The ISC Int'l Journal of Information Security

November 2020, Volume 12, Number 3 (pp. 59–72)

http://www.isecure-journal.org

Selected Paper at the ICCMIT'20 in Athens, Greece.

Identifying Entrepreneurial Influencers on Twitter**

Bodor Almotairy [1,*], Manal Abdullah [1], and Rabeeh Abbasi [2]

[1] *Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia.*

[2] *Quaid-i-Azam University, Islamabad, Pakistan.*

ARTICLE INFO.

Keywords:
Entrepreneurship, Entepreneurial Influencers, Social Network Analysis, Twitter, Social Media, Ranking

Type: Research Article

doi: 10.22042/isecure.2021.271108.629

Abstract

Entrepreneurship involves an immense network of activities, linked via collaborations and information propagation. Information dissemination is extremely important for entrepreneurs. Finding influential users with high levels of interaction and connectivity in social media and involving them in information spread helps disseminating the information quickly. Thus, facilitating key entrepreneurial actors to find and collaborate with each other. Identifying and ranking entrepreneurial top influential people is still in infancy. This paper proposes an E-Rank framework for topic-specific influence theories that are specialized with respect to Twitter. Firstly, it extracts four dimensions to characterize influencers, including user popularity, activity, reliability, and tweet quality. Afterwards, it uses linear combinations of these dimensions to assign influence score to each user. Experimental results on a real-life dataset containing 233,018 Arabic tweets show that E-Rank successfully ranks 8 out of 10 entrepreneurial influencers. Unlike other existing approaches, E-Rank doesn't require any labelled data and has lower computational cost. To ensure the effectiveness and efficiency of E-Rank, three validation techniques were used (1) to compare the detected influencers with the real-world influencers, (2) to investigate the spread of information of the detected influencers, and (3) to compare the quality of E-Rank results with other ranking methods.

1 Introduction

An entrepreneurial ecosystem or entrepreneurship ecosystem refers to the economic environment that influences the regional entrepreneurship [1]. Ecosystem includes commercial chambers, business incubators and accelerators, entrepreneurs, small and medium enterprises (SMEs), and universities. Entrepreneurs might have the capability to innovate and operate, but they require a human and social capital to disseminate their products and work. They need to collaborate with the other stakeholders in entrepreneurial ecosystem to support them.

In the last few decades, social media have taken the world of entrepreneurship by storm. Social media constitute virtual communities, that allow users to sign up for a public profile and establish a network of relationships with people of same interests [2]. These new technologies have a significant impact on how entrepreneurs operate and how they interact with

* Corresponding author.

**The ICCMIT20 program committee effort is highly acknowledged for reviewing this paper.

Email addresses: balmetere0002@stu.kau.edu.sa, maaabdullah@kau.edu.sa, rabbasi@qau.edu.pk

ISSN: 2008-2045

each other [3]. They provide new ways of business-to-business (B2B) communication, information sharing and thus link companies to the different players in the ecosystem [1, 3, 4].

It is noteworthy; however, that social media have more to offer to entrepreneurs, than mere communication and networking. Their impact on news, politics, economy and marketing should not be underestimated [5]. For instance, social media have modernized business management and strategic thinking, and they have introduced a new form of B2B and business-to-ecosystem communications [6]. It is for this reason that social media are being hailed as great assets for individual entrepreneurs who are wary of entering the market [7]. They allow the entrepreneurs to diversify their communication tactics, claim new customers and manage crises [5]. In today's competitive and complex business world, entrepreneurs need to be constantly present on social media to interact with their customers and communicate with different stakeholders [8].

One way to understand social media is through Social Network Analysis (SNA). SNA is defined as a social formation comprising a subset of actors and the interaction between these actors. SNA is a multidisciplinary field of study that partakes of sociology, social psychology, graph theory and statistics. A social network functions at several levels, from one-to-one personal relationships to international relations. Social networks play a pivotal role in problem-solving, corporate management, and personal motivation. SNA involves several tasks, such as identifying important people and finding communities in a network. Such tasks are essential to extract knowledge from networks and to solve problems. [9].

An important aspect of SNA is to study how information diffuses in a network. Information diffusion process - alternatively called information propagation, information spread, or information dissemination - refers to the way information flows and moves between individual and communities within the same social network [10, 11]. Researchers have developed several models to understand the diffusion processes, these models involve discovering the key players in information diffusion [12]. Messages from key persons in the network (R 2012), such as leaders and managers, are more likely to be followed and shared by followers, and would thus reach the whole community via small world [13] and word-of-mouth [14] effects. Understanding the pattern of the information flow and finding some influential users with high levels of interaction and connection in social media and use them to initiate the information spread can diffuse the information more quickly [15]. Twitter and other social media platforms produce valuable opportunities to localize and connect with key nodes of entrepreneurship. The primary research works in entrepreneurial and social media have focused on determining the different actors in the ecosystems and on demonstrating the uses of social media in the ecosystem [1, 3, 7, 8]. This paper is based on two-step flow model of communication (TFMC) [16] when influencers first select data from various media platforms and then relay it to the public. There are many algorithms to detect influencers in social media networks. These algorithms are varied. Some of these algorithms use simple Twitter metrics, while others depend on complex models. Many algorithms are based on the PageRank based techniques, while others consider the contents of the messages, the timeline of tweets, or concentrate on specified topics. This paper proposes an E-Rank framework which linearly combines various metrics to detect influencers. The framework produces a ranked list of influencers, with highly influential users on top of the list. The main contributions of this paper are as follows:

- Collecting entrepreneurial ecosystem dataset from Twitter
- Validating the research needs by proving the influential pattern of the data flow
- Proposing the E-Rank framework to detect entrepreneurial influencers by assigning scores to each user in the dataset
- Identifying the features that capture the entrepreneurial influencers
- Identifying the effective normalizing method in case of detecting entrepreneurial influencers

The rest of the paper is structured in the following way. Section 2 includes a literature review. Section 3 explains the modules of the proposed framework. Section 4 provides the experimental setup, while the evaluation of the framework is presented in Section 5. Section 6 provides a brief discussion of the E-Rank results. Section 7 is the general conclusion of the study.

2 Literature Review

This section defines influential users and the importance of identifying them on social media and reviews the proposed algorithms in previous literatures to identify and predict influential users on social media.

2.1 Identifying Influential Users on Social Media Network

It is assumed that messages from important people are more likely to be received and spread by their followers (R 2012). This way, these messages reach a wide audience due to the word-of-mouth [14] and small-world [13] phenomena. It is essential to comprehend the patterns of the information flow and to find influential users with high levels of interaction and

topological connections in social media networks and use them to increase the diffusion of information more quickly [15].
On Twitter, one recurrent issue is how to identify influential users. This issue is of the utmost importance when we consider the high number of users who choose to be inactive or provide no additional data [17]. Moreover, user identification criteria are numerous and so are the techniques used to classify them.

2.2 Who Is an Influential User?

The concept of user influence is a fuzzy one. Despite the amount of research done so far, there seems to be no exact definition of the concept. Accordingly, new measures are being devised every day to come to an understanding of the issue. Influential users are also called authoritative actors [18], prestigious [19], opinion leaders, or innovators [20]. They have also been linked with topical experts for specific domains [21, 22].
There exist other classifications of users based on the influence spread. For instance, a distinction is usually made between opinion leaders, influential users and discussers as per impact and activity [23]. Another distinction can be made between inventors (those users who launch a new topic) and spreaders (users whose job is topic dissemination) [24]. A further way to classify users focuses on disseminators (those users that spread their influence and presents structural holes), engagers (users whose task is to manage and simplify relations with third parties) and leaders (top disseminator-engagers) [25].
Another important group of users is made up of celebrities [26]. The criteria to classify celebrities differ from those used with influencers. Twitter users are further grouped as per their popularity as either passive users or broadcasters [27] (several followers and few followees) and acquaintances (same amount of followers and followees). Some experts make the difference between popular, influential, listener, star and highly-read users, based not only the accessible metrics but also the content of their tweets, as well as their cognitive and psychological traits [28]. According to [29], influential users are usually linked with hub nodes, but the influence can be travel through multi-levelled peripheral node clusters.

2.3 Influencer Identification Algorithms

A review of the identification algorithms helps to differentiate between three types: (i) network-based [30–41]; (ii) machine-learning based [20, 21, 42]; and (iii) linear combinations [43–45]. The first one relies on network topology and the dynamics models to identify leaders, while the second focuses on finding relevant features to determine the target nodes, the third used linear companions of features.

2.3.1 Network-based Algorithms

Cha *et al.* [33] designed the Indegree measure method (the user's number of followers) and two user functions: retweeting and mentioning. They classified users correspondingly by, mention count, retweet count, and Indegree count. The conclusions of this study indicate that Indegree measure is not telling about the user's influence and that the importance of a user is better seen in the sum of his retweets and mentions. This is in contrast to the findings of another study established a link between how many followers a user has and their centrality [34]. To overcome the degree centrality obstacles, Li *et al.* [41] proposed a novel centrality termed clustered local-degree (CLD), which combines the sum of the degrees of the neighboring of specific nodes and its clustering coefficients to rank spreaders. To overcome PageRank deficiences, Alp *et al.* [40] advanced a Personalized PageRank algorithm using the score of spread and various basic methods. They employ the actions of the user and specific topics to discover current influencers. Zhuang *et al.* [41] have devised a method named SIRank, which calculates the influence diffusion of users in social media by considering user features, such as retweet time intervals and position of users in information cascades to detect influential spreaders in random walk similar to PageRank's original concept. Jinyoung Kim [34] uses betweenness to identify the patterns of information distribution for key users. They found that the number of influential users' followees and followers are significantly associated with their central status in the hashtag network studied. Arularasan *et al.* [30] developed DKIE model to detect the influential users based on their relationships and the discussed topics. Finally, to identify the topics, DKIE clustered the influential users based on the WordNet ontology and the measurement of N-gram similarity. Sun *et al.* [38] used k-shell decomposition on retweet network to determine if the political domain on Twitter indicates the fame, intentions, and the impact of politicians during so called Malaysia's first social media election. Political parties whose representatives had the highest centrality in network won the presidential elections. Sheikhahmadi *et al.* [36] believe that k-shell alone does not have enough efficiency since it gives the nodes within the same shell the same rank. Another shortcoming of this method is that it offers only one node-ranking indicator. To overcome this limitation, they put forth an advanced technique that uses K-core to identify first super-spreader nodes, with an eye to the degree, and the diversity of the friends of the nodes. Cappelletti and Sastry [31] believe that everything on

Twitter is very fast, So they devised a way to rank influential users in real-time during big events. The rank they devised made use of the notion of information amplification. Tinati *et al.*[39] employ the dynamic communication behavior of users on Twitter with no inclusion of the network topology measure. The method makes use of the influence topology of Edelmam to create a classification model [46]. They applied this influence topology on Twitter dataset to establish a network where the role and influence of users is reflected in the inter-user interactions. Ma *et al.* [35] look for the initial spreader first by localizing the dense group and at the same time selecting the initial spreader from each dense group. This method is validated through the susceptible-infectious (SI) model to demonstrate effectiveness and efficiency.

2.3.2 Machine Learning Algorithms

Machine learning algorithms are used to predict influential users, particularly using supervised machine learning. A robust set of features is required to predict results effectively. Labeled datasets are also required to train a machine learning model. Most of the studies on predicting influential users [20, 21, 42] have devised significant characteristics to ameliorate the general model of prediction. Cossu *et al.* [42] investigate a set of conventional featuressuch as user information, tweet characteristics, stylistic features, topology, and othersto identify influential users on Twitter. They proposed several ML approaches based on SNA and Natural Language Processing to classify Twitter users as influencers or not. This study concludes that conventional features provide insignificant results. The authors also proposed a set of new features with enhanced performance. In another study [21] a number of features are identified to guide an SVM. The characteristics use three ways of aggregation: score-, list-, and SVM-based aggregation. The ACQR framework proposed by Chai *et al.* [20] also uses SVM. They proposed the following discriminatory attributes for identifying influential users: reputation, centrality, activeness, and quality of post.

2.3.3 Linear Combinations of Features

Some studies are based on Twitter API metrics, Yuan *et al.* [45] proposed Weighted Ranking Algorithm (WRA), which two measures the QualityScore and ActivityScore. It considers the tweets on specific topic, the total users' tweets, and the probability of retweeting that tweets. Aleahmad *et al.* [43] proposed OLFinder algorithm. It first extracts the hot topics in a domain, then calculates two scores; a competency score and popularity score. The competency score depends on the hot topics while the popularity score is based on users' in-degree. The influence is calculated based on a linear combination of those two scores. Another study [44] tried to characterized the user influence on Twitter to identify the characteristics that created influence and boundaries in a community. The authors used user data such as number of followers, followees, and lists, as well as tweet data such as the total number of tweets per year. The research concludes that the experts have a large number of followers, and they are present in a large number of lists, they tend to post new tweets and replies to other experts, and they are unlikely to re-tweet.

2.4 Research Gaps

As mentioned in this Section, user influence is measured based on various factors and by using various techniques. The state-of-the-art algorithms include network-based [30–41] and machine learning algorithms [20, 21, 42]. Few previous studies measure the user's influence based on the linear combination of selected features [43–45]. Although this approach is simple and scalable. The difficulties are on determining the features which used in the linear co for a given user ranking problem. For example, the features used to detect academic influencers not be relevant when trying to rank the political influencers [47]. In fact, discovering such features to rank the entrepreneurial influencers using a linear combination are not widely explored. Therefore, the linear combination approach was chosen for this research purpose.

3 The Proposed E-Rank Framework

E-Rank is framework aims to detect the entrepreneurial influencers in Arabic Twitter. The strength of the E-Rank framework lies in the following: first, it ranks the entrepreneurship influencers based on a simple metrics. Second, it ranks users such that influencers are ranked higher in the ranking list, this feature is important as users concern top ranked results. Figure 1 shows E-Rank proposed framework. The framework has the following components: (1) data collection, (2) feature selection, (3) data preprocessing, (4) users' annotations and (5) influencers ranking. Data collection component collects data from Twitter to establish a ground truth. The second component is for feature selection. Preprocessing encodes and scales the data. E-Rank labels the users which serves as input for the final component, which is influencers ranking. In the following the details of each component.

3.1 Data Collection

Detecting the entrepreneurial influencers on Twitter requires data about entrepreneurial ecosystem stake-

holders' accounts and their entrepreneurial tweets. As shown in Figure 1, the data collection processes. starts by identifying the accounts entrepreneurial ecosystem stakeholders on Twitter. Based on these accounts, the keywords are determined to start crawling the tweets and to retrieving the users profile data. The next subsections describe the data collection procedures in detail.

3.1.1 Twitter Accounts and Keywords Identification

The entrepreneurial ecosystem stakeholders' accounts were identified and checked manually. The accounts are categorized into six categories based on Andonova *et al.* 2019 [48]. Stakeholders include government sector, universities, startups, entrepreneurs, accelerators and incubators, and unofficial accounts like news and initiatives account. Table 1 shows the description of stakeholders' accounts and the number of identified accounts. There are 658 total accounts. Table 2 shows the keywords, which were used to crawl the Tweets. These keywords contain the most famous entrepreneurial Saudi hashtags and other related keywords. The keywords are identified by determining the top frequent and used words in startups_saudi_forum (الملتى_السعودي_للشركات_الناشئة). The selected hashtag is the most active Saudi entrepreneurial hashtag. Twenty-two keywords were selected.

3.1.2 Tweets Crawling

Tweet crawling is the process of tweets gathering. Search API [1] is used to retrieve tweets that already exist, but it is limited to the last 5,000 tweets for each search query. The dataset was collected using hashtags and keywords identified from the accounts identified in section 3.1.1 during Jan 2, 2018 to Des 31, 2018. As a result, we ended up with a total of 233,018 tweets from 656 users.

This paper requires retrieving the basic user information to characterize the users. Twitter REST API [2] was used to get data of the users. The API returns a user object that has up to 38 attributes including: name, screen name, description, location, number of friends and followers and others. For this research, we manually add a boolean attribute official, which is true, if the a count represents an official account as per the website of the stakeholder. The data was stored in Mongodb [3] in JOSN format. MongoDB is an open source NoSQL database system built for storing semi-structured data.

Table 1. The stakeholder's accounts description

Stakeholder	Description	Total of accounts
Government sectors	The Twitter accounts government sectors related to the business and economic issues in SA such as Saudi commercial chambers, Saudi national banks, general authority for SMEs, the human resources development fund, associations of entrepreneurship, entrepreneurship and the development startups centers, Saudi industrial development fund, local private sector development unit, government initiatives, etc.	76
Universities	The Twitter accounts of Saudi universities incubators, accelerators, centers, agencies, clubs, institute, forums, and initiatives for entrepreneurship	47
Startups	The Twitter accounts of Saudi startups mentioned in Saudi Forum for Startups 2017 [4] reported by Wadi Makkah in Umm Al- Qura university. The startups which do not have a Twitter account were avoided.	223
Entrepreneurs	The Twitter accounts of the founders, co-founders, and CEOs of the Saudi startups, managers and supervisors of entrepreneurship institutes, entrepreneurial counselors and ministers, investors, academia and researchers in entrepreneurship	252
Accelerators & incubators	The Twitter accounts of Saudi profit and non-profit entrepreneurial accelerators and incubators.	43
Unofficial	The Twitter accounts of Saudi entrepreneurial news, clubs, magazines, and unofficial initiatives	17

Table 2. Crawling keywords

Sr	Arabic Keywords	Translation	Sr	Arabic Keywords	Translation
1	الملتقى_السعودي_للشركات_الناشئة	startups_saudi_forum	12	الشركات	companies
2	الشركات_الناشئة	startups	13	الابتكار / الإبتكار	innovation
3	رواد_الأعمال	entrepreneurs	14	مسرعات	accelerators
4	ريادة_أعمال / ريادة	entrepreneurship	15	مسرعة	accelerator
5	مسرعات_الأعمال	bussiness_accelerators	16	استشارات	consulting
6	شركات_سعودية_ناشئة	saudi_startups	17	حاضنة	incubator
7	رياديون / رواد	business pioneers	18	حاضنات	incubators
8	مشروع	project	19	المؤسسات	enterprises
9	مشاريع	projects	20	مؤسسة	enterprise
10	الأعمال	business	21	التمويل	funding
11	الأسر المنتجة	productive families	22	التسليف	credit

3.1.3 Research Validation

An exploratory data analysis (EDA) was conducted to validate the research needs. It proves that entrepreneurial stakeholders rely on specific entrepreneurial accounts to get information. The exploratory data analysis was conducted at two levels: tweet level and user level. The tweet level EDA aims to prove that specific tweets get diffused and get higher interaction than others. The user level EDA proves that there are specific users who influence entrepreneurial stakeholders more than others.

Tweets level EDA: Table 3 shows the statistics of the dataset in respect of retweets, replies, and favorites. Almost all standards of retweets and favorites are close to each other, except for the max value. Also, max value of favorites exceed the retweets by 6.910. This observation is in agreement with [44] that passive members form a major group of the Twitter audience. The passive members engage by liking (Favorite) tweets that they read, supporting influencers

[1] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets - Last accessed, November 2, 2019
[2] https://dev.twitter.com/rest/public - Last accessed, November 2, 2019
[3] https://www.mongodb.com - Last accessed, November 2, 2019

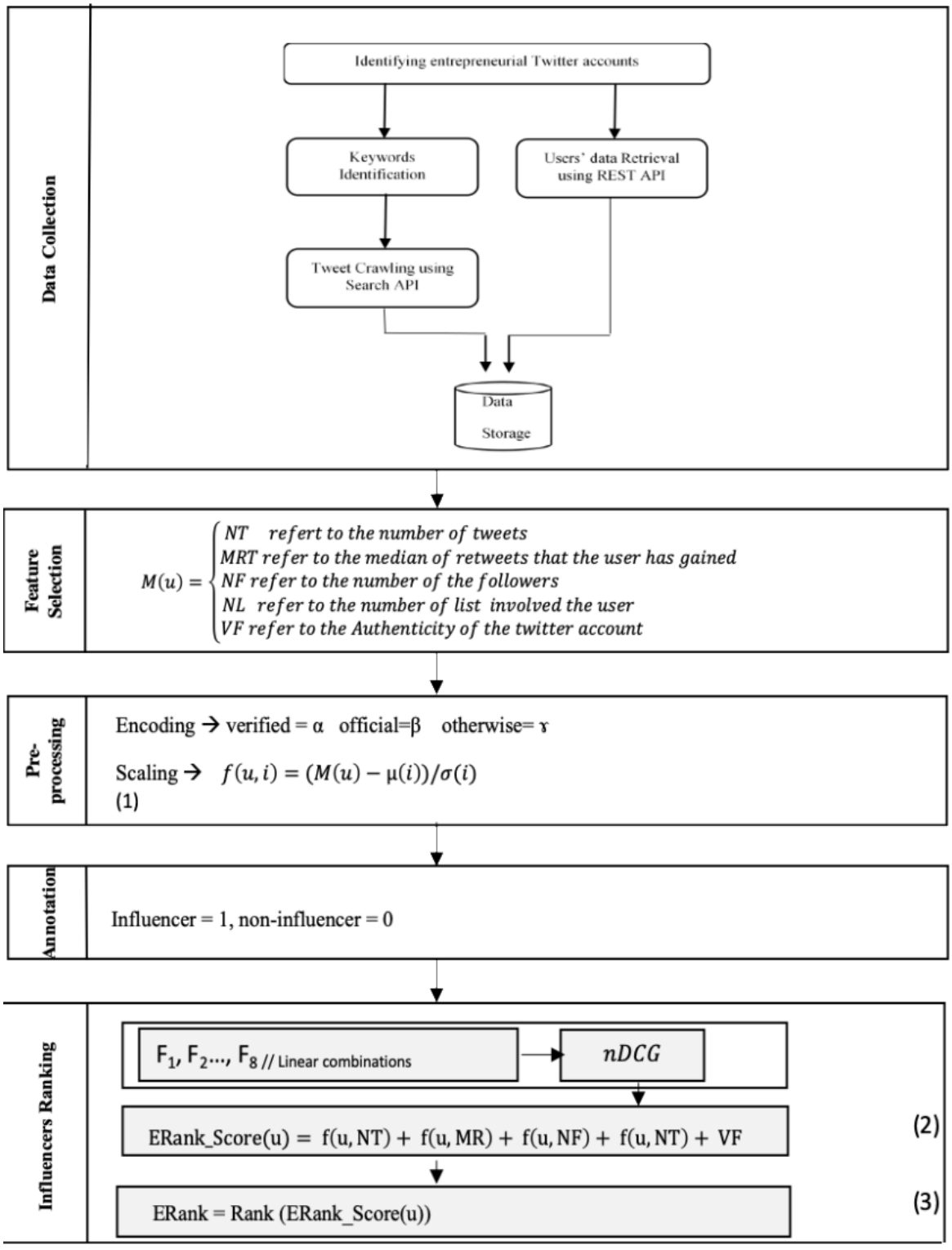

Figure 1. The proposed E-Rank framework

to continue their work. Despite the differences between replies and retweets, they approximately have a close max value; which entails that some tweets ignite replies. Figure 2 shows a log-log plot for each of number of retweets, replies, and favorites, where they follow power-law distributions [49], meaning that the few of tweets gain a large number retweets, replies, and favorites. Most of the distribution arises for tweets with very few retweet, replies, and favorites. We can conclude that there are small number of

Table 3. Statistical standards of the dataset

Statistical	Retweet	Reply	Favorite
Total	1502205	301020	1368362
Mean	6.601	1.323	6.013
Std. Dev.	53.45208	33.16619	56.0431
Max	9490.0	9901.0	16400.0

tweets ignite retweets, repliers and favorites. This

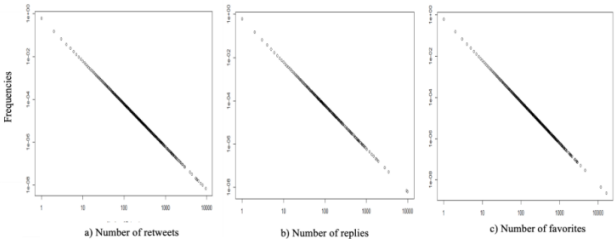

Figure 2. The log-log distribution plot for the number of (a) retweets, (b) replies, and (c) favorites

insight leads to question, does the tweet which ignites the retweets also ignites the other users' reactions (reply and favorites)? To answer this question, a bubble chart in Figure 3 a is used between the count of retweets and favorites which were observed together was plotted. The bubbles represent the number of tweets. The bubble becomes bigger when the cases frequency increases. The most frequent cases are when retweets and favorites in the range of 0 to 50, indicating that most of the tweets do not ignite the other users' reaction or has a few reactions. Figure 3b represents the same concept between retweets and replies. Figure 3 concludes that the tweets which ignite retweet also ignite favorites and replies. Those tweets are few based on the size of the dataset, which consists of 233,018 tweets.

Users level EDA: This section focuses on exploring the data in respect to the users in each stakeholder. We calculated the tweets and retweets for each user, in each stakeholder. Figure 4 shows the histogram distributions of the users' retweets. All the plots are skewed to the right, at the same time; all of them have very few outliers. The outliers indicate that some users got more retweets than other users. Startup has user with very high retweets. There are also users from other stakeholders have high count of retweets. Therefore, the answer of the question is yes, there are users who attract entrepreneurial users more than others.

3.2 Features Selection

As mentioned in section 2.2, the concept of user influence is a fuzzy one, it seems to be no exact definition of the concept. However, the commonly adopted ap-

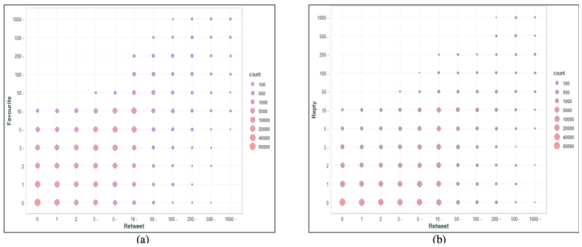

Figure 3. Count of (a) retweets vs. favorites (b) retweets vs. replies

proaches are based in identifying the proper features, i.e., the data characterizing the users. Since this paper focuses on detecting influencers for the information diffusion on a microblog service, we define influencers as the persons who are able to effectively spread information within the network. We use the following four dimensions for measuring influence, each diminution included different features which used to rank the influencers.

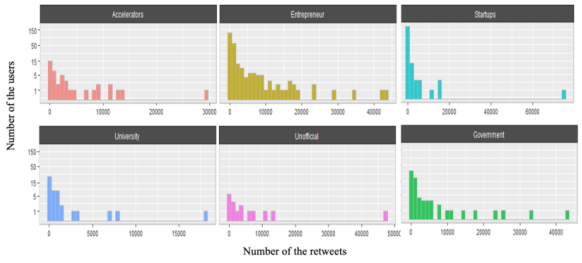

Figure 4. Distributions of the stakeholders' retweets

3.2.1 Popularity Dimension

Popularity dimension has the following two variables:
Number of followers (F): The number of followers accumulated by a user depends on his fame and activity level. The influence of a user can be associated with the amount of their followers because their tweets reach a wide audience [44].
Number of times listed (L): Twitter users can create lists of users. A list usually contains a set of related users. If a user is Listed by many users, it means that users expressly value the person related to the topic of the list, which is an indication of influence [44].

3.2.2 Activity Dimension

Number of tweets (T): User's activity on Twitter can be measured by the number of tweets he posted. Users who post a few tweets are mostly information seekers (Java *et al.* 2007). On another hand, influencers in specific topic tend to be more active and post high number of tweets related to this topic [50].

3.2.3 Tweet Quality Dimension

Median of retweets (MR): The quality of each tweet can be measured in terms of its diffusion and liking in the network, by using characteristics of retweets and favorites [50]. According to Chorley *et al.* [51], retweet is the best quantitative measure when we decide to read a tweet or not.
Median of favorite (MF): Passive members, the largest group, participate by liking (favoriting) tweets that they consume, encouraging others to continue their actions. The large number of retweets and favorites are indications of influence [44, 50]. For this paper, the median of retweet and favorites is used; sum and average make sense in case of normal distribution. Therefore, a better way would be to use the median.

3.2.4 Authenticity Dimension

Verified and official account (VF): Identifying the entrepreneurial influencers differs from identifying the influencers for marketing purpose. The entrepreneurial influencers must be in a place of trust, because of their tweets about crucial issues such as funding, government regulations and others. Therefore, this paper assumes that the users in the entrepreneurial ecosystem will be influenced by official or verified accounts more than others.

3.3 Data Pre-processing

Data preprocessing transforms the raw data for further processing [52]. We use the encoding and scaling for data preprocessing.

3.3.1 Data Encoding

Encoding converts categorical variables to numerical. The verified and official features are encoding into a single variable VF as shown in Equation 1. Thus, if the account is official and verified, it will get $VF = 1$. If the account is official but not verified or vice-versa it will get $VF = 0.5$. The value of VF will be zero, if the account is neither official nor verified:

$$\alpha = \begin{cases} 0, & if\ account\ not\ verified \\ 0.5, & if\ account\ is\ verified \end{cases}$$

$$\beta = \begin{cases} 0, & if\ account\ is\ not\ official \\ 0.5, & if\ account\ is\ official \end{cases}$$

$$VF = \alpha + \beta \tag{1}$$

3.3.2 Data Normalization

Scaling, also called normalization, is the process of transforming the data of different ranges into a uniform scale so that they can be compared [53]. In absence of a consensus on effectiveness of the scaling methods, both Z-score and Min-Max were used to scale the features. The results were compared to determine the effectiveness of each method in ranking entrepreneurial influencers.
Z-score is one of the most popular normalization methods. It is able to handle the outlier issues. it relies on the mean and standard deviation of the target population as shown in the Equation 2, where (μ) is the mean of the population and is (σ) the standard deviation of the population [53].

$$Z_i = \frac{x_i - \mu}{\sigma}. \tag{2}$$

An alternative approach to Z-score standardization is Min-Max. It scales the data to a fixed range, usually [0,1]. Min-max scaling is affected by outliers, as an extreme value can make other values relatively small. Min-Max scaling for range [0,1] is done using the Equation 3. Where x_i is the original value, S_i is the normalized value, S_{min} is the smallest value of the variable (feature), and x_{max} is the largest value [53].

$$S_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}. \tag{3}$$

4 User's Annotation

To ensure the reliability, three expert coders were hired to annotate the top 200 users. Top 200 users were chosen according to the amount of retweets they have gained. The first two coders independently annotated the users as entrepreneurial-influencers or non-influencers. Cohen's kappa was used to measure their agreement between two raters [54]. Cohen's kappa is calculated as shown in Equation 4.

$$k = \frac{P_0 - P_e}{1 - P_e} = 1 - \frac{1 - P_0}{1 - P_e}. \tag{4}$$

Where P_0 = the agreement observed between coders and P_e the probability of chance agreement.
Cohen's Kappa showed a "good" agreement with a kappa value of 0.633, reflecting 85% agreement between the two annotators. Third expert annotated the users independently where the first two coders had disagreed. Based on the three coders' judgment, the dataset contained 28 influencers.

5 Ranking Influencers

Because there is no agreement on the characteristics of ranking the entrepreneurial influencers, influencers characteristics differ from area to another, the features in section 3.2 can be combined to rank entrepreneurial

influencers. Therefore, we assume that:

Assumption 1: T, MR, and VF are considered main characteristics, so they cannot be excluded; T reflects the user's activity; a user cannot be an entrepreneurial influencer if he is not active. MR reflects the information diffusion, which is the main purpose of this research. We assume that the entrepreneurial influencers should be trusted users. Equation 5 reflects the F_1 linear combination of influence:

$$F_1 = T + MR + VF. \tag{5}$$

Assumption 2: F may affect positively in ranking the influencers; the influence of a user can be associated with how many followers they have because of the potential reach of their tweets. Equation 6 reflects the F_2 linear combination of influence:

$$F_2 = T + MR + VF + F. \tag{6}$$

Assumption 3: L affects positively in ranking the influencers, if a user is Listed by many users; it means those many users expressly value the influential user. Equation 7 reflects the F_3 linear combination of influence:

$$F_3 = T + MR + VF + L. \tag{7}$$

Assumption 4: MF affects positively in ranking the influencers; the largest group of passive users participate by favoriting tweets that they consume. Equation 8 reflects F_4 the linear combination of influence:

$$F_4 = T + MR + VF + VL. \tag{8}$$

Assumption 5: NL, NF, and MF can be combined with the main characteristics NT, MR, and VF to increase the ranking performance. Equation 9 - Equation 12 define the linear combinations F_5, F_6, F_7, and F_8:

$$F_5 = T + MR + VF + L + F + MF, \tag{9}$$
$$F_6 = T + MR + VF + F + L, \tag{10}$$
$$F_7 = T + MR + VF + MF + L, \tag{11}$$
$$F_8 = T + MR + VF + MF + F. \tag{12}$$

6 E-Rank Implementation

E-Rank was implemented on a real-life dataset from Arabic Twitter as described in detail in Section 3.1.1 and Section 3.1.2. It contains 658 Twitter accounts. Only 28 accounts were labeled as entrepreneurial influencers as described in detail in Section 3.1.1 and Section 3.1.2. Each linear combination was calculated on the Z- scale and MinMax scaling versions of the features. Top ten users with highest scores were analyzed to measure the effectiveness of the ranking mechanisms. To evaluate the quality, the Normalized Discounted Cumulative Gain ($nDCG$) [55] was calculated for each combination. ($nDCG$) normalized version of (DCG)which measures the quality of list that has been ranked. It assumes that the highly relevant documents that appear down in a result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. In fact, ($nDCG$)s an advantage compared to many other measures, since it has a discount function over the rank, while other popular methods like precision, recall, accuracy, and f-measure consider all the positions. This feature is particularly important for measuring ranking methods as users consider top ranked documents much more than others [55] ($nDCG$) calculated Equation 13, Equation 14 and Equation 15 as

$$nDCG_p = \frac{DCG_p}{IDCG_p}, \tag{13}$$

where

$$DCP_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}, \tag{14}$$

$$IDCP_p = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{log_2(i-1)}. \tag{15}$$

Where REL the list of documents ordered by relevance in the corpus up to position p, and rel_i. is the graded relevance of the result at position i.

7 Experiment Results

Table 4 shows the results of all the linear combinations. In each column, number "1" represent the influencers who is ranked correctly while "0" represents wrong user who was ranked as influencers wrongly. For example, F_1(z_score) ranked the first and second users correctly as influencers, while the third user has been ranked as influencer wrongly. As shown in the Table 4, $F_6 = T + MR + F + L + VF$ with Z-score scaling produces the best results having $nDCG$ value of 0.839. the results also show that z_score normalizing technique provides better result than Min_Max as shown in Figure 5. This result is discussed in detail in Section 8. Figure 6 shows the pseudocode of the E-Rank based on the best combinations. For each user, The T, MR, VF, F, and L are inputted to be normalized and then to measure the influence score of users. Then, the users are ranked based on E-Rank_Scores. Highest E-Rank_Scores refers to highest influence.

Table 4. the result of the all linear combinations

	F1		F2		F3		F4		F5		F6		F7		F8	
	Z-score	Min_Max	Z-score	Min_Max	Z-score	Min_Max	Z-score	Min_Max	Z-score	Min_Max	Z-score	Min_Max	Z-score	Min_Max	Z-score	Min_Max
User 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User 2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User 3	0	0	1	1	1	0	0	0	0	1	1	0	1	1	0	0
User 4	0	1	1	0	1	1	1	0	1	0	0	1	1	0	1	1
User 5	1	1	0	1	0	1	0	1	1	1	1	1	0	1	1	1
User 6	1	1	1	0	1	0	1	1	1	0	1	0	1	0	1	0
User 7	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1	0
User 8	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0
User 9	1	0	0	1	1	1	0	0	1	0	0	0	1	0	0	1
User 10	1	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0
nDCG	0.653	0.548	0.779	0.62	0.708	0.605	0.532	0.586	0.82	0.628	0.839	0.672	0.708	0.623	0.69	0.605

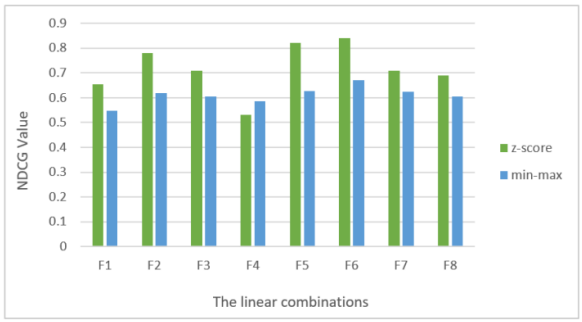

Figure 6. Comparison between Z-score and Min-Max normalization results

Input: T, MR, F, L, VF for each user u $\in$ U

Output: ERank

for each user u $\in$ U **do**

$$M(u) = \begin{cases} T \quad refert\ to\ the\ number\ of\ tweets \\ MRT\ refer\ to\ the\ median\ of\ retweets\ that\ the\ user\ has\ gained \\ F\ refer\ to\ the\ number\ of\ the\ followers \\ L\ \ refer\ to\ the\ number\ of\ list\ \ involved\ the\ user \\ VF\ refer\ to\ the\ Authenticity\ of\ the\ twitter\ account \end{cases}$$

$f(u,i) = (M(u) - \mu(i))/\sigma(i)$ // standardization

$ERank\_Score(u) =$ f(u, T) + f(u, MR) + f(u, F) + f(u, T) + VF

repeats

ERank $\leftarrow$ Rank(ERank_score(u))

return ERank

Figure 5. Pseudocode of the ranking procedure

7.1 E-Rank evaluation

Three methods were used to illustrate the usefulness and correctness of the E-Rank. First, by comparing the ranked influencers with the real-world annotated influencers made by expertise in Section 4. Second, by investigating the spread of information of the ranked influencers. Third, by comparing the E-Rank results quality with other ranking methods.

7.2 Comparing With Real Influencers

Comparing the results of influencers ranking methods with the real-life influencers is one of the common ways used in many literatures [21, 38–40, 44]. The results of F_6 combinations overlaps with eight out of ten Saudi entrepreneurial influencers, at the same time. So, highly influential users appeared on top of the list. The two accounts which are wrongly ranked as influencers are belong to well-known entrepreneurial government accounts, but their influence is not strong.

7.3 The Spread of the Information

This research defines influencers as the persons who can effectively spread information within the network for the information diffusion purpose. Thus, to evaluate the effectiveness of the E-Rank, the spread of information (Retweet) was measured with respect to the user E-Rank score. The number of retweets was plotted against the users' scores as shown by in Figure 7. As seen in the figure. The number of the retweet increase as the user's score increase. This proves that users with higher E-Rank scores can spread information more efficiently in a network.

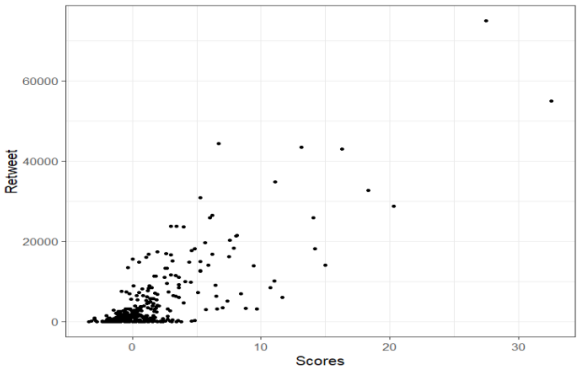

Figure 7. Analysis of retweet count against authors' E-Rank score

7.4 Comparing with State-of-the-art Researches

E-Rank performance was compared with some of a state-of-the-art machine learning classifiers confidence. In learning algorithms, different confidence defines the probability of input to fall in classes [56]. If a class has

high probability, then it has high confidence. SVM, GaussianNB, and logistic regression were applied to classify the users as influencers or none-influencer. All the features discussed in Section 3.2 were considered. To avoid the misleading results, the features were correlated with each other to eliminate any correlated features. We resulted with five features, they are: number of topic tweet, number of the followers, median number of retweets, verified, and official as shown in Figure 8. For each classifier, the users sorted descending according to the classifier's confidence value of class "influencer".

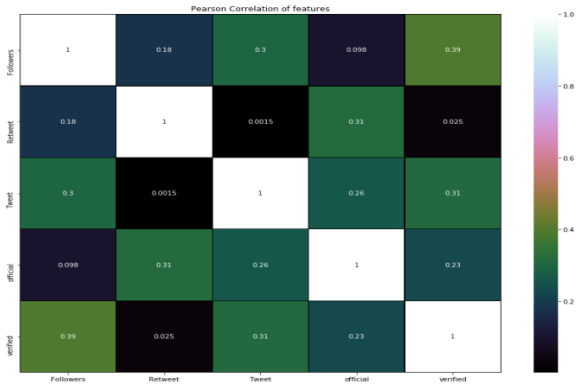

Figure 8. The correlation between various features of Twitter

The linear combinations produced by Asadi and Agah [44]. was also applied to the dataset and compared with E-Rank results. Equation 16 represents the linear combination, where N_F is the follower's number, $\frac{N_{FW}}{N_F}$ is ratio of number of friends to number of followers, N_{listed} is the number of lists that include the user's account, and N_T is the number of topic tweets posted by the user. The constants α, β, γ and η are weighted parameters. f_i is the influence score of influencer i:

$$f_i = \alpha \cdot N_F + \beta \cdot \frac{N_{FW}}{N_F} + \gamma \cdot N_{listed} + \eta \cdot N_T \quad (16)$$

The top ten users off all the previous methods were analyzed, the $nDCG$ has calculated to evaluate the effectiveness of the results. Figure 9 shows the results of the state-of-the-art methods against E-Rank. The E-Rank outperform the other ranking methods where the Z_score normalization works better than Min_Max in all ranking methods. The GaussianNB achieved the better results followed by SVM.

In Fact, the machine learning classifiers did not work well due to two reasons. First, they require robust set of features to work effectively [20, 21, 42], while in this research the features are few and simple, this is the strength of E-Rank, it uses few and simple features while gives high performance. Second, the nature of data is imbalanced, the influencers are only 28 users while the none-influencers are 630.

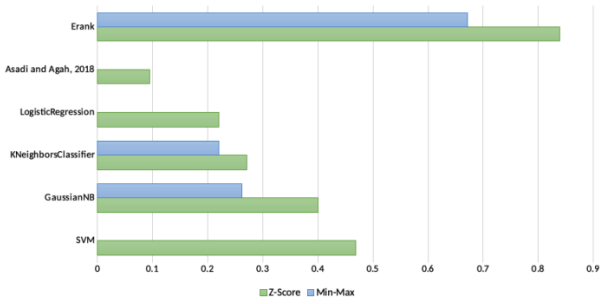

Figure 9. Comparing the performance of E-Rank with SVM, GaussianNB, Logistic Regression, and Asadi and Agah (2018)

8 Discussion

This section discusses the main findings based on the results. Z-score scaling performs better than Min-Max as evident by Figure 5. This is due to the ability of Z-score to handle extreme values better than min-max. In fact, all the features used in this research are likely to have extreme values. Using only the main metrics in F_1, which are number of topic tweets (T), median of the retweet (MR), and verified and official score (VF), resulted in listing verified/official and active accounts on top of the list, but also listed non-entrepreneurial-influencers.

Adding the number of the followers (F) in F_2 to the main metrics came up with famous accounts but also non-influencers, such as some of the government initiatives. On the other hand, it came up with good but not the best results, it overlaps with seven previously identified influencers and ranked the four top influencers correctly, resulted in $nDCG$ equal to 0.779.

Adding the number of the lists (L) in F_3 to the main metrics works as adding the number of the followers in F_2. The results overlap with seven previously identified influencers and rank the top four influencers correctly. The value of $nDCG$ equals to 0.708, which is lower than the value of F_2, because some influencers appear lower in the ranked list.

The results of adding the median of favorites (MF) to the main metrics in F_4 contain only four previously identified influencers. Adding (MF) ranks the famous startups accounts as influencers. By analyzing the ranked accounts qualitatively, we found that tweets from these accounts get a lot of favorites from of non-entrepreneurial users due to the services they provide. In agreement with the previous point, Figure 5 shows that adding (MF) of F_8 to the metrics of F_2 affects negatively, the value of $nDCG$ decreased from 0.779 to 0.69, it ranked the non-influencers and famous startups. At the same time, adding (MF) of F_7 to the metrics of F_3 does not affect the value of $nDCG$. Combining all metrics together in F_5 came up with second best results, it contains eight influencers

with $nDCG$ equal to 0.82. It shows that adding (MF) ranks a non-influencer and a famous startup as influencer, which appears higher in the ranked list.
Finally, F_6, which combines all metrics except MF comes up with the best value of $nDCG$ equal to 0.839, containing eight influencers, at the same time those influencers are ranked higher. This result concludes that T, $MR4$, FV, F, and L are important to rank the entrepreneurship influencers since each of them represents one of the very important characteristics of the influencers. On the other hand, MF affects negatively on ranking results. As mentioned in previous points, most of the favorite actions come from the non-entrepreneurial users and it targets the famous startups such as @AppMrsool.

9 Conclusions

In this research, authors proposed E-Rank framework to rank entrepreneurial influencers on Arabic Twitter. The dataset was extracted from 658 Saudi entrepreneurial Twitter accounts resulting in 233,018 tweets. We extracted metrics from four dimensions to rank the entrepreneurial influencer included user's popularity represented in the number of the followers and lists that include them, user's activity represented in the total of user's tweets related to entrepreneurial issues, tweet quality which measured by the number of retweets and favorites the user has gained, and user's authenticity. Then, different linear combinations of these dimensions are evaluated. The results proved that all the dimensions are important to rank entrepreneurial influencers except for the number of favorites. The proposed E-Rank framework achieved good performance, it successfully ranked 8 out of 10 influencers in top 10 results considering their order in the ranked list.

References

[1] Conor Drummond, Helen McGrath, and Thomas O'Toole. The impact of social media on resource mobilisation in entrepreneurial firms. *Industrial Marketing Management*, 70:68–89, 2018.

[2] Daria J Kuss and Mark D Griffiths. Social networking sites and addiction: Ten lessons learned. *International journal of environmental research and public health*, 14(3):311, 2017.

[3] Eun Kyung Park, Raphael Mateus Martins, Daniel Hain, and Roman Jurowetzki. Entrepreneurial ecosystem for technology start-ups in nairobi: Empirical analysis of twitter networks of start-ups and support organizations. In *DRUID*, 2017.

[4] I Fiegenbaum and O Mohout. The power of twitter: Building an innovation radar using social media. In *Proceedings of the XXVI ISPIM Conference–Shaping the Frontiers of Innovation Management*, pages 1–17, 2015.

[5] Daniel A Gruber, Ryan E Smerek, Melissa C Thomas-Hunt, and Erika H James. The real-time power of twitter: Crisis management and leadership in an age of social media. *Business Horizons*, 58(2):163–172, 2015.

[6] Jan H Kietzmann, Kristopher Hermkens, Ian P McCarthy, and Bruno S Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3): 241–251, 2011.

[7] Eileen Fischer and A Rebecca Reuber. Online entrepreneurial communication: Mitigating uncertainty and increasing differentiation via twitter. *Journal of Business Venturing*, 29(4):565–583, 2014.

[8] Philip Gratell and Carl Johan Dahlin. How does social media affect entrepreneurial leadership: A qualitative study on entrepreneurs perceptions regarding social media as a tool for entrepreneurial leadership, 2018.

[9] Sagar S De, Satchidananda Dehuri, and Gi-Nam Wang. Machine learning for social network analysis: A systematic literature review. *IUP Journal of Information Technology*, 8(4), 2012.

[10] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357 (4):370–379, 2007.

[11] Yu Zhang and Yu Wu. How behaviors spread in dynamic social networks. *Computational and Mathematical Organization Theory*, 18(4):419–444, 2012.

[12] XD Wu, Yi Li, and Lei Li. Influence analysis of online social networks. *Chinese journal of Computers*, 37(4):735–752, 2014.

[13] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[14] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60 (11):2169–2188, 2009.

[15] Mani R Subramani and Balaji Rajagopalan. Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300–307, 2003.

[16] Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications*. Transaction publishers, 1966.

[17] Isabel Anger and Christian Kittl. Measuring in-

fluence on twitter. In *Proceedings of the 11th international conference on knowledge management and knowledge technologies*, pages 1–4, 2011.

[18] Mohamed Bouguessa and Lotfi Ben Romdhane. Identifying authorities in online communities. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):1–23, 2015.

[19] Daniel Gayo-Avello. Nepotistic relationships in twitter and their impact on rank prestige algorithms. *Information Processing & Management*, 49(6):1250–1280, 2013.

[20] Wen Chai, Wei Xu, Meiyun Zuo, and Xiaowei Wen. Acqr: A novel framework to identify and predict influential users in micro-blogging. In *Pacis*, page 20, 2013.

[21] Nian Liu, Lin Li, Guandong Xu, and Zhenglu Yang. Identifying domain-dependent influential microblog users: A post-feature based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[22] Shao Xianlei, Zhang Chunhong, and Ji Yang. Finding domain experts in microblogs. *ser. WEBIST*, 14, 2014.

[23] Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks. In *International Symposium on String Processing and Information Retrieval*, pages 111–117. Springer, 2012.

[24] Jingxuan Li, Wei Peng, Tao Li, Tong Sun, Qianmu Li, and Jian Xu. Social network user influence sense-making and dynamics prediction. *Expert Systems with Applications*, 41(11):5115–5124, 2014.

[25] Miguel del Fresno Garcia, Alan J Daly, and Sagrario Segado Sanchez-Cabezudo. Identifying the new influences in the internet era: Social media and social network analysis. *Revista Española de Investigaciones Sociológicas*, (153), 2016.

[26] MS Srinivasan, Srinath Srinivasa, and Sunil Thulasidasan. Exploring celebrity dynamics on twitter. In *Proceedings of the 5th IBM Collaborative Academia Research Exchange Workshop*, pages 1–4, 2013.

[27] Gerasimos Razis and Ioannis Anagnostopoulos. Influencetracker: Rating the impact of a twitter account. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 184–195. Springer, 2014.

[28] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. In the mood for being influential on twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 307–314. IEEE, 2011.

[29] Stefan Räbiger and Myra Spiliopoulou. A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications*, 42(5):2824–2834, 2015.

[30] AN Arularasan, Annamalai Suresh, and Koteeswaran Seerangan. Identification and classification of best spreader in the domain of interest over the social networks. *Cluster Computing*, 22 (2):4035–4045, 2019.

[31] Rafael Cappelletti and Nishanth Sastry. Iarank: Ranking users on twitter in near real-time, based on their information amplification potential. In *2012 International Conference on Social Informatics*, pages 70–77. IEEE, 2012.

[32] Manuel Castriotta, Michela Loi, Elona Marku, and Luca Naitana. Whats in a name? exploring the conceptual structure of emerging organizations. *Scientometrics*, 118(2):407–437, 2019.

[33] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

[34] Jinyoung Kim. How did the information flow in the# alphago hashtag network? a social network analysis of the large-scale information network on twitter. *Cyberpsychology, Behavior, and Social Networking*, 20(12):746–752, 2017.

[35] Songjun Ma, Ge Chen, Luoyi Fu, Weijie Wu, Xiaohua Tian, Jun Zhao, and Xinbing Wang. Seeking powerful information initial spreaders in online social networks: a dense group perspective. *Wireless Networks*, 24(8):2973–2991, 2018.

[36] Amir Sheikhahmadi and Mohammad Ali Nematbakhsh. Identification of multi-spreader users in social networks for viral marketing. *Journal of Information Science*, 43(3):412–423, 2017.

[37] Igors Skute. Opening the black box of academic entrepreneurship: a bibliometric analysis. *Scientometrics*, 120(1):237–265, 2019.

[38] Hong-liang Sun, Eugene Chng, and Simon See. Influential spreaders in the political twitter sphere of the 2013 malaysian general election. *Industrial Management & Data Systems*, 2019.

[39] Ramine Tinati, Leslie Carr, Wendy Hall, and Jonny Bentwood. Identifying communicator roles in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1161–1168, 2012.

[40] Zeynep Zengin Alp and Şule Gündüz Öğüdücü. Identifying topical influencers on twitter based on user behavior and network topology. *Knowledge-Based Systems*, 141:211–221, 2018.

[41] Kechen Zhuang, Haibo Shen, and Hong Zhang. User spread influence measurement in microblog. *Multimedia Tools and Applications*, 76(3):3169–3185, 2017.

[42] Jean-Valère Cossu, Nicolas Dugué, and Vincent Labatut. Detecting real-world influence through twitter. In *2015 Second European Network Intelligence Conference*, pages 83–90. IEEE, 2015.

[43] Abolfazl Aleahmad, Payam Karisani, Maseud Rahgozar, and Farhad Oroumchian. Olfinder: Finding opinion leaders in online social networks. *Journal of Information Science*, 42(5):659–674, 2016.

[44] Mehran Asadi and Afrand Agah. Characterizing user influence within twitter. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 122–132. Springer, 2017.

[45] Jinfeng Yuan, Li Li, Le Luo, and Min Huang. Topology-based algorithm for users' influence on specific topics in micro-blog. *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE*, 10(8):2247–2259, 2013.

[46] Jonny Bentwood. Distributed influence: Quantifying the impact of social media. *Hentet den*, 15, 2008.

[47] Jean-Valère Cossu, Vincent Labatut, and Nicolas Dugué. A review of features for the discrimination of twitter users: Application to the prediction of offline influence. *Social Network Analysis and Mining*, 6(1):25, 2016.

[48] Veneta Andonova, Milena S Nikolova, and Dilyan Dimitrov. What is an entrepreneurial ecosystem? *Entrepreneurial Ecosystems in Unexpected Places*, pages 3–16, 2019.

[49] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46 (5):323–351, 2005.

[50] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on twitter: A survey. *Information processing & management*, 52(5):949–975, 2016.

[51] Martin J Chorley, Gualtiero B Colombo, Stuart M Allen, and Roger M Whitaker. Human content filtering in twitter: The influence of metadata. *International Journal of Human-Computer Studies*, 74:32–40, 2015.

[52] A Famili, Wei-Min Shen, Richard Weber, and Evangelos Simoudis. Data preprocessing and intelligent data analysis. *Intelligent data analysis*, 1(1):3–23, 1997.

[53] S Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[54] Robert Gilmore Pontius Jr and Marco Millones. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429, 2011.

[55] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6, 2013.

[56] Scikitlearn. `https://scikit-learn.org/stable/modules-/generated/sklearn.svm.libsvm.predict_proba.html`, Accessed 22 December 2019.

Bodor Almotairy is Ph.D. student. She has completed her master's degree in Information System Department at the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia in 2020. She obtained her bachelor's degree with first honor from King Abdulaziz University. Her research field's interest includes data science and social network analysis.

Manal Abdullah received her Ph.D. in computers and systems engineering, Faculty of Engineering, Ainshams University, Egypt, 2002. She has experienced in industrial computer networks and embedded systems. Her research interests include artificial intelligence, performance evaluation, WSN, network management, big data analysis, and streaming data analysis. Currently, she is professor in Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia.

Rabeeh Abbasi completed his Ph.D. from University of Koblenz-Landau, Germany in 2010. He is working as an associate professor at the Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan. He has a vast research experience in the fields of social media analytics and social network analysis. His research focuses on leveraging positive aspects of social media including social media's use in saving lives, understanding events, and analyzing sentiments among many others.