ECharacterize: A Novel Feature Selection-Based Framework for Characterizing Entrepreneurial Influencers in Arabic Twitter

https://doi.org/10.3991/ijim.v14i10.14807

Bodor Moheel Almotairy(✉), Manal Abdullah
King Abdulaziz University, Jeddah, Saudi Arabia
balmetere0002@stu.kau.edu.sa

Rabeeh Abbasi
Quaid-i-Azam University, Islamabad, Pakistan

Abstract—Social media are widely used as communication platforms in the world of business. Twitter, in particular, offers valuable opportunities for collaboration due to its open nature. For that, many entrepreneurs employ Twitter for different reasons, such as mobilizing financial resources, get funding, and increase their innovation capabilities. Therefore, they keep looking for local entrepreneurial accounts to help them. Messages from entrepreneurial influencers -opinion leader- increase the information diffusion to entrepreneurs, helping them to find more opportunities. Discovering the characteristics of entrepreneurial influencers in Twitter networks becomes extremely important since it reflects the way to reach entrepreneurs. In the present paper, we propose a novel framework called ECharacterize based on feature selections techniques to discover the characteristics of the entrepreneurial influencer in the Saudi context in a robust manner. The framework extracts abundant influencers' features and then employs seven state-of-the-art ranking methods to determine the characteristics of the most relevant influencer. It robustly aggregates the lists to come out with the accurate final list using Robust Rank Aggregation. The framework examined on 233,018 real-life Arabic tweets. The results show the ability of the proposed method to distinguish between the influencers by their popularity, reliability and activity level.

Keywords—Twitter, characteristics of influencers, entrepreneurial influencers, robust ranked list.

1 Introduction

In the last decade, a variety of social media platforms have brought the new world of information. Starting by Myspace, which disappears with Facebook and Twitter. Then, a life-sharing social network such as Instagram, Snapchat [1] and others give us the capability to share our voice, make some new friends, become more social positively, and giving a chance to share some cultural information and build an extensive knowledge the unknown places and cultures. Overall, these social networks allow people to be connected whenever and everywhere.

Nowadays, many entrepreneurs employ social media – Twitter in particular – for different reasons, such as mobilizing financial resources [2], connecting with potential investors in an attempt to get funding [3], connecting with other startups [3]. Another use of social media lies in consulting with advisors for knowledge creation [4], the process of innovation [5] and innovation capabilities [6], which allows them to find more opportunities. Entrepreneurs in Twitter look for information from various sources, especially the local accounts, helping them to find and interact with other stakeholders in the entrepreneurial ecosystem. Entrepreneurs in their early-stage look for information from various sources, especially the local accounts. Since they need the advice and consulting provided by entrepreneurial ecosystem stakeholders [5], therefore, there is a need to enhance the information follow for them [5].

Messages from key persons in the network [7], 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 [8] and word-of-mouth [9] effects. There are many studies that focus on ranking Twitter influencers [10]. However, how can we influence these networks? With the help of some particular accounts known by influencers particular which allow Twitter, for example, to have the chance to interact and increase information diffusion with accounts followers (audiences) become efficient.

All these reasons drive us to think about an efficient manner to detect influencers. Thus, it is difficult to determine the appropriate features for a given study case. For example, some features are used to detect academic influencers not be relevant if it is used to rank the political influencers. [10]. For this reason, there is a need to determine the appropriate features of an entrepreneurial influencer.

In the literature, many methods have been proposed and developed to determine the most pertinent attributes for particular applications. Ranking methods is a technique for attribute selection used to emphasize the most relevant attributes. Also, it represents a critical task for information retrievals, such as search engines, advertisement systems, and recommender systems [11]. Rank aggregation (RA) can be defined as a process that combines multiple ranked lists and gives as output one accurate ranking list [12]. Because of the RA, it becomes easy to integrate information from individual genomic studies [12].

This article is motivated by the lack of literature to identify the characteristics of Twitter's entrepreneurial influencers, particularly for users of Saudi Arabic. It reviews the Twitter influencers' characteristics in detail to establish the link between these characteristics and the Saudi entrepreneurial influencers. Then it proposes a novel framework called ECharacterize to determine the essential characteristics, robustly. The ECharacterize framework extracts abundant influencers feature a range of characteristics against different research fields such as natural language processing, retrieval information and social network understanding. It is built on eight state-of-the-art ranking and aggregation methods to ensure its efficiency. The ECharacterize was examined on a real-life data set, reaching a total of 233,018 Arabic tweets from 656 Saudi entrepreneurial ecosystem stakeholders. Finally, the results were evaluated by three different state-of-the-art algorithms for machine learning, supervised prediction models to prove its efficiency and correctly.

The rest of this paper is organized as follows: the theoretical literature review is discussed in section 2. In section 3, we explain the framework phases of its evaluation. Section 4 discusses the obtained results and interpret the phenomena. Finally, a conclusion and perspective work are presented in section 5.

2 Literature Review

This section reviews many interesting features could be used to characterize Twitter users in subsection 2.1. All the discussed features are extracted to characterize entrepreneurial influencers in the ECharacterize framework. Then, the ranking methods which embedded in the ECharacterize framework are explained in section 2.2, followed by explanation of the aggregation method in 2.3.

2.1 Features

The features are grouped into five categories. In Fact, the categorization of features does not follow any standard. So, authors usually tend to categorize them thematically. The next subsections describe these features in detailed based in their group.

User profile: The first group gathers features related to user profiles. Feature 1(Verified) indicates if the users' account verified by Twitter [13]. Feature 2 (Description length) is the number of characters written by the user to describe himself. In fact, this feature is considered an excellent feature to indicate the user presence on Twitter and his online presence. Generally, corporate accounts and professional bloggers tend to fill their profile[14]. Feature 3,4, and 5 (URLs, usernames (mentions), and hashtags) are appearing in the textual profile description. Previous studies [14] [15] show that some users use these features to indicate their professional, distinguished roles to gain visibility in a specific area. Feature 6 (Profile age) could be related to the user's visibility on Twitter since it needs some time to have an influential position [14].

Activities and publications: Publishing activity category focuses on the ways the influencers behaves regarding publishing the tweets. Feature 7 (Tweet count) represents the total number of tweets he posted in general, while Feature 8 (Topic Tweet) corresponds to the number of tweets he posted related to the entrepreneurial issues. Tweet count and topic tweet represent the user activity on Twitter [14] [15].

Interaction and responsiveness: This category focus feature describes how the user interacts with people. Feature 9, 10, and 11 are related to the reactions caused by the user's tweets. These features can be used as indicators to the tweet quality, and the high-quality tweet may cause a tremendous other reaction. Feature 9 (Retweet) represents the total number of retweets of the user's tweets [15]. Feature 10 (Favorite) is the number that the user's tweet marked as a favorites. [15]. Feature 11 (Reply) represents how many times the user's tweet replies by others [15]. Feature 12 (User Favorites Count) represents another type of interactions, and it considers the total number of favorites chosen by the user[15].

User relationship: Relationship category describes how the user is popular and famous on Twitter and connect to the rest of the Twitter users. Feature 13 and 14 clarify how much others prefer the user' tweets on Twitter. Feature 13 (Follower) is the number of user's followers[16], while Feature 14 (List) is the count of lists include the user's account [16]. On the other hand, feature 15 (Friends) correspond to how much the user seeks information from others [16].

Lexical Aspects: The features of this category can be investigated in order to figure out the lexical aspects. These features are beneficial to distinguish users based on the ways users describe themselves on Twitter. For instance, if users belong to the same class used to describe themselves in the same way, the selected features will be useful and allow their identification. Features 16 focuses on the Parts of Speech (POS), while feature 17 focus on Named Entities recognition NER. POS and NER are Natural Language Process (NLP) techniques [17]. NLP is a field of linguistics in computer science with artificial intelligence that concerns with the interactions and defines the languages used by human in a comprehensive way to the computers[17]. The Part-of-Speech (POS) tagger is a process of tagging a sentence to a list of words. In general, eight main parts define the in which: adjectives, interjections, prepositions, nouns, adverbs, verbs, conjunctions and pronouns as cited in [17]. The profile may include more than one part. The output of this stage is tagged profiles (T.P) as shown by equation 1.

$$T.P = \{V1...n, N1...n, Adv1...n,..., Adj1...n\} \tag{1}$$

The Named Entity Recognition (NER) aims to classify named entities mentioned in a specific text into some predefined categories for example "cities", "companies", "organization", "individuals", "product " and others. The NER gives a wealth knowledge and meaning to the given text to be understandable. Thus, these feasters can be used to discover the relation between the named entities mentioned in the profiles and the users' influence. The output of this NER is Named Entities in profiles (N.E.P) shown by equation 2.

$$N.E.P = \{P1...n, O1...n, L1...n\} \tag{2}$$

2.2 Ranking Methods

The ranking is one of the significant problems in the field of information retrieval, which aims to assign a score to a set of objects (for example documents), this rank will be used to sort these objects. For the feature, ranking is used to give a score to each feature in order to figure out the most relevant one for a specific study. Depending on its application, the ranking may give an idea about the relevance, importance of the studied case[11]. In the literature, several methods for features ranking have been proposed [11]. Based on state of the art, SVM-RFE, Correlation, Information gain, Chi-squared, Gain ratio, and Random forest are chosen in order to rank the entrepreneurial features. We describe in the next subsections these methods briefly.

Random Forest: In data science, Random Forests RF are considered accessible, accurate, robust, and easy to use machine learning methods. RF proves its effectiveness in assess features importance. RF uses decision tree strategies which rank features according to its contribution in improving the node purity, decreasing impurity over all trees. Also, they provide a helpful feature called feature importance. The feature importance finds the most effective variable in the dataset [18].

In the decision trees, every node is considered a feature condition to divide the dataset into two sets: training and test. So, during training a tree, they compute how much each feature decrease the impurity. This could help us in the classification stage because it is based on both information gain/entropy. For regression trees, it is known by variance. Finally, the feature list is ordered based on this measure[18].

SVM-Recursive feature elimination: Support Vector Machines Recursive Feature Elimination (SVM-RFE) is a well-known approach for ranking. As mentioned in the study of Guyon et al. [19], this approach has shown superior classification results compared with other methods. Generally, this method is used to evaluate the importance of each variable. SVM-RFE can also find the best combination possible for the feature in order to have the best classification performance [20]. Moreover, this method uses a recursive way to classify some samples from the dataset with SVM then selects the best fit and ensure the tradeoff between accuracy and feature number. [19].

Information Gain: Information Gain, on the other hand, is one of the ranking methods that give a weight for the feature by measuring the gain vis-a-vis the class. It performs the feature selection based on Claude Shannon theory[21], based on the information value for the analyzed message. The formula can be expressed as follow:

$$IG = H(Y) - H\left(\frac{Y}{X}\right) \tag{3}$$

Where H (Y|X) is the uncertainty about Y for a given X and H(Y) is the entropy of Y. IG is a symmetrical measure, where the information gained with Y to X is the same as with X to Y. IG biases to high branching features even if it is not valuable for the study. Because of this bias, it is recommended to select a large number value for the attributes before performing the IG method.

Gain Ratio: The gain ratio is an extension of IG with less bias since it take into consideration the size and number of branches when choosing a feature[22]. This is done by normalizing the IG by "intrinsic information" of a split. intrinsic information is a positional information created by splitting the dataset into n portions. Gain Ratio is given by equation 4

$$GainRatio(A) = \frac{Gain(A)}{Splitinfo(A)} \tag{4}$$

Where $Splitinfo(A)$ is intrinsic information. GR biases to unbalanced splits in which one partition is smaller than the other.

Symmetrical Uncertainty: The symmetrical uncertainty SU criterion, giving by equation 5, is explained in order to compensates the inherent bias of IG [22].

$$SU = 2\left(\frac{IG}{H(Y)+H(X)}\right) \tag{5}$$

The values of SU are selected and normalized to [0,1] range. If SU value is 1, that means this feature can be predicted successfully, else its value is 0, there is no correlation between X and Y. This method is pretty similar to GR in the bias because its selection is based on the features with lower values.

Correlation: The selection of characteristics based on correlation is the basis of symmetric uncertainty (SU) [23]. It is a symmetrical measure that can be used to measure the correlation between characteristics and characteristics. The value of symmetrical uncertainty ranges [0 to 1]. Thus, one indicates that one variable (either X or Y) ultimately predicts the other variable. The value of 0 indicates that both variables are entirely independent. The Pearson correlation coefficient is defined as the following equation 6 to predict Y.

$$R(i) = \frac{cov\ (X_i,Y)}{\sqrt{var\ (X_i),var(Y)}} \tag{6}$$

where cov and var designate, respectively, the covariance and the variance.

Chi-squared: Chi-square is one of the standard methods which is used to select feature [24]. As described in formula 7, this method evaluates feature values by calculating its statistic chi-squared. Starting by a hypothesis H0 which assume that there is no relation between a set of features (two or more) and perform the test by the following formula:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{7}$$

Where Oij is the observed frequency and Eij is the expected (theoretical) frequency, asserted by the null hypothesis. The higher the value of χ2, the greater the evidence against the hypothesis H0 is.

2.3 Ranking aggregation

Ranking aggregation is the process of aggregating many ranked lists generated by individual rankers to one ranked list. This gives a better rank and resort the list based on the new rank[12]. In general, Rank Aggregation (RA) is an ensemble-based method for feature selection. Using this technique gives more accurate results with different kind of data as reported in [12]. Furthermore, the RA method can perform in both supervised and supervised methods, but overall, the unsupervised RA methods are mainly used in the literature [12]. Overall in this field of study, there are many studies to rank features using aggregation, we cite, for example, median, highest ranked, sum, mean, and lowest rank aggregation [25]

Robust Rank Aggregation (RRA) is an aggregation method proposed by Dittman at el. 2013 [25] to aggregate results of many ranking methods in an unbiased manner. RRA is considered one of the statistically stable and computationally efficient algorithms. Authors proposed RRA to prioritize genes lists in genomic data analysis applications. RRA assigns an importance score for each gene, providing a robust way to retain only the relevant genes in the final list. RRA looks at how the feature is positioned in the ranked lists and compares it to the baseline case where all ranked lists are shuffled randomly. Then, RRA assigns a *P*-value for all features to decide their significance and for re-ranking the feature.

3 ECharacterize Framework

This research proposes ECharacterize framework in order to discover the traits which make certain users more influential in entrepreneurial ecosystem on Twitter. The ECharacterize assigns importance scores to each influencers feature; then, the features are evaluated by prediction validation. The feature scores have generated by aggregating the ranked lists created by seven state-of-the-art feature ranking methods. Figure 1 shows the ECharacterize framework components. Next subsections explain the components in detail.

3.1 Data Collection

A real dataset was collected from Twitter. The Twitter Search API[1] was used to crawl the data from Saudi entrepreneurial hashtag “startups_saudi_forum /الملتقى_السعودي_للشركات_الناشئة” during Jan 2, 2018, to Des 31, 2018. Based on the collected tweets, Twitter REST API[2] was used to get data of the users’ profiles. As a result, we ended up with a total of 233,018 tweets from 656 users.

3.2 Features Extraction

All the seventeen discussed features in section 2.1 were extracted. Stakeholder, official, and contact channels are new features that added for this paper purpose. Those features are not discussed before in the literature. Stakeholder feature represents entrepreneurial stakeholder category which Twitter account belongs to. The stakeholders are categorized into six categories based on Andonova et al. 2019 [26]. They include government sector, universities, startups, entrepreneurs, accelerators and incubators, and unofficial accounts like news and initiatives. The official feature represents if the account is official or not. 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 accounts. Contact Channels represents the availability of contact channel in the profile, increasing the profile reliability. We categorized the two new features “official” and contact channels in the profile features. Stakeholders are considered a separated feature.

[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

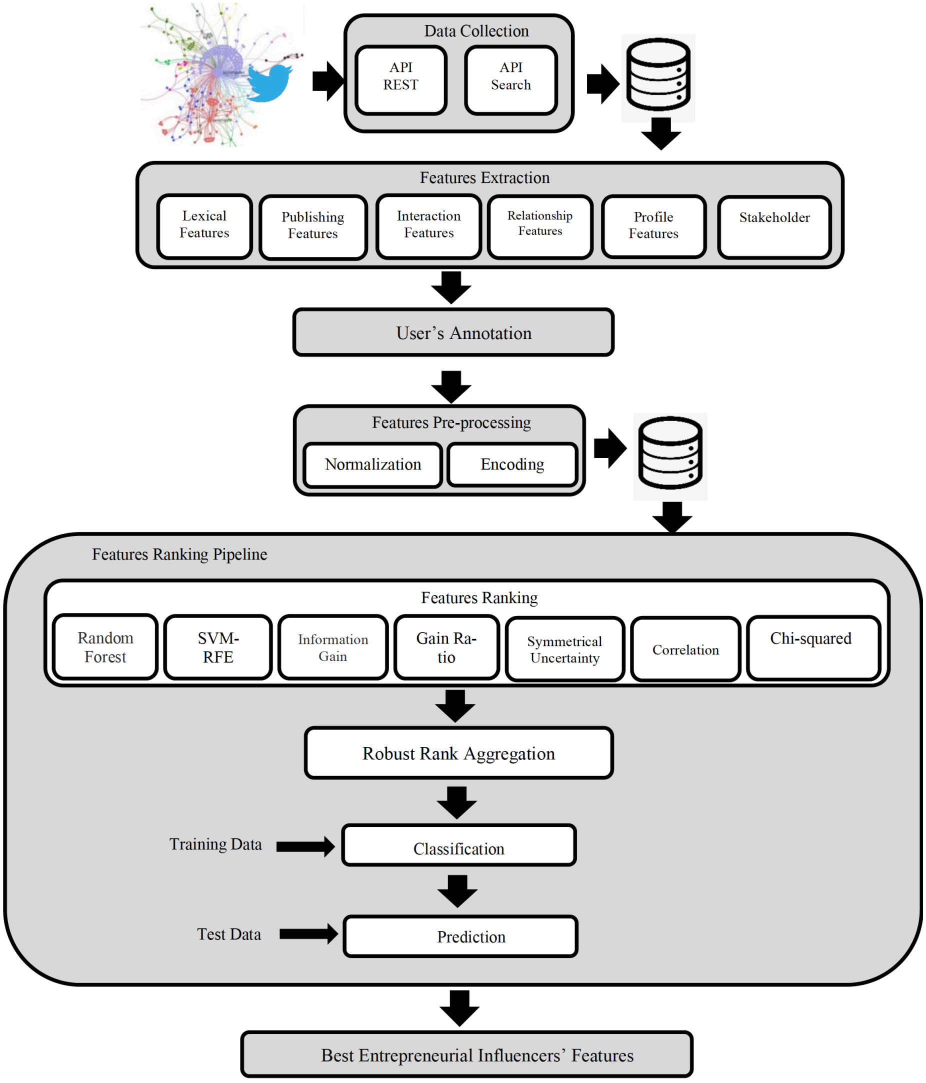

Fig. 1. ECharacterize Framework

MADAMIRA was used to apply NER and POS [27]. MADAMIRA is one of the state-of-the-art Arabic, accurate, and fast text processing morphological analysis for Arabic text. MADAMIRA can find the named entities in three categories they are Person (PER), Organization (ORG), and Location (LOC), we consider each category a separated feature. Regarding POS, MADAMIRA can find all the eight parts of speech, representing eight new features. The number of users tagged, and the number of hashtags were calculated by counting the @ and # symbols in the tweets. Profile description length extraction step, the words and spaces were kept, everything else was removed. Then, the number of letters were counted. This step must be done after extracting the features like hashtags; user tagged because this step will also remove '#' and '@'. Finally, we result in 35 features; they are listed in table 1.

Table 1. The extracted entrepreneurial influencers features

Category	The Features
User's profile	Verified, Official, Profile age, Contact Channels, Description length, and URLs, usernames (mentions), and hashtags appearing in the textual profile description
Publishing activity	Tweet count (all tweet the user-posted) and Topic Tweet (the number of entrepreneurial tweets the user-posted)
User's interaction	Retweet, favourite, reply (the count of other users' reactions on the user's post), and User Favorites Count (considers the total number of favourites selected by the user)
User' relationship	Followers Count, List Count, and Friend Count
Lexical Aspects	Named Entities features (PER, ORG, and LOC), and Part of Speech features (nouns, pronouns, proper nouns, adjectives, verbs, adverbs, prepositions, conjunctions and interjections)
Stakeholder	Each stakeholder category represents a separated feature (government sector, universities, startups, entrepreneurs, accelerators and incubators, and unofficial accounts)

3.3 User's Annotation

To ensure reliability, three expert coders were hired to annotate the top 200 users. Top 200 users were chosen according to the number 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 [28]. Cohen's Kappa showed a 'good' agreement with a kappa value of 0.633, reflecting 85% agreement between the two annotators. The third expert annotated the users independently, where the first two coders had disagreed. Based on the three coders' judgment, the dataset contained 28 influencers.

3.4 Data preprocessing

Data preprocessing transforms the raw data for further processing [29]. Based on the collected dataset, there are no missing data, and we did not remove outliers since they reflect some influencers' characteristics. This research used the encoding and normalization for data preprocessing.

- Normalization: It is the process of transforming the data of different ranges into a uniform scale so that they can be compared [30]**.** Z-score was used to scale the features due to its ability to handle the outliers.
- Encoding is the process of converting categorical variables into numerical. Binary encoding technique was used to encode the verified and official features, '0' represents the account which is not verified or official, while '1' represents verified and official the account. The stakeholder feature was encoded using one-hot technique. One-hot encoding is binary style of categorizing, each categorical variable has one element for each label with the class label is 1 and all other elements are 0.

3.5 Features ranking and aggregation

In this paper, researchers consider seven commonly used features ranking methods based on learning algorithms, statistical and entropy-based with excellent performance in various domains. These are random forest, SVM-RFE, information gain, gain ration, symmetrical uncertainty, correlation, and chi-squared [11]. Robust Rank Aggregation (RRA) algorithm was used to aggregate the seven lists produced by the ranking methods. RRA returns the final aggregated list with associated *P*-value score of each feature. The *P*-value is used for deciding their significance and thus re-ranking the feature. Figure 2 shows the aggregated results and its *P*-value scores. *P*-value score becomes significant (smaller than 0.05) as the features become more important. Table 2 shows the results of all the ranking and aggregation methods. The numbers indicate to the position of the feature in the list, and the final column shows the RRA associated score (*P*-value).

Table 2. The result of all ranked methods and RRA method.

	RF	**SVM-RFE**	**Correlation**	**IG**	**GR**	**SU**	**Chi**	**RRA**	***P*-value**
FollowersCount	4	2	5	5	6	6	4	1	6.98E-06
listedCount	9	3	6	6	2	2	5	2	0.000187
All_Tweet	7	10	7	3	5	5	7	3	0.000427
Favorite	8	1	4	2	4	3	1	7	0.009612
UserFavoritesCount	11	15	15	8	9	8	12	5	0.009416
Reply	3	20	2	4	1	1	2	6	0.009612
Retweet	6	5	1	1	3	4	3	4	0.009612
Tweet	1	11	3	7	7	7	6	8	0.014028
Verified	30	4	8	11	8	11	8	9	0.017151
ProfileAge	12	19	19	12	18	16	13	10	0.054688
Desclength	2	18	23	10	11	10	10	11	0.143959
ORG	18	9	16	20	26	24	9	12	0.545206
Official	15	31	9	15	12	12	27	13	0.601293
FriendsCounts	5	25	22	9	10	9	11	14	0.69346
Verb	10	30	11	14	16	14	21	15	0.754631
Adjective	13	28	12	17	25	21	22	16	0.934387
Noun	14	26	30	13	22	17	20	17	1
Preposition	16	24	29	18	21	18	23	18	1
Mentions_in_profile	17	21	32	21	24	22	15	19	1
Unofficial	19	35	14	30	20	28	31	20	1
Startups	20	6	34	22	19	20	34	21	1
Hashtag_in_profile	21	29	28	23	30	25	16	22	1
LOC	22	14	31	26	31	30	18	23	1
University	23	8	20	34	33	33	35	24	1
Pronoun	24	33	17	19	13	13	25	25	1
ContactChanel	25	13	18	28	29	27	14	26	1
Accelerators	26	27	25	32	32	32	30	27	1

Conjunction	27	23	33	24	23	23	24	28	1
URL_in_Profile	28	17	13	29	28	29	26	29	1
ProperNoune	29	16	35	16	15	15	19	30	1
Entrepreneur	31	12	24	33	34	34	32	31	1
PER	32	27	27	27	27	26	17	32	1
Adverb	33	32	26	31	14	31	28	33	1
Interjection	34	34	21	35	35	35	29	34	1
Government	35	7	10	25	17	19	33	35	1

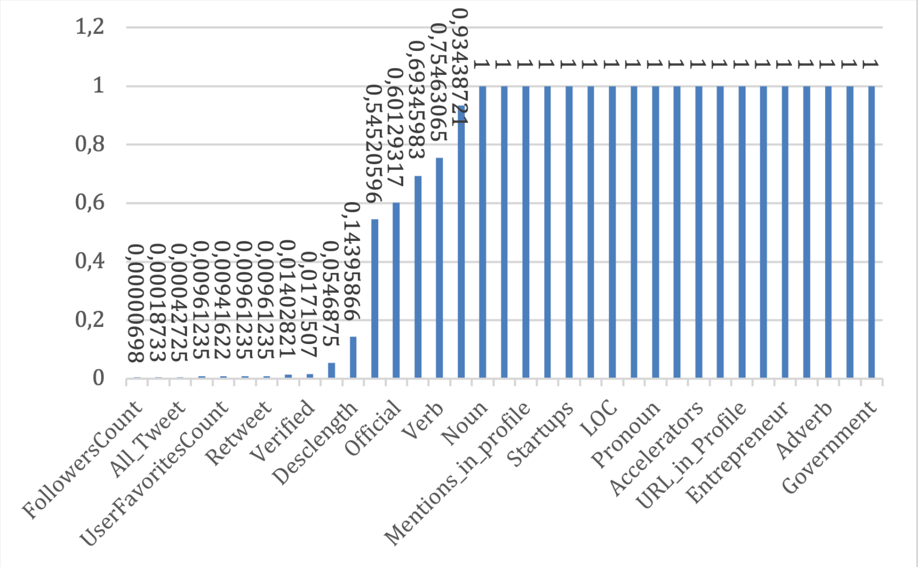

Fig. 2. The aggregated list features and associated score (P-value)

3.6 Evaluation

To evaluate the final aggregated list, researchers used the concept of an incremental feature selection (IFS) [31]. In IFS, supervised machine learning algorithms are used to evaluate the features which sorted according to its importance. It works as follows: the algorithm is trained on only the first best attribute, then the top 2, then top 3 and continue until finishing all the features. In each iteration, the algorithm returns the accuracy. In this paper, we used precision as evaluation metrics [32]. As shown in equation 8 precision is the number of true positives (the number of correctly predicted influencers) divided by the total number of elements classified as positive class (influencers) (the sum of correctly and incorrectly predicted influencers) [32]. We used it due to its ability to deal with imbalanced class distribution. In this research case, there are 28 influencers out of 200 users.

$$\text{Precision= True\_Positive/ (True\_Positive+ False\_Positive)} \qquad (8)$$

Three different types of state-of-the-art algorithms trained in a train-test fashion, they are Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF). The algorithms were fed the aggregated list incrementally. Figure 3 shows the precision results of all iterations. Each number represent the number of features in the iteration. For example, '1' means the best feature (highest significant *P*-value), while '2' means the two best features. The significant features are the first nine features.

As shown in the figure, the performance of NB starts with 0.86896 in the first iteration and then it increased incrementally until it reaches its highest performance with 0.948365 in the ninth iteration, then it becomes stable. SVM provides better performance reaching 0.95254 from the first and second iterations, but its performance declined in the third iteration to reach 0.8711111, then its performance is stable to the final iteration. Compared with SVM and NB, RF started with the lowest performance reaching 0.8292397, then the performance increased incrementally until it reaches its highest performance in the ninth iteration equal to 0.910169, then it declined and became stable. Table 3 shows the performance of the three-algorithms based on precision for the nine significant features.

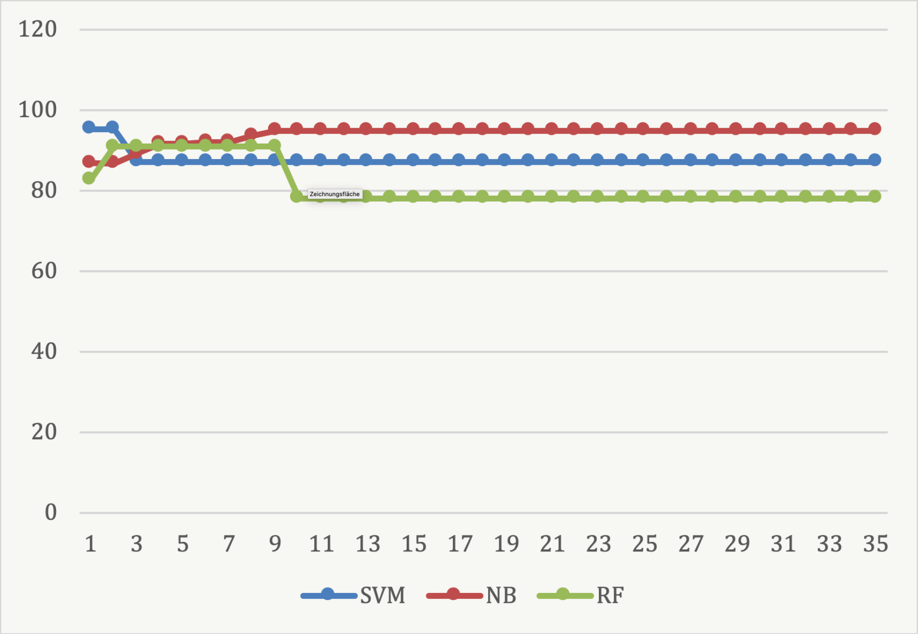

Fig. 3. The performance of the models based on precision

Table 3. Performance of the three-algorithm based on precision for the nine significant features

	1	2	3	4	5	6	7	8	9
SVM	95.25423	95.25423	87.11111	87.11111	87.11111	87.11111	87.11111	87.11111	87.11111
NB	86.896552	86.78362	89.25925	91.71608	91.71608	92.09876	92.09876	93.72549	94.83651
RF	82.92397	91.01694	91.01694	91.01694	91.01694	91.01694	91.01694	91.01694	91.01694

4 Discussion

Only the first nine features with significant *P*-value are considered the essential features of entrepreneurial influencers since 0.05 is used as the cutoff for significance. The 'number of followers' is considered the most crucial characteristic of the entrepreneurial influencers followed by the number of the list. These two features reflect the importance of entrepreneurial influencer's popularity. The user's popularity may be increased by activity level. Therefore, we found the influencer's activity 'All Tweet' is the third essential features. This result is in agreement with Asadi at el. 2018 [16] who found that most of the influencers conversations ranged across different topics as personal experiences, travel, or politics.

Ranking 'Favorite' as the fourth entrepreneurial influencers reflects that many of influencers' audience is made up of followers who act as observers than participants in the conversation. The 'User Favorite Account' and 'Reply' are ranked as the fifth and sixth most essential features, reflecting the influence of the influences' interaction level. This also agrees with Asadi at el. 2018 [16] who reported that the majority of influencers spent their time in interaction with their audience. The quality of tweets 'Retweet' feature is the seventh feature distinguishes the entrepreneurial influencers. This is a logical result since the entrepreneurial users especially the beginner entrepreneurs, and the founder of Small and Medium Enterprises SMEs usually look for the information guide them. This result corresponds with the result of Kuffo at el. 2018 [33] who found that entrepreneurs rely more on local sources for information. The actively level of influencers again proves its importance in term of 'Tweet' feature which ranked as the eighth most important feature distinguish the entrepreneurial influencers .it is the number of influencers tweet related to entrepreneurial issues. This find agrees with As Kuffo at el. 2018 [33] who found in his research entrepreneurship-focused sources are more popular among entrepreneurs. Finally, the profile features, 'verified' is ranked on the ninth position on the ranking list, reflecting how much the influencers' account must be reliable.

5 Conclusions and future work

In this paper, researchers focused on the problem of detecting valuable features of entrepreneurial influencers on Twitter, in particular, Saudi's influencers. At the first stage, a wide range of features are collected in order to be investigated for the performed research. These features are coming from several research domains such as social media analysis, natural language processing, and retrieval information studies. It then proposed a robust framework called ECharacterize to rank the most relevant features distinguish the Saudi entrepreneurial influencers. Three state-of-the-art machine learning supervised algorithms are used to evaluate the final results to ensure the correctness and efficiency. Based on the experimental, we can highlight following main results. First, the entrepreneurial influence is based on the number of followers and the number of followers who have added those influencers to a list. Second, the level of activity distinguishes those account either on term of entrepreneurial tweets or general tweets. Third, their continue conversation are selected on the basis of evidence that they keep strong influence, passive members who participated by liking tweets are also considered. Finally, the influence also related to the reliability of the account.

6 References

[1] D. Kuss, M. Griffiths, D. J. Kuss, and M. D. Griffiths, "Social Networking Sites and Addiction: Ten Lessons Learned," Int. J. Environ. Res. Public Health, vol. 14, no. 3, p. 311, Mar. 2017. https://doi.org/10.3390/ijerph14030311

[2] S. Shane, "The Importance of Angel Investing in Financing the Growth of Entrepreneurial Ventures," Q. J. Financ., vol. 02, no. 02, p. 1250009, Jun. 2012. https://doi.org/10.1142/s2010139212500097

[3] F. Jin, A. Wu, and L. Hitt, "Social Is the New Financial: How Startup Social Media Activity Influen Funding Outcomes," Acad. Manag. Proceedings., p. 13329, 2017. https://doi.org/10.5465/ambpp.2017.13329abstract

[4] A. Papa, G. Santoro, L. Tirabeni, and F. Monge, "Social media as tool for fa-cilitating knowledge creation and innovation in small and medium enterprises," Balt. J. Manag., vol. 13, no. 3, pp. 329–344, Jul. 2018. https://doi.org/10.1108/bjm-04-2017-0125

[5] Y. Motoyama, S. Goetz, and Y. Han, "Where do entrepreneurs get information? An analysis of twitter-following patterns," Small Bus. Entrep., vol. 30, no. 3, pp. 253–274, 2018. https://doi.org/10.1080/08276331.2018.1435187

[6] C. Riverola and F. M. On, "Entrepreneurs' Bricolage and Social Media," in 2018 IEEE International Conference2018, .

[7] C. D. M. M. C. R, "Identifying influential and susceptible members of social networks," Science (80-)., vol. 329, no. 0036–8075, pp. 1194–1197, 2012.

[8] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' net-works," Nature, vol. 393, no. 6684, pp. 440–442, Jun. 1998. https://doi.org/10.1038/30918

[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 11, pp. 2169–2188, Nov. 2009. https://doi.org/10.1002/asi.21149

[10] J. V. Cossu, V. Labatut, and N. Dugué, "A review of features for the discrimination of twitter users: application to the prediction of offline influence," Soc. Netw. Anal. Min., vol. 6, no. 1, Dec. 2016. https://doi.org/10.1007/s13278-016-0329-x

[11] I. Sangaiah, A. Vincent, A. Kumar, A. Balamurugan, and I. Sangaiah, "An Empirical Study on Different Ranking Methods for Effective Data Classification," J. Mod. Appl. Stat. Methods, vol. 14, no. 2, p. 7, 2015. https://doi.org/10.22237/jmasm/1446350760

[12] X. Li, X. Wang, and G. Xiao, "A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications," Brief. Bioinform., vol. 20, no. 1, pp. 178–189, Jan. 2019. https://doi.org/10.1093/bib/bbx101

[13] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, “Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?” IEEE Trans. Dependable Se-cur. Comput., vol. 9, no. 6, pp. 811–824, 2012. https://doi.org/10.1109/tdsc.2012.75

[14] K. Lee, P. Tamilarasan, and J. Caverlee, “Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Me-dia.”

[15] G. de-la-Ramírez-Rosa, E. Villatoro-Tello, H. Jiménez-Salazar, and C. Sánchez-Sánchez, “Towards automatic detection of user influence in twitter by means of stylistic and behavioral features,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8856, pp. 245–256, 2014. https://doi.org/10.1007/978-3-319-13647-9_23

[16] M. Asadi and A. Agah, “Characterizing User Influence Within Twitter,” in International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2018, pp. 122–132. https://doi.org/10.1007/978-3-319-69835-9_11

[17] J. Mustafi, “Natural Language Processing and Machine Learning for Big Da-ta,” in Techniques and Environments for Big Data Analysis, Springer, Cham, 2016, pp. 53–74. https://doi.org/10.1007/978-3-319-27520-8_4

[18] K. J. Archer and R. V. Kimes, “Empirical characterization of random forest variable importance measures,” Comput. Stat. Data Anal., vol. 52, no. 4, pp. 2249–2260, Jan. 2008. https://doi.org/10.1016/j.csda.2007.08.015

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Mach. Learn., vol. 46, no. 1–3, pp. 389–422, 2002. https://doi.org/10.1023/a:1012487302797

[20] M. D. Shieh and C. C. Yang, “Multiclass SVM-RFE for product form feature selection,” Expert Syst. Appl., vol. 35, no. 1–2, pp. 531–541, Jul. 2008. https://doi.org/10.1016/j.eswa.2007.07.043

[21] W. P. Alston and F. I. Dretske, “Knowledge and the Flow of Information.,” Philos. Rev., vol. 92, no. 3, p. 452, Jul. 1983.

[22] M. a. Hall and L. a. Smith, “Practical feature subset selection for machine learning,” Comput. Sci., vol. 98, pp. 181–191, 1998.

[23] I. Guyon and A. M. De, “An Introduction to Variable and Feature Selection André Elisseeff,” 2003.

[24] H. Liu and R. Setiono, “Chi2: feature selection and discretization of numeric attributes,” in Proceedings of the International Conference on Tools with Artificial Intelligence, 1995, pp. 388–391. https://doi.org/10.1109/tai.1995.479783

[25] R. Kolde, S. Laur, P. Adler, and J. Vilo, “Robust rank aggregation for gene list integration and meta-analysis,” Bioinformatics, vol. 28, no. 4, pp. 573–580, Feb. 2012. https://doi.org/10.1093/bioinformatics/btr709

[26] V. Andonova, M. S. Nikolova, and D. Dimitrov, “What Is an Entrepreneurial Ecosystem?” in Entrepreneurial Ecosystems in Unexpected Places, Cham: Springer International Publishing, 2019, pp. 3–16. https://doi.org/10.1007/978-3-319-98219-9_1

[27] A. Pasha et al., “MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic,” in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), 2014, pp. 1094–1101.

[28] R. G. Pontius and M. Millones, “Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment,” Int. J. Remote Sens., vol. 32, no. 15, pp. 4407–4429, Aug. 2011. https://doi.org/10.1080/01431161.2011.552923

[29] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, “Data Preprocessing and Intelligent Data Analysis,” Intell. Data Anal., vol. 1, no. 1, pp. 3–23, Jan. 1997.

[30] S. G. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage," Comput. Sci., vol. 74, no. 5, pp. 32–40, Mar. 2015.

[31] H. Liu and R. Setiono, "Incremental Feature Selection," Appl. Intell., vol. 9, no. 3, pp. 217–230, 1998.

[32] P. A. Flach PETERFLACH, "An Analysis of Rule Evaluation Metrics Johan-nes F urnkranz," in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 202–209.

[33] L. Kuffo, C. Vaca, E. Izquierdo, and J. C. Bustamante, "Mining Worldwide Entrepreneurs Psycholinguistic Dimensions from Twitter," in 2018 International Conference on eDemocracy & eGovernment (ICEDEG), 2018, pp. 179–186. https://doi.org/10.1109/icedeg.2018.8372352

7 Authors

B.Moheel Almotairy 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.

M.Abdulaziz Abdullah. received her PhD in computers and systems engineering, Faculty of engineering, Ain Shams University, Cairo, 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 pattern recognition. Dr Abdullah published more than 120 research papers in various international journals and conferences. She has also joined many HiCi research projects all over the world.

R.Abbasi completed his PhD 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. He has published more than 35 articles in reputed journals like IEEE Computational Intelligence Magazine, Computers in Human Behavior, Telematics and Informatics, Applied Soft Computing, and Scientometrics and international conferences like ACM HyperText Conference, ACM World Wide Web Conference, Pacific Asia Conference on Knowledge Discovery and Data mining, and European Conference on Information Retrieval.

Article submitted 2020-04-13. Resubmitted 2020-05-15. Final acceptance 2020-05-16. Final version published as submitted by the authors.