IJICTR
International Journal of Information & Communication Technology Research

Volume 11- Number 4 – Autumn 2019 (57 -65)

A New Framework for Discovering Important Posts and Influential Users in Social Networks

Leila Rabiei, Mojtaba Mazoochi*, Farzaneh Rahmani
ICT Research Institute (Iran Telecom Research Center)
Tehran, Iran
(l.rabiei, mazoochi, rahmani)@itrc.ac.ir

Received: 1 May 2019 - Accepted: 16 August 2019

Abstract—**The popularity of social networks has rapidly increased over the past few years. Social networks provide many kinds of services and benefits to their users like helping them to communicate, click, view and share contents that reflect their opinions or interests. Detecting important contents defined as the most visited posts and users whom disseminate them can provide some interesting insights from cyberspace user's activities. In this paper, a framework for discovering important posts (most popular posts by views count) and influential users is introduced. The proposed framework employed on Telegram instant messaging service in this study but it is also applicable to other social networks such as Instagram and Twitter. This framework continuously works in a real social network analysis system named Zekavat to find daily important posts and influential users. The effectiveness of this framework was shown in experiments. The accuracy achieved in the advertisement detection model is 89%. Text-based clustering part of the framework was tested based on the human factor verification and clustering time is less than linear. Graph creation based on publishing relationships is more effective than mention relationship and in this process influential users can be identified in a precise manner.**

Keywords-social networks; clustering; LSH; machine learning; important posts;influential users

I. INTRODUCTION

Since information is rapidly disseminated through social networks, studying social networks and understanding the relationships between users have turned to noticeable research subjects. In 2018, an estimated 2.65 billion people were using social media worldwide, a number projected to increase to almost 3.1 billion in 2021 [1]. Due to the importance of social networks, researchers are interested in finding important posts that are of interest to users and as a result of their high number of visits. Because these posts represent topics that users are interested in, and can have inappropriate social consequences if there is a fake news or topic with social anomalies. Identifying these posts can also be helpful in managerial decision making. Furthermore, it is necessary to identifying users who have published Important posts. As a result, users who can spread the information to most people on the network can be identified. These users are called influential users. In addition to maximizing information dissemination, finding influential users is widely used in marketing [2] as well as hindering the dissemination of unwanted contents [3].

Social network is an online platform which people use to build social relationships whit other people who share similar personal or career interests, activities, backgrounds, news or real-life connections. Social networks vary in format and number of features. Facebook, YouTube and WhatsApp are among the most popular social networking services based on

* Corresponding Author

number of active users, as of April 2020 as published by Statista [4].

In the following a description for importance of Telegram is presented to show why this messaging service is selected for the first time to deploy the proposed framework. Telegram is a cloud-based instant messaging and voice over IP service. This messaging app focuses on speed and security. In March 2018 Telegram stated that it had 200 million monthly active users [5]. On march 14, 2019, Pavel Durov (Telegram founder) claimed that "3 million new users signed up for Telegram within the last 24 hours"[6].

In September 2015, Telegram added channels. Channels are a form of one-way messaging where admins are able to post or broadcast messages. In fact, a channel can have an unlimited number of subscribers. Telegram channels can be public or private. Public channels have a username. Anyone can find them in Telegram search and join. Private channels are closed societies someone need to be added by the creator or get an invite link to join. 15 billion Telegram messages were being sent daily in February 2016- the last time Telegram made this stat public. Telegram is the most popular messaging app, namely Ethiopia, Iran and Uzbekistan [7].

An estimated 50 million Telegram users in Iran as of April 2018. At this point, the app was banned in Iran, though it can still be accessed through VPNs. Telegram penetration in Iran estimated at around 56%. This is down from 60 % pre-ban, but well up from 47% in its immediate wake [8]. There are 35 channels with up to 1 million subscribers made up of Persian language with content geared around news, entertainment, education, shopping and official channel of gold Telegram (@mono_ir : 4.9m), Donyayetaraneh (@dtaraneh : 4.3m), melobit (@melobit : 3.6m), akharinkhabar (@akharinkhabar : 3.5m), daily news of Iran (@iranfnews : 3.4m) are top 5 channels with most subscribers based on Telegram analytics [9]. Due to this statistics, Telegram is known as an important social networking and instant messaging service in Iran. Therefore, finding top posts that published in Persian Telegram channels and influencer users (channels in Telegram) is valuable. In this context it is necessary to model Telegram as a graph of channels.

With the growth of advertising opportunities on social networks and messaging services such as Instagram, Twitter, Telegram, etc., advertisement posts are published every moment in these platforms. For example, popular channels and bots in Telegram have become a strong way for mobile advertisement to introduce new products, services and etc. to their audience. This posts receive high view count via publishing from different channels and bots. Generally, these posts are of a little value in terms of contents.

Finding top posts that published in Persian Telegram channels is important. One factor to know the importance of posts is the number of views a post receives via view counter measure. Telegram analytics provides posts ranking based on views, however there are advertisements posts in this ranking and near duplicate versions of posts are not included. A lot of posts that broadcasted in Telegram are advertisement posts. These posts forwarded by many channels and therefore view number is high in these posts, however usually these posts are of low importance.

Each message in a Telegram channel has a view counter that gets updated when the message is viewed, including its forwarded copies. Posts published by public Telegram channels can be forwarded by other channels. In this case, view counter of initial post will increase due to each visit of members. In another case, each channel can copy the content of post published by a channel, add and subtract content and resend as a new post. It is worth noting that the view count is restarted and gets updated for this new post separately from the initial post. Moreover, republishing of posts via copy or forward can be considered as relationships between channels as well as channels that published top posts for the first time are influenced channels.

In this paper which is an extended version of our previous work [10] a new framework for detecting important posts in both views and content(non-advertisement) and influential users, important spreaders whom sharing important posts, is introduced.

In this version of framework, a module for creating graph and finding influential users is added to the framework. In this module Telegram graph can be created based on publishing duplicate or near-duplicate posts in public channels. Formation of graph based on this kind of interaction is a new method that is introduced and used in the proposed framework. This framework employed for Persian language Telegram public channels and their posts. However, the framework can be employed to other social networks and other messaging services. It is worth noting that the proposed framework continuously works in a real social network analysis system named Zekavat to find daily important posts and influential users. Zekavat system includes data gathering from Telegram, Instagram, Twitter and web domains. This high volume of data is stored in the big data platform. Elastic Search technology allows to search for text from the integrated data. In this system, after preprocessing of Persian text (official or slang), descriptive analysis including statistical analysis and graph-based analysis are presented. Finding trends for user activities, daily and hourly monitoring of user activities, recognizing hot topics, finding influential users, detecting communities and co-occurrence graphs are among this level of analysis. Predictive analysis based on machine learning methods and artificial intelligence are also provided such as content classification, publisher classification and sentiment analysis. Analytical dashboards allow users to monitor results of all analysis. There are also restful APIs provided by Zekavat.

In the proposed framework after gathering specified data (here Persian posts published in public channels), advertisement posts are identified based on a machine learning-based model and excluded from the total number of posts. Then similar posts are identified and clustered based on the similarity of their textual content. Views for similar posts are aggregated and

assigned to the cluster representative, which is the first published post based on the timing of posts for each group. Top posts are identified by ratings based on total views. Moreover, after clustering of posts based on their content similarity, graph of users (here channels) can be formed based on sharing similar posts by different users in each cluster. To find influential users on social networks first a graph of users and their relationships should be created. The proposed framework introduced a new kind of interaction graph based on sharing similar posts between users. In this manner edges can be represented as sharing original posts known as cluster representative previously by another user in that cluster. As a result, influential users can be identified based on publishing top posts in the constructed graph.

The rest of this paper is organized as follows. Related works in different parts of the framework is overviewed in Sec II. The proposed framework will be introducing in Sec. III. Experimental results corresponding to each part of the framework will also be shown in Sec. IV. Conclusion and future works will be given in Sec. V.

II. Related works

In Telegram, channels are a form of one-way messaging where admins are able to post or broadcast messages. In fact, a channel can have an unlimited number of subscribers. Telegram channels can be public or private. Public channels have a username. Anyone can find them in Telegram search and join. Private channels are closed societies someone need to be added by the creator or get an invite link to join.

Advertisements are spam posts that repeatedly published in the same format or content over a short period of time [11]. Publishing this type of posts on social networks and instant messaging services like Telegram have become more ordinary with economic purposes. By extracting meaningful features from the text using Natural Language Processing (NLP), it is possible to conduct spam detection using various machine learning techniques [12]. Supervised learning can be used to detect advertisements posts considering as a classification problem of separating posts into two classes: advertisement and non-advertisement posts. In this approach, labeled and classified data are used to train the model.

Different machine learning techniques have been used by researchers to find out the spam content [13]. Researchers in [14] have been focused on classification and summarization of opinions using NLP and data mining techniques to identify opinion spams. Based on the analysis of 5.8 million reviews and 2.14 million reviewers from amazon.com, they showed that opinion spam in reviews is widespread. In that research, such spam activities were analyzed. They used more than 2.5 million reviews from social network users as training data and achieved over 90% accuracy in spam detection using Support Vector Machine (SVM) and Naïve Bayesian algorithms based on AUC criteria. In [15] three different features include POS tags, LIWS and bigrams are used to spam detection. Classification using SVM, Naïve Bayesian and Logistic Regression models is cross validated using K-fold method. The process of finding the best hyper parameters is usually called the model selection phase in the machine learning literature. One of the most popular resampling techniques is the k–Fold Cross Validation (KCV) procedure [16], which is simple, effective and reliable. The combined model LIWC+ BIGRAMS+ SVM is 89.8% accurate at detecting deceptive opinion spam [17-18]. [19] proposed a new approach for performing spam detection in Arabic opinion reviews by merging methods from data mining and text mining in one mining classification approach. F-measure is improved up to 99.59%. Their approach is based on the three models NB, K-NN and SVM. NBs achieved promising results compared to other supervised methods. The NB classifier receives 99.20% accuracy and F1-score 99.59%. In [20] neural network, SVM and decision tree classification methods are trained using features such as message length, number of @ in message, number of links in message, number of Telegram links in message, time of sending message(hour), time of sending message(minutes), is a forwarded message to detect advertisement posts. Neural network achieved 83.7% in F-measure.

In another way, there are many posts with similar and nearly similar content in Telegram. In detecting popular posts, identifying similar posts is important to find duplicate or near duplicate content. Different approaches for identifying similar text can be roughly classified into four main groups based on the level of analysis applied to text: character-based approaches, token-based approaches, structure-based approaches, and knowledge-based approaches. Each of these four categories presents different techniques for measuring the similarity between two texts.

Different algorithms for character-based approaches are N-Gram, Smith-Waterman, Lowenstein distance, hamming distance, hash technique, cosine similarity, etc. [21]. In token-based approaches, cosine similarity methods, Dice correlation coefficient, similarities like Jaccard, etc. The structure-based approaches, are based on the information obtained from very large corpuses. These approaches include techniques such as latent semantic analysis (LSA), explicit semantic analysis (ESA), etc. Finally, in semantic approaches or knowledge-based approaches, the semantic similarity between text words is commonly calculate using ontologies like WordNet or FarsNet (in Persian).

As mentioned, there are different techniques in character-based approaches. In Lowenstein distance method, each text is considered as a string of characters, and the distance between them is calculated using the longest common substring (LCS). Distance (similarity) is the number of insertions, deletions, or shifts necessary to convert source text to destination text. Cosine similarity is based on vector space. Each text is converted into vector space. Then. its vector distance is measured against the vector of other texts. In Hamming distance, the number of places where different characters exist is counted, and this value is considered as the distance between two texts. This distance can be applied to two texts of the same length.

In the hashing method, also called fingerprinting, texts are converted to vectors of numbers of equal

length. One of the important points in measuring similarity between two texts is scalability, meaning that as the text grows, the temporal and computational complexity of the similarity will not increase. Also, number-based similarity is much better and faster than text-based similarity. The fingerprinting method is currently the most common method used to detect scientific fraud while it is also very fast [22]. In a recent research [23], The hash value is used to measure the similarity between short texts. In this paper, different methods of measuring similarity between two texts with different granularities, such as cosine similarity or Lowenstein distance, etc., have been tested and concluded that for short texts, the hashing method has a significant advantage over other methods in terms of accuracy and runtime. In another study done in [24], phrase hash is used to measure the similarity of the short texts. Also in [25], an improved hash method is used for similarity measurement.

MinHash is an approximation, which improves the performance significantly while maintaining the accuracy of similarity computation reasonably. MinHash produces a signature, an h-dimensional vector, from text feature where h is a value much smaller than m (the dimensionality of a feature). Performance can be improved because the similarity is computed by comparing only a few elements when we use signatures [26]. In order to build the signature $Sj=<s_1,s_2,s_3,\ldots>$, MinHash needs random permutations, which define the exchange sequence of elements in a vector. After the permutation, the J^{th} element S_j set as the lowest index among those elements whose value is 1 in the permuted vector.

A number of random hash functions is given $F_j:V\rightarrow R$ assigning a real number to each visual word. Let X_a and X_b be different words. The random hash functions have to satisfy two conditions: $F_j(X_a)\neq Fj(X_b)$ and $P(F_j(X_a)<F_j(X_b)) = 0.5$. The functions F_j also have to be independent [26].

LSH is a hashing-based algorithm for identifying near or similar posts. LSH divides a signature into multiple lower-dimensional vectors, called a band. By comparing the bands from different signatures, we know whether the same parts of two signatures would be identical. For fast comparisons of bands, LSH uses the hash function, which assigns the identical bands to the same hash bucket [26]. A similarity algorithm almost tries to reduce this complexity to a linear one. Sublinear complexity is achieved by reducing the number of comparisons needed to find similarities.

One of the issues in this work is modeling social networks (specially, instant messaging services like Telegram) as graphs. Social networks are considered as graphs where the nodes are users and the edges represent the social connections among them. These graphs are called social graphs. Edges also can show the interactions between users like mention, tag, reply, etc. These kinds of graphs are called interaction graphs [28]. Formerly, Telegram graph was created by researchers [20]. This graph is created based on mention relation i.e. channels which mentioned another channel or were mentioned by another one. In the mentioned research, it was shown that 89% of edges are spam mentions, 8% are ham mentions and 3% are self-mentions. The size of nodes in the graph indicates number of followers of the channel (i.e. nodes). Intuitively the number of followers of the channel can be considered as an indicator of the channel popularity and its quality. Therefore, there is no relationship between the degree of nodes in Telegram mention graph and its number of followers. Specifically, many popular channels in Telegram are standalone nodes or they have a low degree. On the other hand, a lot of nodes in the largest connected component of mention graph i.e. high-degree nodes, are not popular channels. A considerable portion of edges are spam mentions which do not have any actual content [20].

After creating the graph, in order to find influential nodes, there are considerable number of algorithms in the literature. These algorithms can be classified into six categories [12] include local measure, short path based methods, iterative calculation based methods, coreness based measures, machine learning based methods and others. A famous local-based algorithm is degree centrality[13], in which the user with the highest degree is the influential user. Degree centrality is the simplest centrality measure. Time complexity of this measure for unweighted network is O(m) where m is the number of edges. It is clear that this measure is less complex and useful in many applications. The degree of each node is calculated as follows:

$$\deg(V) = \sum_{w\ in\ neighbour(v)} e(v.w) \qquad (1)$$

where e (v, w) is the edge between v and w. In directed graphs, in-degree or out-degree can be considered as the score of the node. Many algorithms use degree centrality to rank the nodes. Diffusion-degree algorithm [29] is one of them that considers the degree of the node, degree of its followers and propagation probability to rank the nodes.

III. THE PROPOSED FRAMEWORK

In this section the proposed framework is described in details. This framework consists of 5 main modules, data gathering, a module for removing advertisement posts, post clustering module based on text similarity and LSH algorithm, finding important posts based on views and the last module is for creating graph and discovering influential nodes. The proposed framework is depicted in Fig. 1

A. Gathering Posts

Data gathering is the first module that provides data for other modules. In this module posts are collected from target social network using offered APIs or scraping if the target social network provided via web (e.g. Instagram). Posts published in Telegram channels can be gathered using Telegram API. The number of Persian public channels has not been published by Telegram officially. Over one million channels are discovered during data gathering process, although many of these channels are either not active or have limited activity. Through these channels, about 140,000 active channels are identified and posts published by them are collected on a daily basis. Over 1 million posts can be collected daily.

B. Advertisement Detection

According to studies on spam detection, the most common models are SVM, Logistic Regression, Random Forest and Naïve Bayesian algorithms. In this paper, the four classification algorithms mentioned above are used to detect Telegram advertisement posts and the best model is selected using K-fold cross validation.

In the first step, the text of the posts will be pre-processed to be prepared for feature extraction. Text preprocessing is the most important step in text classification. The language used in social network posts is mostly spoken language. Therefore, use of conversational language preposition and formal language is required simultaneously. Preprocessing for Persian-Language posts in Telegram include normalization, tokenizing these posts by words, stop word removal, and also verb stemming. This step is performed using Persian NLP package which is developed in Zekavat social network analysis system. After preprocessing, the dataset is ready as a tokenized set of words. The TF-IDF method is used to extract features. This method calculates the repetition of words based on their weight in the whole document. After extracting this feature for each post, these features are considered as input to the machine, along with advertisement and non-advertisement labels.

C. Clustering and Finding Top Posts

In the second module, the posts are clustered by content similarity. Then a representative is assigned to each cluster based on time. In this research, the purpose is to identify similar posts structurally and with the same words and characters to identify duplicate or near-

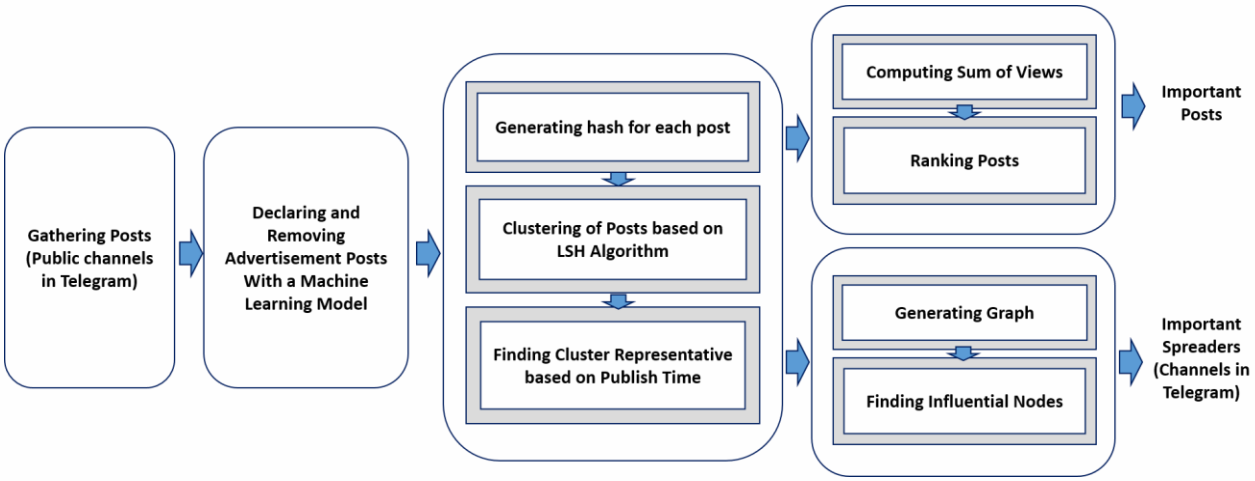

Figure 1. The proposed framework.

duplicate content. For this reason, the character-based approach is used. This approach is used in methods that aim to detect near duplicate, such as scientific fraud detection.

Given the type of text that are short texts, similarity testing using hashing can be the best way and can be very effective. That's why we first compute a hash for each post text content. The collision rate in hash calculation should be as low as possible, even zero, and can work well with the large volume of data discussed here. Therefore, MinHash algorithm is selected, which is shown not only to minimize the amount of collisions, but also to handle large volumes of data stream [28].

Next, the different hashes associated with each post on different channels must be compared. Each post is compared to posts which their length is at most 100 characters different from that post and published within 3 days before the publication time of the post, because usually after 3 days of publication, each post has achieved its position and its diffusion decreases in Telegram channels.

Before starting this step, it is worth noting that duplicate posts within the same channel are not counted as different posts and in fact duplicates are removed. Next, different channel posts are compared. Here, an appropriate number should be considered as threshold value for detecting similarity. LSH algorithm is used for clustering. This algorithm considers clusters based on the Jaccard similarity criterion for hashed posts with a similarity value above 0.42. The degree of similarity can be varied according to requirements. It is worth noting that by reviewing the timing of each post in the same cluster, we can determine what channel was the first to publish this post. The number of posts in each cluster also indicates the large number of visits and the inclusion of that cluster.

D. Forming Graph and finding Influential users

Formerly, Telegram graph was created based on mention relation i.e. channels which mentioned another channel or were mentioned by another one[20]. However, a considerable portion of edges are spam mentions which do not have any actual content. Therefore, sharing relationship between channels seems to be more convenient. This kind of relationship is established after deleting and removal of advertisement messages. Also, publishing of similar posts by different channels is considered based on timing of publishing. A sample graph based on this kind of relationship in Telegram is demonstrated in Fig. 2.

Nodes represent public channels and edges show sharing-relationship between them in Figure 2. It means

if channel A shares a post formerly shared by another channel B as original post and channel B was selected as representative in clustering module of framework, there is an edge from A to B. Edge weight shows the number of posts shared between them during this time. In order to construct this graph, as mentioned before, in clustering module posts are clustered by content and each post assigned a cluster number. More over a cluster representative is defined.

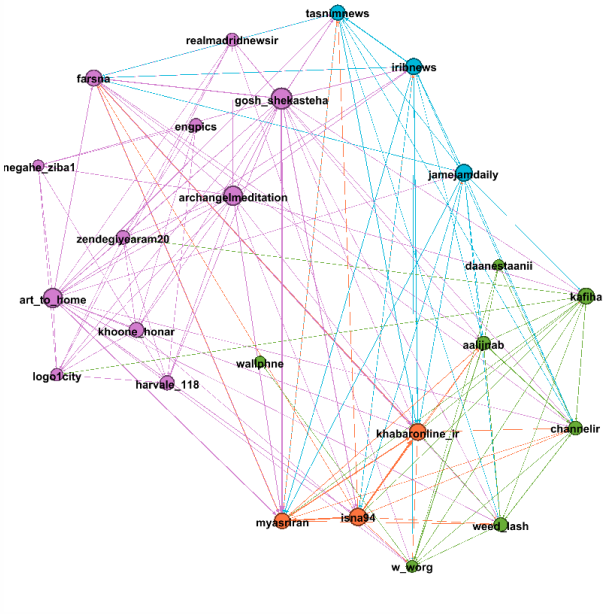

Figure 2. A sample Telegram Graph based on sharing relationship.

Representative post is a post that is similar in content to other cluster posts and has been published earlier than other posts in that cluster. Publisher channel for each post in the cluster is known in the clustering module. The channels whose posts are selected as the cluster representative are considered as the source nodes, and edges are established to other channels which their posts are located in that cluster. In this way we considered edges between channels who publish a post from another channel defined as representative channel previously.

The number of posts publish from a source channel specifies the weight of the edge between channels. Out-degree of nodes shows the number of posts from the channel published by other channels. Alternatively, the number of views (sum of views in a cluster) for all posts of a channel can be defined as the out degree of a node. In this module degree-centrality is used to specifying influential nodes because this algorithm is very simple and it covers our need to identify influential users based on important posts. However, any other algorithm for detecting influential users can be applied in this module. An extended version of this module can identify communities of users in the generated graph using modularity based community detection algorithms.

IV. EXPERIMENTAL RESULSTS AND DISSCUSSION

To evaluate the performance of the proposed Framework, experiments were conducted on real Telegram posts in Persian language.

A. Advertisement detection

We manually labeled a randomly chosen subset of our data collection including 20000 posts. Labels are in two classes of advertisement and non-advertisement.

Finally, there are 8000 and 12000 posts in mentioned classes, respectively. We used 80% of these posts to train the model. 5-fold cross validation on this 80% of data is used for setting hyper parameters and selecting the best model Samples for each class are represented in Table I.

Fig. 3 shows the average classification accuracy based on TF-IDF feature. Results in Figure 3 shows that Logistic Regression classifier is better than other classifiers in this problem. Logistic Regression achieved F_1 89% in cross validation. Precision, recall and F_1 for two classes is illustrated in Fig. 4. These results are on 20 % unseen data using the selected model.

F_1 for advertisement and non-advertisement classes was 87% and 92%, respectively, while the mean F_1 for entire model was reported 89%. Model's weakness in identifying advertisement posts is because of diversity of these type of posts. Therefore, the features selected for training the model failed to be sufficiently comprehensive for the training samples.

TABLE I. SAMPLES IN EACH CLASS OF DATASET.

Class	Telegram Post(Text Content)
ADV	با خرید یا تمدید اشتراک در قرعه‌کشی یک میلیون تومنی هفتگی 📣فیلیمو#تماشا و دانلود قسمت دوازدهم سریال جذاب و پرطرفدار #هیولا با #اینترنت_رایگان در ✸ hayola12 درصد تخفیف فیلیمو بر روی اشتراک سه ماهه✓ از طریق لینک زیر یا کد تخفیف ۴۰ 🎁فیلیمو و جایزه ۳۰۰ میلیون تومنی تابستانه فیلیمو شرکت کن 👇👇👇➡ اشتراک بخر یا تمدید کن✓ فقط برای شرکت در قرعه‌کشی تا دوشنبه هفته آینده فرصت داری قسمت دوازدهم #هیولا رو تماشا کنی وارد شو www.filimo.com/r/hayola12@filimo
ADV	راهنمایی خرید از چین و حمل کالابرای 💻🕹✈ پیش فروش۱۵روزه✈□□فروش فوری تحویل یک روزه🚚عرضه و واردات انواع ماینر بیت کوین به صورت عمده🛩 همکاران گرامی در اسرع وقت✈➡09021763614➡09357405606➡09181763614➡09183740634👇👇👇👇👇👇👇👇👇👇https://t.me/joinchat/aaaaafhrue-r-s55rwgppw
Non-ADV	تصویب صورتهای مالی بانک تجارت صورتهای مالی بزرگترین شرکت بورسی ایران با حضور بیش از ۷۳ درصد سهامداران به تصویب رسید.

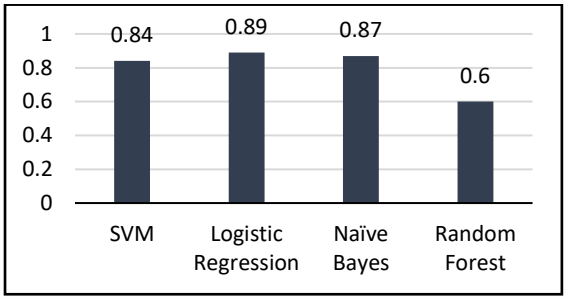

Figure 3. Mean F_1 in KVC for different classifiers.

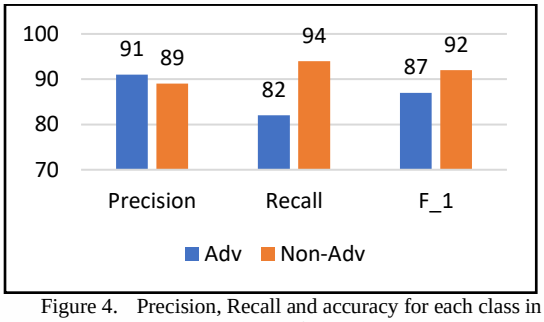

Figure 4. Precision, Recall and accuracy for each class in classifier.

The problem of model integrity can be solved by adding more training examples in the advertisement class.

B. *Similarity Detection and Clustering*

After Detecting and Removing Advertisement posts, clustering of duplicates and near duplicates is conducted. Initially, hash is calculated, for text in each post based on MinHash algorithm. This hash with a unique index are sent to LSH algorithm in order to clustering. First post is placed in one cluster; when the next post is entered, if the similarity is less than the threshold value, recognized as similar and placed in the previous cluster. But if there is no similarity between them, the new post will be placed in the new cluster. This process will continue by entering subsequent posts. Finally, after running the algorithm for posts in each time interval, clusters of similar posts are found, and the total number of views within the cluster is considered as the number of visits to cluster representative post. For each cluster, the first post published according to the time of publication is considered to represent that cluster. Also, for each post, the number of posts that have copied from is considered as its Penetration Rate. Clustering time is presented in Fig. 5. This time is calculated on a computer with Intel (R) Xeon 7-core CPU and 8 GB main memory. Jet Brains PyCharm2019 is utilized as development environment and the programming language was python3. Clustering time resulted from LSH algorithm shows less than linear time when number of posts increases. A sample of clusters generated in the framework in Table II represents that results are promising.

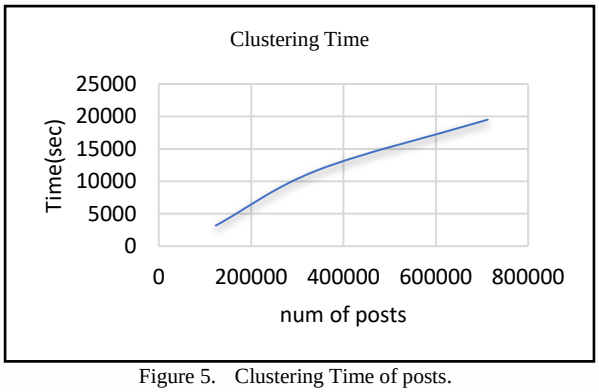

Figure 5. Clustering Time of posts.

C. *Graph creation and finding influential users*

In order to construct Telegram graph based on publishing relationship, results of the previous module called clustering module was utilized. A sample graph which is made from Table II is displayed in Fig. 6. In results represented in Table II. one post from "euronewspe" was selected as representative and mother channel field was set to 1. Therefore, this channel can be selected as source node in the corresponding graph and there are edges from this node to other channels in this cluster which published posts similar to the source node. In this way the whole Telegram graph created for gathered data. Since degree centrality is used to find influential nodes in the graph, influential nodes are channels that maximum number of posts originally published by them, later published by other channels in Telegram. In a more precise way, sum of views for the origin posts can be selected as degree for source nodes. Therefore, influential channels are channels where their posts get maximum number of views.

Experiments showed that there are channels that have a similar function to the bots. These channels generally feed on bats similar to themselves. Results of clustering shows that in some clusters, the cluster representative post was published with only one timestamp difference before the other posts in that cluster and the other posts in that cluster were all published at the same time. In other cases, in some clusters, all the posts in that cluster are published at exactly the same time, which is not possible except for bots. As a result, the introduced framework can be employed to identifying channels that work as bots.

V. CONLUSION AND FUTURE WORK

In this paper a new framework for discovering important posts and influential users was presented and experiments proved the effectiveness of this framework. The accuracy achieved in the advertisement detection model is 89% which is better than research in [20]. The mentioned research reached 79.9% accuracy in machine learning methods and 80.5% accuracy in neural networks. In text-based clustering the algorithm shows good performance based on the human factor verification and clustering time is less than linear. Graph creation based on publishing relationships is more effective than mention relationship and in this process influential nodes can be identified in a precise

manner. Results of this framework have been reviewed and validated on a daily basis by human users. These validations indicate the high accuracy of the output of the framework. This framework is applicable to other social networks such as Instagram and Twitter. For future research, other hashing algorithms such as Universal hashing can be used. For post clustering we can use semantic methods. In addition, more advanced classification methods can be proposed to detect advertisement posts.

ACKNOWLEDGMENT

This research was fully supported by project number 4160970510 in ICT Research Institute (Iran Telecom Research Center). We are thankful to our colleagues who provided their expertise that greatly assisted the research.

TABLE II. RESULT OF CLUSTERING FOR NEAR DUPLICATE POSTS.

post	cluster_id	mother_channel
◆ اظهار نظر عجیب رئیس جمهور آمریکا ترامپ: «اگر ایرانی‌ها به دنبال پهپادشان هستند لباس غواصی بپوشند و در کف اقیانوس دنبال آن بگردند.»/یورونیوز? AkhbareFori@	52	0
اظهار نظر عجیب رئیس جمهور آمریکا ترامپ: اگر ایرانی‌ها به دنبال پهپادشان هستند لباس غواصی بپوشند و در کف اقیانوس دنبال آن بگردند.? #بین_المللی khabarsarasari@	52	0
ترامپ: «اگر ایرانی‌ها به دنبال پهپادشان هستند لباس غواصی بپوشند و در کف اقیانوس دنبال آن بگردند.» جزئیات را بخوانید: https://bit.ly/2SxPKQq euronewspe@	52	1

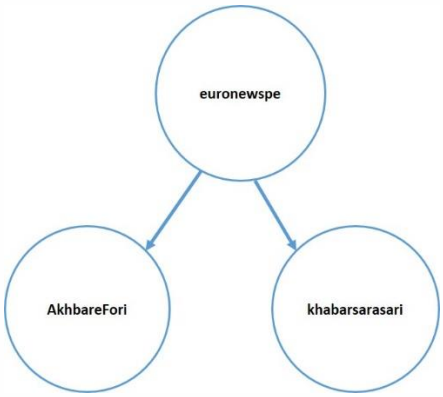

Figure 6. Corresponding Graph for result of clustering.

REFERENCES

[1] "number of worldwide social network users" https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, Retrieved 5 May. 2020.

[2] Zhu, Zhiguo. "Discovering the influential users oriented to viral marketing based on online social networks." Physica A: Statistical Mechanics and its Applications 392, no. 16 (2013): 3459-3469.

[3] Richardson, Matthew, and Pedro Domingos. "Mining knowledge-sharing sites for viral marketing." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 61-70. ACM, 2002.

[4] "global social netorks ranked by number of users" , https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, Retrieved 5 May. 2020.

[5] "200,000,000 Monthly Active Users", https://telegram.org/blog/200-million , Retrieved 4 Nov. 2019.

[6] Lomas,Natasha. "Telegram gets 3M new signups during Facebook apps' outage" TechCrunch. Retrieved 4 Nov 2019

[7] "The global state of digital in 2019 – Hootsuite", https://hootsuite.com/resources/digital-in-2019 , Retrieved 4 Nov. 2019

[8] ISPA Iranian Student Polling Agency, http://ispa.ir/Default/Index/en, Retrieved 4 novamber 2019

[9] Iranian Telegram-channels. Statistics, Analytics, Top charts. Telegram Analytics, http://Ir.tgstat.com, Retrieved 24 July 2020.

[10] Leila Rabiei, Farzaneh Rahmani, Mojtaba Mazoochi, Meissam Kheyrollah Nejhad, " A New Framework for Discovering Important Posts in Social Networks ", 10th International Conference on Information and Knowledge Technology (IKT 2019)

[11] Zhao, Laijun, Qin Wang, Jingjing Cheng, Yucheng Chen, Jiajia Wang, and Wei Huang. "Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal." Physica A: Statistical Mechanics and its Applications 390, no. 13 (2011): 2619-2625.

[12] Al-Garadi, Mohammed Ali, Kasturi Dewi Varathan, Sri Devi Ravana, Ejaz Ahmed, Ghulam Mujtaba, Muhammad Usman Shahid Khan, and Samee U. Khan. "Analysis of online social network connections for identification of influential users: Survey and open research issues." ACM Computing Surveys (CSUR) 51, no. 1 (2018): 16.

[13] Tewari, A., & Jangale, S. (2016). Spam Filtering Methods and machine Learning Algorithm-A Survey. International Journal of Computer Applications, 154(6).

[14] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.

[15] Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 219–230). ACM, Stanford, CA

[16] Jung, Y., & Hu, J. (2015). AK-fold averaging cross-validation procedure. Journal of nonparametric statistics, 27(2), 167-179

[17] Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 309–319). Association for Computational Linguistics

[18] Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ (2007) The development and psychometric properties of LIWC2007

[19] Hammad ASA (2013) An Approach for Detecting Spam in Arabic Opinion Reviews. Doctoral dissertation, Islamic University of Gaza

[20] Dargahi Nobari, A., Reshadatmand, N., & Neshati, M. (2017, November). Analysis of Telegram, An Instant Messaging Service. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2035-2038)

[21] Udagawa, Y. (2013). Source code retrieval using sequence based similarity. arXiv preprint arXiv:1308.3554.

[22] Rafieian, S. (2016). Plagiarism checker for Persian (PCP) texts using hash-based tree representative fingerprinting. Journal of AI and Data Mining, 4(2), 125-133.

[23] Gao, J. (2018, November). Detecting Short Near-Duplicates with Semantic Relations. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) (pp. 122-125).

[24] Huang, X. (2015). A Mixed Generative-Discriminative Based Hashing Method. IEEE Transactions on Knowledge and Data Engineering, 28(4), 845-857.

[25] Gionis, Aristides, Piotr Indyk, and Rajeev Motwani. "Similarity search in high dimensions via hashing." In Vldb, vol. 99, no. 6, pp. 518-529. 1999.

[26] Hwang, W. S., Park, J., & Kim, S. W. (2015, January). A method for recommending the latest news articles via MinHash and LSH. In Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication (p. 60). ACM.

[27] Chum, O., Philbin, J., & Zisserman, A. (2008, September). Near duplicate image detection: min-hash and tf-idf weighting. In Bmvc (Vol. 810, pp. 812-815).

[28] Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Measurement and analysis of online social networks." In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 29-42. ACM, 2007.

[29] Pal, Sankar K., Suman Kundu, and C. A. Murthy. "Centrality measures, upper bound, and influence maximization in large scale directed social networks." *Fundamenta Informaticae*130, no. 3 (2014): 317-342.

Leila Rabiei received her B.Sc. degree in Computer Engineering from Islamic Azad University of Tehran, Iran, and her M.Sc. degree in Computer Engineering from Iran University of Science and Technology, Tehran, Iran. She is currently works as a researcher and project manager in the Communication Technology Department of ICT Research Institue (ITRC), Tehran, Iran. Her research interests include Big Data Analysis, Data Mining and Social Networks Analysis.

Mojtaba Mazoochi received his B.Sc. degree in Electrical Engineering from Tehran University, Iran in 1992. He received his M.Sc. degree from Khajeh Nasir Toosi University of Technology, Iran in 1995 and his Ph.D. degree from Islamic Azad University, Tehran, Iran in 2015 in Electrical Engineering (Telecommunication). He is an assistant professor and deputy of IT faculty in ICT Research Institute (ITRC), Tehran, Iran. His research interests include Data Analytics, Quality of Service (QoS), and Network Management.

Farzaneh Rahmani received her B.Sc. degree in Computer Engineering from Bahonar University, Kerman, Iran and her M.Sc. degree in Computer Engineering from Tarbiat Modares University and Ph.D. degree from ICT Research Institue (ITRC), Tehran, Iran. Her research interests include Social Networks Analysis, Machine Learning, Natural Language Processing and Computer Vision.