                   |                        |
| Sum                              | 85.65%                 |
| Product                          | 85.65%                 |
| Max                              | 84.39%                 |

We can conclude that the combined classifier outperforms each of the single classifiers, as shown in Figures 4-a, 4-b and 4-c for the three subsets respectively, which is due to the diversity of tastes in users’ posts. The second category of experiments focuses on detecting influential users. For this purpose, we use three different methods: The first method is Influential Posts’ Ratio which is a very simple approach based on enumeration of influential posts. If the number of influential posts is greater than 50% of the total posts, the user is detected as an influencer, otherwise the user is considered not to be an influence. The second method is Mean Scores which calculates the mean score of all posts for each user. If the mean score is greater than 0.5, the user is an influencer, and otherwise not. The results for these two methods are explained in the Tables 4 and 5 for DB_2 and DB_3 respectively. The third method is a statistical method for measuring the influence of an individual with respect to the scores of their posts. As explained in the methodology, 5 statistical scores are used to build a feature vector for each user, and then by using an SVM with an appropriate threshold, the users are classified into influential and non-influential sets. Table 6 shows the accuracy in finding influential users of different methods of feature extraction over the two subsets. In Table 6 the accuracy of identifying influential users is shown using three values: accuracy of all, accuracy of influential users and accuracy of non-influential users, which are briefly referred to as (Acc_all), (Acc_Inf), (Acc_Non Inf) respectively. We note that the number of influencers in all datasets is lower compared to non-influencers, as influential people in real world are also fewer than ordinary people. This seems to cause a problem with supervised machine learning, as the system shows a tendency towards non-influential users. We tried to solve this problem by using equal error rate (EER) based on thresholds.
to take approximately equal accuracies for influencers and non-influencers. We combined the scores in the user level for all methods using heterogeneous combined classifiers, on the part of the dataset on one specific topic (fashion) and obtained the results shown in Table 7.

V. DISCUSSIONS

The analysis of results obtained in this article suggests that identifying the influential users based on detecting influential posts is not a simple task, as our proposed method focuses only on feature extraction of a user’s posts while disregarding the topology of the network, which can be a very powerful tool for finding influencers. Although the proposed method is computationally more extensive due to the feature extraction step and the use of machine learning methods, the results were also more accurate and more robust than other methods for identifying influential users in a social network. The comparison of our results with those of the method of identifying influential posts on Instagram based on caption and Hashtag Analysis [25] over the same datasets showed that our results outperform those obtained from the textual content. Similarly, we used two methods: counting the number of influential posts, and the average score of posts based on a certain threshold, to identify the influential users. The best accuracies obtained from the Influential Posts’ Ratio method on DB_2 and DB_3 are 95.36% and 88.08% for influential users respectively, while the best accuracy from textual content analysis is 83.64%. Also, the accuracies obtained from the average likelihood of posts method on DB_2 and DB_3 are 91.45% and 87.36% for influential users respectively, while the best accuracy from textual content analysis is 83.18%, as shown in Tables 4 and 5 respectively. In addition, the accuracy of identification of the influential posts for the general dataset (DB_2) is 67.16% for local features and 64.65% for holistic features, which is better than the accuracy of hashtag and caption with accuracies of 65.51% and 63.75% respectively on the same dataset. However, the best accuracies of influential post detection on the dataset consisting of posts on a specific topic, fashion, (DB_3) is 81.96% for local features and 71.28% for holistic features, which are better than the accuracies of hashtag and caption, 65.37% and 61.22% respectively as shown in Table 2. From this comparison we can conclude that the analysis of users’ visual content is better than the analysis of hashtags and caption of the users’ posts for identification of influential users in social media. Also, we conclude that the contents of users’ posts were equally important or better in investigating the influential users than in structural graph-based networks.

VI. CONCLUSION

The significance of influential users in social media in every aspect of daily life, such as marketing, companies, brands, and social/political communities is acknowledged by everyone. Since the proposed method aims at identifying influential users based on their generated content, and due to the fact that Instagram is among the most popular social networks for sharing visually generated contents, in this study, the experiments were conducted on the Instagram platform. However, nothing in the proposed method prevents it from being applied to other social media platforms, popular for photo sharing. Our results showed the success of this model in investigating and detecting influential posts through analyzing the contents of the posts regardless of the graph and structure of social networks. The major aim of this article was to find features and properties of the influential users in social networks through establishing a labeled dataset as the ground truth, therefore we gathered three datasets of ordinary users and users which have been verified influencers in different topics. We also introduced a statistical analysis method to identify influential users, where the experimental results presented on Instagram showed the content of users’ posts is equally significant or even more beneficial in investigating influential users than in structural graph-based networks.

A potential direction for future works would be using the proposed model as a prototype in investigating influential users and posts. The contents of users’ posts play an effective role in detecting influential users and posts; therefore, we suggest combining this method with the method of identifying influential users using text to obtain a better model with a higher accuracy in influential user detection.

Another direction could be performing an analysis of the features extracted from the posts of influential users to guide others how to become influencers themselves.

REFERENCES

[1] J. Heidemann, M. Klier, and F. Probst, “Online social networks: A survey of a global phenomenon,” Comput. Netw., vol. 56, no. 18, pp. 3866–3878, Dec. 2012.
[2] T. N. Dinh, D. T. Nguyen, and M. T. Thai, “Cheap, easy, and massively effective viral marketing in social networks: Truth or fiction?” in Proc. 23rd ACM Conf. Hypertext Social Media (HT), 2012, pp. 165–174.
[3] D. Kempe, J. Kleinberg, and É. Tardos, “Influential nodes in a diffusion model for social networks,” in Proc. Int. Colloq. Automata, Lang., Program., 2005, pp. 1127–1138.
[4] L. C. Freeman, “A set of measures of centrality based on betweenness,” Socometry, vol. 40, no. 1, p. 35, 1977.
[5] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the Web,” in Proc. WWW, 1999. [Online]. Available: https://www.semanticscholar.org/paper/The-PageRank-Citation-Ranking-%3A-Bringing-Order-to-Page-Brin/eb82d3055849cd23578096462ba19b5319da556
[6] Q. Li, T. Zhou, L. Lü, and D. Chen, “Identifying influential spreaders by weighted LeaderRank,” Phys. A, Stat. Mech. Appl., vol. 404, pp. 47–55, Jun. 2014.
[7] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, “Identifying influential nodes in complex networks,” Phys. A, Stat. Mech. Appl., vol. 391, pp. 1777–1787, Feb. 2012.
[8] Y.-H. Fu, C.-Y. Huang, and C.-T. Sun, “Using global diversity and local topology features to identify influential network spreaders,” Phys. A, Stat. Mech. Appl., vol. 433, pp. 344–355, Sep. 2015.
[9] Q. Ma and J. Ma, “Identifying and ranking influential spreaders in complex networks with consideration of spreading probability,” Phys. A, Stat. Mech. Appl., vol. 465, pp. 312–330, Jan. 2017.
[10] S. Gao, J. Ma, Z. Chen, G. Wang, and C. Xing, “Ranking the spreading ability of nodes in complex networks based on local structure,” Phys. A, Stat. Mech. Appl., vol. 403, pp. 130–147, Jun. 2014.
[11] Y. Liu, M. Tang, T. Zhou, and Y. Do, “Identifying influential spreaders in complex networks, the role of neighborhood,” Phys. A, Stat. Mech. Appl., vol. 452, pp. 289–298, Jun. 2016.
[12] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, “Measuring user influence in Twitter: The million follower fallacy,” Icwm, vol. 10, nos. 10–17, p. 30, 2010.
[13] C. Xiao, Y. Zhang, X. Zeng, and Y. Wu, “Predicting user influence in social media,” J. Netw., vol. 8, no. 11, pp. 2649–2655, Oct. 2013.
