User profiling for big social media data using standing ovation model

Muhammad Al-Qurishi[1] **· Saad Alhuzami**[1] **·**
Majed AlRubaian[1] **· M. Shamim Hossain**[1,2] **·**
Atif Alamri[1,2] **· Md. Abdur Rahman**[3]

Received: 15 April 2017 / Revised: 22 October 2017 / Accepted: 8 November 2017 /
Published online: 22 November 2017

Abstract Online Social Networks (OSNs) have recently been the subject of numerous studies that have attempted to develop effective methods for classifying and analyzing big content. Some of the key contributions of these studies to current scientific understanding include the identification of underlying topics within content (posts and messages), determination of each user's influence and contributions, c) measurement of content quality, and extraction and analysis of users' motives and preferences. We aimed to develop an integrative solution entailing a combination of these methodological advances within a single framework that could facilitate attribution and differentiate OSN members. Specifically, we examined peer effects within Twitter and assessed the propensity of members to alter their views on commonly discussed matters based on their exposure to alternative views expressed by respected and influential members. We availed of abundant available resources

✉ M. Shamim Hossain
mshossain@ksu.edu.sa

Muhammad Al-Qurishi
qurishi@ksu.edu.sa

Saad Alhuzami
saad4q@gmail.com

Majed AlRubaian
malrubaian.c@ksu.edu.sa

Atif Alamri
atif@ksu.edu.sa

[1] Chair of Pervasive and Mobile Computing, Collage of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[2] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[3] Department of Computer Science, University of Prince Muqrin , Madinah al munawarah, Saudi Arabia

and tracked historical interactions of selected users to create a workable model that captured differences in opinions. The resulting solution enables peer influence within the online environment to be quantified and the level of investment of identified social media users in particular topics to be assessed.

Keywords Data processing · Big data mining · Analysis of textual content · Twitter

1 Introduction

The earliest social media platforms date back to the mid-1990s, although the extent of their potential impact on mainstream media was not apparent at that time [13]. It was inconceivable then that less than two decades later, leading networks would have hundreds of millions of active members or that their values on Wall Street would extend to ten-digit figures. Practically every facet of social life and human behavior has been transformed by this technological advance, and it is now difficult to remember a time when individuals could not express their opinions on a range of events via social media. Through the advance of social media, a huge trove of data has been generated that can be used for information mining and artificial learning. Consequently, social media has emerged as one of the most studied subjects in the field of information technology during the last decade, which has been marked by numerous seminal discoveries [9, 10].

Our aim in this study was to develop a new solution for deconstructing communication patterns within any online social media platform. More precisely, this solution relates to on the phenomenon of peer effect (social influence) that exists in any socially cohesive group, but is particularly pronounced within virtual communication systems. Microblogging platforms like Twitter empower their members to interact actively with social media content by posting their own entries or re-posting existing content, mentioning a particular post or member, or simply agreeing with ("liking") the original message. All of these modes of interaction can be useful for understanding mutual relations between network members and their impacts on each other's attitudes [3, 18].

The solution outlined in this paper is based on an existing model, the Standing Ovation model (SOM), developed by Miller and Page [25] in 2004. For this study, we attempted to map the model's concept to differentiate individual members of the Twitter community based on their actions within this online social network (OSN). The model was deemed adequate for the task, because the OSN fulfills the requisite methodological criteria of complexity and adaptability. It captures various dimensions of online communication, entails a heterogeneous environment, provides diverse incentives, and enables knowledge adoption. Moreover, its multidisciplinary nature offers significant advantages. The SOM can be applied to all of the abovementioned data types, and is therefore considered a particularly appropriate technique for analyzing Twitter activity and incorporating a number of dimensions in descriptions of each member.

Examples of the SOM's application can be found in everyday life. For example, when a well-received lecture ends in a loud round of applause through hand clapping, some members of the audience may rise from their seats and be joined by many others in a massive display of respect for the lecturer. Numerous factors influence each individual's decision to join in this kind of group action. In this paper, we demonstrate that the SOM is a very useful framework for solving problems of variable difficulty. In general, the SOM focuses on activities and their underlying incentives, creating a matrix of interactions that allows for better prediction of the future activities of Twitter users.

The rest of the paper is organized as follows. The following section outlines the background and presents a review of the literature on SOM-related problems and topics of relevance to this study. In the third section, we provide a detailed description of the proposed framework. The fourth section focuses on the model's implementation and provides details on an experiment that we conducted using Twitter social media content. In the final section, we offer conclusions based on our study.

2 Related studies

2.1 Twitter social network

From the time of its launch in 2006, Twitter has undergone phenomenal expansion, acquiring a vast membership and attracting the attention of the scientific community [22]. Java et al. [20] examined data derived from 70,000 users and subsequently divided the entire data content into four clusters: daily chatter, conversations, sharing of information or links, and news reporting. They also attempted to develop a linear analysis of Twitter's expansion. Moreover, they demonstrated that the characteristics of this OSN are independent of network size, providing insights regarding members' distribution within each location. Another study [11] conducted on a sample of 100,000 members developed three categories of members: "sources" followed by a large number of members, "participants" demonstrating a balanced ratio between incoming and outgoing connections, and "recipients" who receive content from many other members, but who are rarely followed themselves. Other studies [4] have attempted to draw connections between individual users and particular themes, or have assessed Twitter's social relevance rather than its technical features. From the perspective of its members, this network is similar to the blogosphere, the key difference being the establishment of formal rules that favor brief postings rather than lengthy text for the former. It also has practical applications and is being widely embraced by businesses because of its marketing potential and because it offers a convenient method of direct communication between businesses and their customers or potential employees [4].

2.2 User profiling

The term "user profiling" denotes all actions associated with obtaining, studying, and applying data related to user behavior within a network. The resulting profile can facilitate more precise delivery of content. The profiling process usually comprises critical phrases entailing themes that particular members are likely to seek [12]. A more detailed classification may include data on common activity patterns, for example, pages visited or the frequency of logins [37]. Some scientists believe that deciphering connections between members and their mutual interactions can shed light on each user's genuine interests. Such insights have demonstrably enhanced the predictive power of numerous theoretical models and practical systems. Although previous studies have extensively explored this method for deducing users' interests, we nevertheless decided to remove any possible doubts by verifying it through a test performed on a sample extracted from a live network. Our chosen methodology was to describe a user's characteristics and activities prior to leveraging this knowledge within a comprehensive classification system. A manually operated process was conducted to determine the original scope of characteristics, and lessons learned during this phase were applied to automate the process during the subsequent phase. To confirm our theoretical assumptions about the classification criteria, we executed a ten-fold cross-validation procedure, which provided empirical proof of the effectiveness of our strategy.

2.3 Mining social media content

When OSN users share information, their mutual relationships can be used for selecting content and displaying algorithms. It is possible to select content for a particular member based on the influence of individuals who are being followed by this member [19]. Relationship-based selection has its advantages, but suffers from the issue of limited available data. By contrast, this issue is not encountered when selection methods focusing on content rather than connections are used. Some researchers have also attempted to use a combined methodology, for example, by building a hierarchy of categories based on an analysis of accumulated data about members' activities within the network [29].

Nearly 80,000 new blogs are created daily, with more than a million published content pieces appearing within the same timeframe [41]. It becomes clear from a consideration of the content created and shared through social platforms that the online environment contains an abundance of information about people, including their opinions and preferences on a diverse range of subjects. The attitudes conveyed within these channels can be understood as indicators of positions of support or opposition relating to a broad spectrum of local or international issues [14]. These attitudes are often articulated in very strong terms and can be used to drive various marketing initiatives linked to the issues in question. The application of classic techniques can be of value in performing the task of attitude detection, but they are unsuitable for working with massive amounts of data that are typically gathered from social platforms. Consequently, the development of innovative data extraction methods is a crucial task that our study attempted to accomplish.

3 Preliminaries

In this section, we provide brief background information on the SOM, peer effects, and other key terms used in this study.

3.1 Standing ovation definition

A basic definition of standing ovation pertains to the phenomenon that is often witnessed at the conclusion of spectacular performances or speeches [3]. This form of social behavior is widely encountered in many venues. Because of its familiarity, its importance may not be immediately apparent. However, we believe that this paradigm can be applied in the context of online social media, as it provides a convenient way to dissect and analyze current trends within a dynamic and highly connected environment such as Twitter [25, 37].

System modeling that is based on the standing ovation phenomenon does not entail a strict procedure that must be consistently followed. The use of a wide spectrum of different methods can be appropriate under the right circumstances, with statistical analysis, programming, and verbal content management being of particular value. In this study, incoming data required for the SOM framework was collected with the help of existing data filtering algorithms.

The SOM framework applied in our study operated under the following four major assumptions:

- *Every interaction has a property* $\boldsymbol{Q}$. *Whereas interaction can be generally defined as any activity or expressed opinion, interactions in the specific context of this study are defined as tweets.*

- *Every member of the audience is targeted with the message **S**, expressed in the following equation:*

$$\boldsymbol{S} = \boldsymbol{Q} + \boldsymbol{e} \tag{1}$$

where **e** is defined as imprecision, which varies from user to user.

- *If **S** initially exceeds a user's limit (threshold), then that user will stand up.*
- *As the process continues, users will stand up if more than **X%** of the audience is already in a standing position.*

3.2 Measuring user influence

Social impact or influence can be achieved either through direct means or through a chain of side activities that influence the final outcome. Though every day and scientific usage of this term is prevalent, its precise measurement is an extremely complex task, and there is no widely accepted method available for its execution that would meet the highest criteria [21, 32].

Some previous studies treated Twitter as a classic news dissemination mechanism, examining the types and degrees of social impacts associated with this mechanism. The main objective of these studies was to acquire an understanding of how one user could motivate others to take a desired action, with three major mechanisms being discernable. The first relates to users who publish large amounts of engaging content and have a broad base of contacts from whom they actively seek inputs. The second relates to users who re-broadcast important bits of content from third-party sources (retweets, in Twitter parlance) to their network contacts. The last mechanism entails responding or discussing previously posted content, which is termed "mentioning" in the jargon of this social network.

3.3 Finding major topics of interest

Users' preferences can be deduced from their online activity. This process has already been well examined within scientific studies [1, 5, 6, 34]. A key study on this topic conducted by Abel [1] attempted to define a system for grading topic-driven networks. This system, which centered on topics, hashtags, and objects, enabled the organization of a dataset compiled from various tweets. The frequency-based method was used in this study to determine the ranking of every examined user. Tao [34] used a similar methodology to develop his Twitter-based User Modeling Service (TUMS) model, which was aimed at differentiating users based on the content of their tweets. Contrasting with our study, Tao's main objective was to present semantic structures rather than describe individual members. Attempts were also made to calculate a composite grade for two or more social networks, some of which influenced the design of this solution.

However, searching for topics of interest within a sample composed of numerous tweets is a very time-consuming activity, because data samples can be so large that the amount of work required is daunting. A more feasible task entails determining what members are talking about on Twitter from a different perspective [24]. One way of doing this is to track topics that appear on the timeline of any particular member and subsequently to deduce their priorities from this input. Here, we utilized a well-known solution for topic identification, namely, latent Dirichlet allocation (LDA) [26, 33].

This approach requires the definition of what constitutes a topic as a first step. Within the scope of LDA, a topic is understood to mean a group of verbal elements. Every topic can be broken down into a list of words that includes the likelihood of a given word featuring in relation to that topic and the inclusion of all of the words for every topic. Even when individual words are identical, they can be assigned different grading coefficients. Thus, for a topic related to athletic competition, the following words and the likelihood of their occurrence, in descending order, could be:" soccer" (25%), "tennis" (15%), "swimming" (10%). .. "Trump" (0.1%), and "Pentagon" (0.05%). By contrast, a topic related to the U.S. government could include the following words and their frequencies: "Trump" (35%), "Pentagon" (10%). .. "soccer" (0.6%), and "tennis" (0.2%). The words appearing at the bottom of the list can be ignored for all practical purposes [26].

3.4 Ranking content by quality

Twitter content can be clustered according to various categories such as educational, casual, and humorous depending on the underlying intention. Rather than focusing on any one cluster, our aim was to develop a model that could be used to estimate the values of all categories [36]. Consequently, for our purposes, any tweet had to satisfy three distinct requirements to be accepted as "engaging." The first requirement was that of good form, which referred to proper composition, error-free language, and ease of comprehension. Tweets that used jargon and evidenced poor grammar or semantic confusion were rated lower. The second requirement was objective focus. This meant that content which had significant informational value and referred to actual events in the world was ranked above content that primarily comprised an individual's views or had no coherent focus [27, 28, 35]. The last requirement was directional value, whereby a higher value was accorded to tweets with embedded links to external sources, enabling additional information on a given topic to be found [15, 30]. Clearly, some links would be considered more valuable than other links. Consequently, the quality of the landing page needed to be assessed.

4 The proposed model

We demonstrated that the proposed solution could deliver results that were positively correlated with the reliability of data compiled from Twitter. This OSN mainly comprises textual elements, but it also includes social features that can be utilized to learn more about the connections between individuals and their friends. Here, we discuss a solution capable of enabling the performance of such an analysis, describing each of its constituent parts and their interactions with the rest of the system (Fig. 1).

The solution based on the SOM hinged on three specific parameters. First, its operation was related to the quality of inputs, which in this particular study were Twitter posts. Next, it was necessary to define the user's threshold relating to the level of quality, which denoted the lowest possible value of content required for the user to sustain interactions with the tweet in question. Because this limit varied from topic to topic, the last parameter that needed to be controlled was topic priority, which could be operationalized as a list of the five highest ranked topics that a user deeply cared about and regularly read about to update his or her knowledge.

In this section, we present the model design, clarifying the mechanism for the mutual coordination of its independent components. Prior to examining mutual relations between

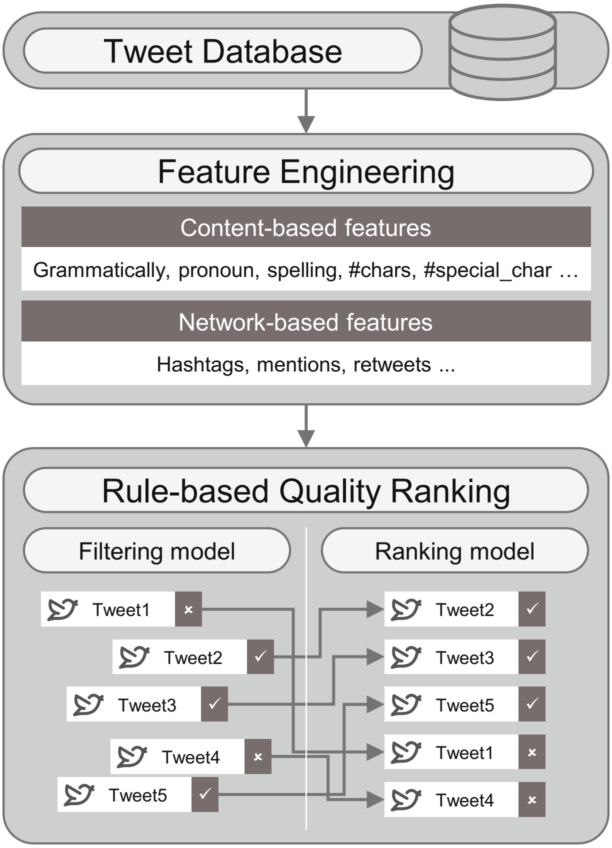


Fig. 1 Overall process of filtering and ranking tweet based on quality analysis

users, it was necessary to implement four separate steps. These were: (1) assessing the influence of each user, (2) investigating the topics and quality of each tweet, (3) determining areas of priority interest for each user, and (4) measuring the thresholds of users relating to particular interests. Only after implementing each of these prerequisite steps could we turn our attention toward the main subject of our study, namely social effects.

It is not necessary to follow the exact sequence of the first two steps, which can be performed in any order. We assessed the influence of each user included in the compiled list based on the three incoming variables shown in Fig. 2. These variables were the total number of connections, the number of re-tweets, and total mentions by others. The following equation was used to measure user influence.

$$UI_u = \left(\frac{\deg_u + RP_u + Men_u + RT_u}{N} \right) \tag{2}$$

Where $\deg_u$, RP_u, Men_u, RT_u respectively denote the number of followers, replies by others, mentions by others, and re-tweets by others in relation to user u. UI_u denotes user influence and N is the total number of users who interact with user u.

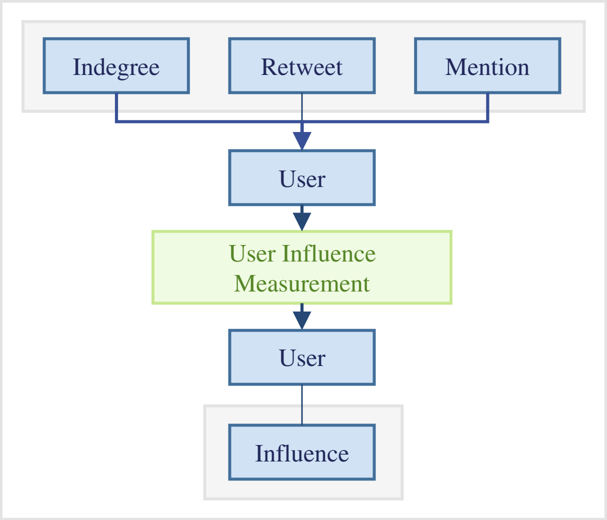


Fig. 2 Measuring user influence

For the next step, tweets were rated according to their quality and were categorized according to the topics that they addressed. These topics assumed importance at a later stage of the analysis entailing deduction of the main interests of each user. Measuring the quality of tweets was required to calculate users' reaction thresholds for particular topics. When estimating the value of a tweet, formal and language elements were considered equally with emotional reactions. Figure 3 depicts the process applied to analyze the topics and quality of tweets.

The above two steps had to be completed before implementing the third step, because a user's interests could only be identified after the raw data had been classified according to topics. The priority interests of users were determined by calculating the number of tweets that featured a certain topic with which the user in question had previously engaged. Following the determination of which tweets referred to a certain topic conducted during a previous step, it became possible to connect a user to the topic by checking his or her profile history (timeline) as shown in Fig. 4. Five topics with which the user had engaged most frequently were used for further reference, although this number was chosen arbitrarily and more topics could have been included.

The last remaining parameter for analyzing peer effect was the reaction threshold of a user for a particular topic. This parameter was obtained by examining all of the content in a user's profile associated with one of the five top-ranked topics. Given that the quality of each content piece had already been determined, the user's reaction threshold for a given topic could be calculated as the average value of all the tweets with which the user had interacted. One weakness of this solution is that when all relevant tweets are considered, it is not possible to differentiate interactions based on legitimate interests and those that convey irony. Figure 5 depicts the procedure for obtaining the threshold limits of users.

Peer effect within the online environment was examined after all of the listed conditions had been satisfied. The following basic formula was used to determine whether a member would react to a tweet or not:

$$\textit{If}\ \ \mathbf{Q} > \ = \boldsymbol{T}, \textit{there will be a reaction} \tag{3}$$

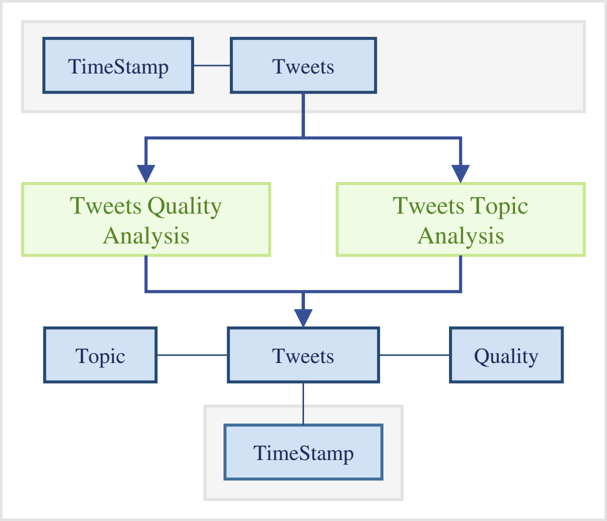


Fig. 3 Analyzing tweets topics and quality

where $\boldsymbol{Q}$ denotes the value of the overall quality and $\boldsymbol{T}$ denotes the threshold limit per user. Situations entailing users' reactions to content valued below their reaction limits are unusual, indicating the presence of external elements that motivate users to deviate from their normal activity pattern. This external element was central for our solution and was formalized as $\boldsymbol{x}$, based on the premise that if $\boldsymbol{x}+\boldsymbol{Q}>\boldsymbol{T}$, then the member will react even though $\boldsymbol{Q}$ may not be greater than $\boldsymbol{T}$.

Factor $\boldsymbol{x}$ denoted peer effect associated with another person within the network. This factor varies from topic to topic, as users trust the opinions of particular individuals in areas in which they consider these individuals to be knowledgeable. Inclusion of this factor in our model introduced a personal dimension into the description of a user and complemented statistical

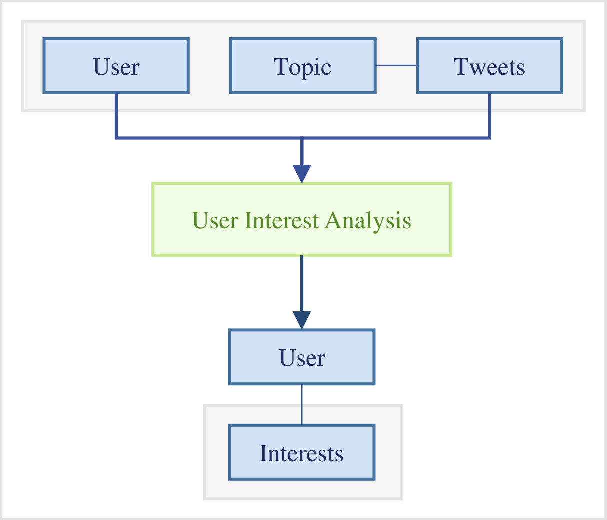


Fig. 4 Identify user preferences and interests

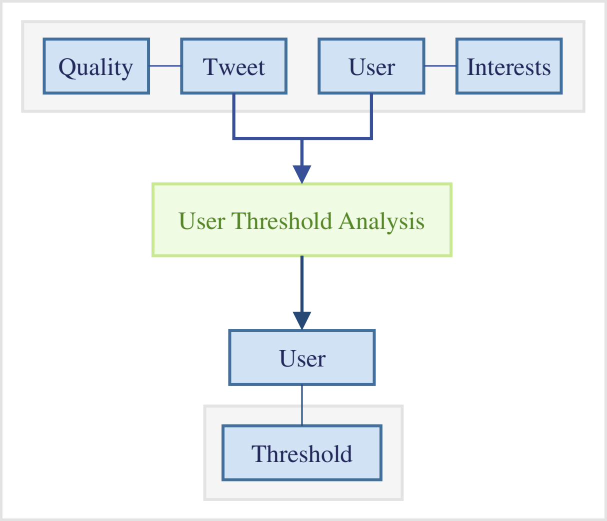


Fig. 5 User threshold analysis

parameters by allowing for the actions of a user to be motivated by respect or regard for other members. Figure 6 illustrates the final step in the analysis of the peer effects of Twitter users.

5 Experimental study

In this section, we discuss the practical implementation of the SOM-based solution and its application to a realistic social media dataset. We carefully describe each step of the empirical testing procedure in unambiguous terms. However, before discussing how the data were harvested and organized, we first briefly revisit the methodology and utilities. Subsequently, the building blocks of the solution (user influence, topic systematization, content quality, and peer effect) and their roles are explained in more detail.

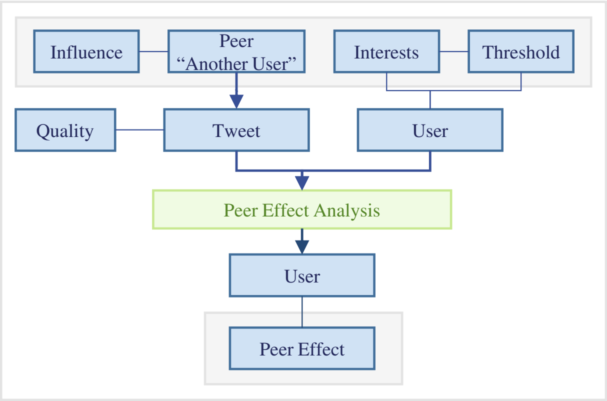


Fig. 6 Peer-effect analysis

5.1 Methods and tools

The system presented here is unique and entails several steps that include compiling the original content, identifying topics prioritized by each user, and implementing quality control to determine the value of each tweet. Mathematical calculations were applied to obtain the values of the peer effect factor as well as user strength coefficients.

Data harvesting for this study was based on several predetermined characteristics. 100 K users were harvested, and 2000 were chosen randomly from the ranks of Twitter users who communicate in the English language and have more than 5000 followers. The original aim was to harvest more than 3200 tweets from each of the selected users. However, the actual number of compiled content pieces was lower because of the Twitter rate limit. Tweets were imported using specialized modules that we developed for our system and were stored in a prepared database for subsequent processing. Between 1600 and 3200 tweets were gathered from each member, and the sample was finalized, with no further modifications being made.

5.2 User influence

Based on Eq. 2, we calculated users' impacts. The solution was designed to enable the general level of impact of each user on his or her followers to be graded based on a score ranging from 1 to 100. The score implied a capacity to prompt other users to react in accordance with the objectives of this study. Table 1 provides an example of the collected profiles after calculating the influence of the users.

5.3 Topic modeling and analysis

After obtaining a complete data sample and calculating the impact of each user, the next operation entailed organizing topics of interest for each user [38]. Accomplishing this operation required the execution of multiple actions that are described below:

1) *Sample filtering*

The compiled sample was transferred to a database generated by the system where it could be adequately cleaned and prepared for further processing. All upper case letters were converted into lower case ones and embedded links were removed, because they held no meaningful information for this study. Grammatical markers and stop words (such as "a" or "and") were eliminated from the sample. Empty spaces between words were also deleted. Following this cleaning procedure, the tweets were ready for a more comprehensive examination.

2) *Words frequency and word cloud*

The next important step was to assess the volume of the total number of appearances of each word in the sample, as this enabled the systematization of topics that featured in the content. This procedure was repeated for each user in the sample. Thus, a list of the most frequently appearing phrases was formulated for each user. Figure 7 depicts a word cloud composed of different topics associated with the elections held in 2016 in

Table 1 Example of the harvested users' profiles (Top 10 users)

#	Friends	Followers	FRF	Posts Count	Weekly Posts Rate	Repost Count	Repost By Others	Signal Strength	Reply Count	Reply By Others	Mentions Count	Mentions By Others	Listed Count	Favorites Count	Account Age	User Influence
0	232	4,361,899	0.0001	83,846	211	7	189	0.9999	1	62	30	270	5271	55	2758	0.133126
1	316	2,766,343	0.0001	395,619	1017	544	636	0.9986	0	0	549	952	5693	8	2723	0.099214
2	4	2,699,485	0	308,629	899	0	2	1	0	0	68	60	4921	98	2387	0.073528
3	2401	1,876,805	0.0013	190,892	427	640	1626	0.9967	0	108	533	4647	11,383	716	3112	0.147054
4	716	1,869,174	0.0004	274,140	531	0	3867	1	0	992	10	5983	19,545	26	3603	0.330672
5	175	1,807,750	0.0001	356,293	852	281	2023	0.9992	46	7	649	2392	27,078	434	2928	0.121922
6	785	1,437,794	0.0005	80,386	171	186	906	0.9977	1	275	371	1992	11,386	175	3264	0.115306
7	762	1,367,535	0.0006	110,494	214	870	355	0.9922	1	227	1664	2583	18,147	642	3583	0.098525
8	58	1,115,006	0.0001	145,238	412	42	17	0.9997	5	3	98	71	1947	1120	2442	0.03158
9	474	1,105,201	0.0004	10,839	33	1310	2409	0.8922	60	22	2507	6657	11,131	740	2283	0.133126

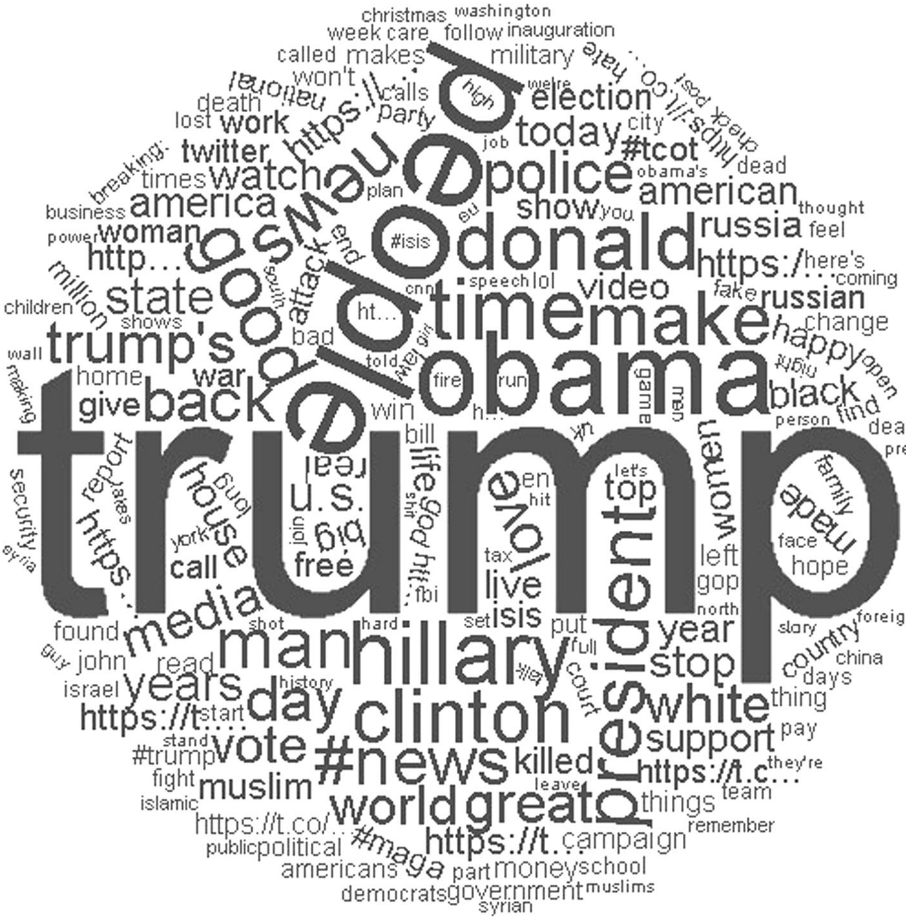


Fig. 7 Words frequency and word cloud (Election 2016)

the United States, and Table 2 presents the most important topics identified along with the top 20 words.

3) *Words Co-occurances and Clustring*

Based on an examination of which words co-occurred within the same tweets, we developed a matrix of verbal expressions, highlighting links between certain words and word groups. Pairings that were found to occur more than 50 times were positioned at the edges of a visual graph, delineating the structure of the word cloud. The resulting figure was then refined to create a hierarchical tree of themes defined as tightly connected groups of words. The LDA solution was the tool of choice for identifying words associated [2, 15, 31, 38] with a particular topic. This algorithm calculates the probability of appearance for every verbal element in a dataset. Consequently, the topic in question necessarily featured high-volume words that appeared in 30% or more of all

Table 2 Selected topics and the most frequent words

Topic 5		Topic 10		Topic 40		Topic 90		Topic 97	
Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
trump	0.042293	trump	0.044453	trump	0.032293	dog	0.012666	game	0.011042
vote	0.022023	ban	0.014002	obamacare	0.024935	dogs	0.008607	sports	0.010576
hillary	0.013644	immigration	0.012669	repeal	0.011419	amp	0.008109	news	0.007578
election	0.011693	order	0.012042	gop	0.010133	animal	0.006409	nfl	0.006891
maga	0.008899	trumpss	0.011396	health	0.010133	zoo	0.006062	win	0.006581
amp	0.008797	wall	0.011318	house	0.009633	rescue	0.005866	coach	0.004501
votes	0.007702	obama	0.01056	obama	0.00918	animals	0.005697	season	0.004252
clinton	0.007368	border	0.010441	jobs	0.008926	baby	0.005421	bowl	0.004087
donald	0.007257	executive	0.00922	care	0.008456	adopt	0.004148	nba	0.003887
president	0.00721	president	0.009111	republicans	0.008357	home	0.004148	yahoo	0.003838
electoral	0.00648	refugees	0.00794	tax	0.008006	cat	0.00381	team	0.003835
voter	0.006478	illegal	0.007206	plan	0.007943	cats	0.00365	football	0.003669
win	0.006478	mexico	0.007145	amp	0.005493	save	0.003552	year	0.003439
people	0.006025	donald	0.00617	congress	0.00545	foster	0.00316	patriots	0.003396
america	0.005947	muslim	0.006086	bill	0.005308	shelter	0.003062	amp	0.003277
won	0.005945	refugee	0.005527	senate	0.005099	video	0.002973	back	0.003102
voting	0.005921	immigrants	0.005436	president	0.004709	killed	0.002964	time	0.003025
states	0.005512	travel	0.005128	insurance	0.004633	share	0.002937	state	0.003016
trump2016	0.005504	sanctuary	0.004901	trumpss	0.004554	love	0.00292	super	0.002964
recount	0.0054	orders	0.004356	donald	0.004484	sign	0.002875	player	0.002936

the tweets on this topic, whereas the percentages for the appearance of these words in other topics could be extremely low. Thus, a sports-related theme could have phrases like “soccer” or “tennis” near the top of the list, but it could also contain words related to the U.S. government such as “Trump” or “Pentagon” toward the bottom. Therefore, only elements ranked highest using LDA were considered for measuring social peer effects.

5.4 Quality of the content

Quality is not objective and clearly depends on personal views. Therefore, the usage of the term here was limited to denoting correctly written tweets that did not exhibit formal or semantic inconsistencies. Because of the need to differentiate between members’ reactions driven by the intrinsic value of a tweet and those influenced by other users, it was necessary to formulate a quantifiable variable. A language-checking function was considered integral for this purpose and was applied in this solution. The implementation of this function entailed two key components. One of these was a database of high-value referential text, while the other was a spell-checking module for comparing the data sample with the referential text. By comparing the number of incorrect words with the total word count, as depicted in Fig. 8, a new variable, the user quality ratio (UQR), was formulated aimed at capturing the average value of a user’s tweets. When a similar procedure was executed with tweets that had been retweeted and replied to, the resulting variables were the retweet quality ratio (RTQR) and the reply quality ratio (REPQR). These variables indicated the user’s reaction threshold or the lowest content value that would still elicit a reaction from the user regarding a particular topic. Figure 9 provides a comparison of the variables UQR, RTQR, and REPQR. The following equation was used to calculate a user’s content quality threshold.

$$QT_u = UQR_u - Avg(RTQR_u, REPQR_u) \tag{4}$$

The procedure described above represents the least complicated method for establishing a baseline for a user’s reaction, which is essential for detecting any effects of social impact in cases where the user reacts to content that does not exceed the threshold. Considerable effort was invested in making this calculation as precise as possible.

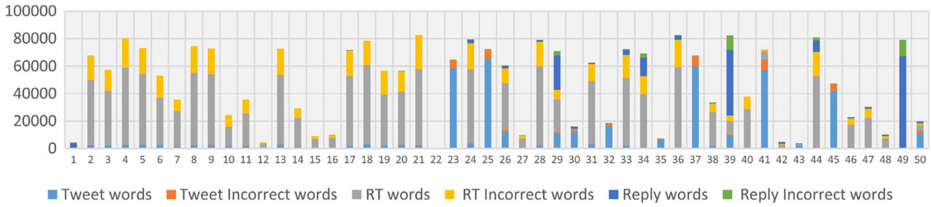


Fig. 8 Corrects words and incorrect words for the top 50 users ranked by their quality for the tweets, RT and replay actions

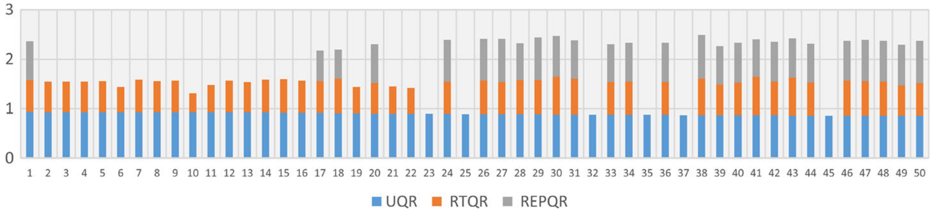


Fig. 9 User quality ratio vs. RT Quality ratio vs. Replay ratio for the Top 50 users

5.5 Peer effect

The last module for our solution was designed to calculate the propensity of each Twitter user to be impacted by the opinions of other users. Though this module was integral for realizing the study's original objective, its operation required flawless execution of all of the preceding modules. Consequently, improving the level of inputs for this module dramatically enhanced its outputs relating to the detection of social effects. The module was used to calculate the strength of peer effects based on the difference between the quality of a user's interactions on a given topic and his or her reaction threshold for the same topic. If the user had a high threshold and periodically exhibited low-quality interactions, this would imply that the external factor $\boldsymbol{x}$ was affecting this user's decision making. Thus, the rule expressed in Eq. 3 should be recalled.

Activity relating to tweets falling below the reaction limit was a clear indicator of an uncommon situation. The underlying premise of our solution centered on our interpretation of this discrepancy, namely that a statement could be updated to include the condition that $\boldsymbol{Q}+\boldsymbol{x}$ must be greater than $\boldsymbol{T}$, with $\boldsymbol{x}$ defined as the peer effect. Based on the performance of a full cycle of the required operations, we concluded that our solution was capable of determining the relation between social effects and the UQR variable and of expressing both of these quantities as percentages. This is demonstrated in Fig. 10.

6 Discussion and analysis

To confirm that our solution using the SOM was capable of delivering reliable results, we needed to scrutinize each phase of its implementation. Based on the findings of this study, we identified several areas requiring attention for the effective enhancement of the system's four modules. Whereas the first module relies on the algorithm expressed in Eq. 2, this is not the only means of calculating user influence. Ascertaining the strength of

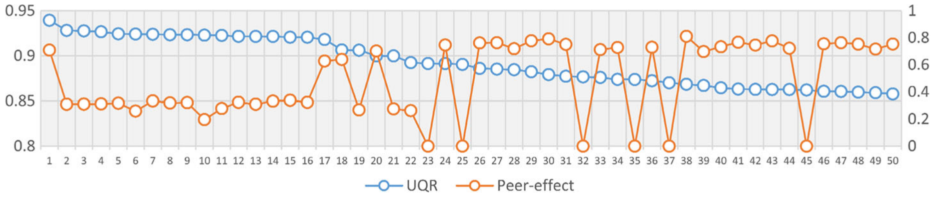


Fig. 10 User quality ration vs. peer-effect for the top 50 users

user impact beyond a reasonable doubt was not possible using the current solution, and we could only assume the precision of this variable. We successfully implemented the second module and demonstrated the solution's capacity to determine a user's interests. The key tool used for topic clustering was an LDA algorithm [2, 31, 38]. The procedure entailed an examination of twice the number of required topics to account for possible impurities within the sample and to avoid words that were not genuinely related to the topic. Using a shortened list, such words were assigned probabilities of association with the theme but were omitted from the final calculations.

The third module was used to determine content quality based on a selected feature as shown in Tables 3. This provided a limited view that could be expanded to include other factors and more flexible definitions of quality. The choice of the reference text for evaluating language is also crucially important. In this case, a compilation of 10,000 words was used that excluded a large number of informal expressions. The fourth and final model entailed considerable limitations, even though it successfully captured the relationship between influencing and influenced users. Its output was unidimensional, being based on a single indicator, which may not have fully captured social effects within the OSN.

7 Limitations and future work

7.1 Importing metadata about users

Twitter requires users to fill out fields in their profiles relating to their identities. Consequently, members' profile pages can offer valuable insights. Some of the key fields include sex, location, and age, but other demographic details could also be of interest. Given that competing OSNs also collect similar types of metadata, the solution could be adjusted to take advantage of this wealth of inputs, thereby increasing the precision of measurements of social pressure.

Table 3 List of features for measuring quality

Symbol	Description
Content-based features:	
GM	Contains grammar mistakes (Boolean)
SM	Contains spelling mistakes (Boolean)
#Char	Number of tweets' characters
#UChar	Number of upper case letters
#LChar	Number of lower case letters
#SChar	Number of special characters
#Words	Number of tweets' words
#Emo	Number of emoticons
Network-based features:	
#HT	Number of hashtags
#ME	Number of mentions
#URL	Number of URLs
#RT	Number of retweets
#RP	Number of replays

7.2 Developing a more precise quality scale

A critical issue that had a bearing on the overall model was the value assigned to each content piece. Because Twitter restricts communication to 140 characters per message, there is a limited amount of text available for conducting an analysis and searching for relevant information. Previous studies have shown that more than half of all tweets carry almost no semantic value for users beyond the immediate circle of the writer's closest confidants for whom the message was originally intended. Automatic determination of which tweets contain significant information of a more general nature is a difficult task that requires the utilization of various techniques. Tracking the total message length and the average number of characters per word as well as checking for the most frequently used terms and/or emojis are some of the methods that could be helpful in this regard. In messages that refer to themes related to economics or science, the presence of numerical elements and mathematical symbols could also indicate that the tweet has more than average value. As this SOM-based solution hinges on accurate estimation of each tweet's true value, the development of more advanced measuring systems for this purpose would be beneficial. Uniform tweet values across different thematic categories represent another limitation of the current solution. Thus, a significant upgrade of the model enabling the determination of separate values for each theme would be particularly advantageous.

7.3 Biographic details

Biographic details recorded for Twitter accounts can provide substantial information about particular members, as well as other similar members. This has significant implications for scientists as well as marketers, as both groups aim to predict the actions of network members [17, 23, 39, 40]. Some studies in this direction have already been conducted. Notably, Jennifer, Aron, and Nirmal were able to predict a user's biographical characteristics based on the structure of his or her connection matrix. The primary method applied in previous studies was monitored transfer in which each member was tagged with demographic descriptors. Multiple variables like age group, nationality, marital status, and educational level can be linked with various behavioral tendencies, enabling more accurate predictions of the future reactions of users across diverse contexts [8, 16].

7.4 Statistical breakdown of social impact

Various types of data are considered for calculating social impact, ranging from new messages, retweeted content from other users, hyperlinks and special characters, and the total size of users' contact networks [7]. Another relevant parameter is the relation between inbound and outgoing connections, which reveals whether a member is more of an influencer or a recipient. This abundance of native information enables the performance of an in-depth analysis of the impacts of mutual links between users. Consequently, we recommend the development of a specialized module for estimating the strength of social impacts. A specialized solution entailing a data mining approach would be considerably more reliable than the influence method that was used in this study, as it could be customized to address the most important concerns associated with the effect being investigated. It would be especially valuable to be able to differentiate between various topics when determining which users are capable of having the most impact within their networks.

7.5 Measuring social impacts

One factor that complicates the measurement of social effects is that every network user is simultaneously located at both ends of a communication stream. The same user could be the source of the social impact in one case, but in a different setting, he or she may be a subject who is influenced by others. All users should be initially tested as recipients, as this fosters a better understanding of their online behavior and contact networks. Only after this step has been performed should these users be treated as influential members in further operations. This is because it is crucial to possess as much background information about users as possible to measure their social influence. Identifying users with the greatest capacity to impact on their surroundings constitutes a further step that would enable the solution to be used on a larger dataset with more accurate and reliable results. For this reason, measuring the level of impact for all users remains one of the most sensitive aspects of the proposed solution.

8 Conclusion

The study was conceived as a complex effort that extended beyond simple information gathering to develop an innovative solution that addressed the issue of peer effects relating to Twitter users' opinions. The proposed solution entailed tracking the casual behavior of network members and transforming the data thus obtained into well-organized quantitative descriptions of particular users' online habits. This enabled the estimation of the most prominent areas of personal inquiry for every user as well as detection of the presence and strength of their social impacts. To accomplish the study's objectives, we consulted and applied the findings of numerous other studies wherever we deemed this appropriate and constructive.

Future work to advance this solution should be directed at enhancing the functionality of its individual modules. Because the system's outputs are contingent on the outcomes of each module, it is recommended that each module should be optimized with the overall objective of improving the predictive power of the solution. The Twitter platform can be helpful in this exercise, as a large number of parameters about members and their activities are tracked and could be utilized to develop a better understanding of how various users are linked to each other and how they affect the thinking of their immediate as well as more distant contacts. These parameters would provide an additional layer of data above verbal content, thus constituting a multidimensional approach for addressing this issue.

Acknowledgments The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through Vice Deanship of Scientific Research Chairs.

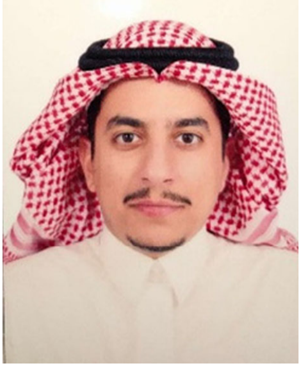
References

1. Abel F, Gao Q, Houben G-J, Tao K (2011) Semantic enrichment of twitter posts for user profile construction on the social web. In: Extended semantic web conference, pp 375–389
2. Alhamid, Mohammed F., Majdi Rawashdeh, Haiwei Dong, M. Anwar Hossain, and Abdulmotaleb El Saddik. Exploring latent preferences for context-aware personalized recommendation systems. IEEE Transactions on Human-Machine Systems 46(4):615–623

3. Al-Qurishi M, Aldrees R, AlRubaian M, Al-Rakhami M, Rahman SMM, Alamri A (2015) A new model for classifying social media users according to their behaviors. In: Web applications and networking (WSWAN), 2015 2nd world symposium on, pp 1–5
4. Benevenuto F, Rodrigues T, Cha M, Almeida V (2009) Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on internet measurement conference, pp 49–62
5. Besel C, Schlötterer J, Granitzer M (2016) On the quality of semantic interest profiles for onine social network consumers. ACM SIGAPP Appl Comput Rev 16:5–14
6. Besel C, Schlötterer J, Granitzer M (2016) Inferring semantic interest profiles from twitter followees: does twitter know better than your friends? In: Proceedings of the 31st Annual ACM symposium on applied computing pp 1152–1157
7. Chaudhary P et al (2016) XSS detection with automatic view isolation on online social network. Consumer electronics, 2016 I.E. 5th global conference on. IEEE
8. Chaudhary P et al (2017) A novel framework to alleviate dissemination of xss worms in online social network (osn) using view segregation. Neural Netw World 27(1):5
9. Chen M, Zhang Y, Li Y, Mao S, Leung V (2015) EMC: emotion-aware mobile cloud computing in 5G. IEEE Netw 29(2):32–38
10. Chen M, Ma Y, Hao Y, Li Y, Wu D, Zhang Y, Song E (2016) CP-robot: cloud-assisted pillow robot for emotion sensing and interaction. Industrialiot 2016, Guangzhou, China
11. Chu Z, Gianvecchio S, Wang H, Jajodia S (2012) Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 9:811–824
12. Dougnon RY, Fournier-Viger P, Nkambou R (2015) Inferring user profiles in online social networks using a partial social graph. In: Canadian Conference on Artificial Intelligence, pp 84–99
13. Eason G, Noble B, Sneddon IN (1955) On certain integrals of Lipschitz-Hankel type involving products of Bessel functions. Philos Trans R Soc Lond A247:529–551 references
14. Fang Q, Xu C, Sang J, Hossain MS, Muhammad G (2015) Word-of-mouth understanding: entity-centric multimodal aspect-opinion mining in social media. IEEE Trans Multimed 17:2281–2296
15. Fang Q, Sang J, Xu C, Shamim Hossain M (2015) Relational user attribute inference in social media. IEEE Trans Multimed 17(7):1031–1044
16. Gupta BB, Agrawal DP (2016) Shingo Yamaguchi, "handbook of research on modern cryptographic solutions for computer and cyber security," IGI global. Publisher, USA
17. Hossain MS, El Saddik A (2008) A biologically-inspired multimedia content repurposing system in heterogeneous network environments. ACM/Springer Multimed Syst J 14(3):135–143
18. Hossain MS, Muhammad G, Al Hamid MF, Song B (2016) Audio-visual emotion-aware big data recognition towards 5G. Mob Netw Appl 21(5):753–763
19. Hossain MS, Alhamid MF, Muhammad G (2017) Collaborative analysis model for trending images on social networks. Futur Gener Comput Syst
20. Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, pp 56–65
21. Jiang J, Wilson C, Wang X, Sha W, Huang P, Dai Y et al (2013) Understanding latent interactions in online social networks. ACM Trans Web (TWEB) 7:18
22. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, pp 591–600
23. Li J, Yan H et al Location-sharing systems with enhanced privacy in mobile online social networks. IEEE Syst J. https://doi.org/10.1109/JSYST.2015.2415835
24. Lo SL, Chiong R, Cornforth D (2016) Ranking of high-value social audiences on twitter. Decis Support Syst 85:34–48
25. Miller JH, Page SE (2004) The standing ovation problem. Complexity 9:8–16
26. Min W, Bao B-K, Xu C, Hossain MS (2015) Cross-platform multi-modal topic modeling for personalized inter-platform recommendation. IEEE Trans Multimed 17:1787–1801
27. Peng, Min, Jiajia Huang, Hui Fu, Jiahui Zhu, Li Zhou, Yanxiang He, Fei Li (2013) High quality microblog extraction based on multiple features fusion and time-frequency transformation. In International Conference on Web Information Systems Engineering, pp. 188–201. Springer, Berlin, Heidelberg
28. Peng M, Gao B, Zhu J, Huang J, Yuan M, Li F (2016) High quality information extraction and query-oriented summarization for automatic query-reply in social network. Expert Systems with Applications 44:92–101
29. Pennacchiotti M, Popescu A-M (2011) Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 430–438

30. Qian S, Zhang T, Xu C, Hossain MS (2015) Social event classification via boosted multimodal supervised latent dirichlet allocation. ACM Trans Multimed Comput Commun Appl (TOMM):11–27
31. Rawashdeh, Majdi, Mohammad Shorfuzzaman, Abdel Monim Artoli, M. Shamim Hossain, and Ahmed Ghoneim (2017) Mining tag-clouds to improve social media recommendation. Multimed Tools and Appl 76(20) 21157–21170
32. Riquelme F, González-Cantergiani P (2016) Measuring user influence on Twitter: A survey. Inf Process & Manag 52:949–975
33. Song J, Zhang Y, Duan K, Hossain MS, Rahman SMM (2016) TOLA: topic-oriented learning assistance based on cyber-physical system and big data. Futur Gener Comput Syst 75(2017):200–205
34. Tao, Ke, Fabian Abel, Qi Gao, and Geert-Jan Houben. "TUMS: twitter-based user modeling service." In Extended Semantic Web Conference, pp. 269–283. Springer, Berlin, Heidelberg, 2011
35. Vosecky, Jan, Kenneth Wai-Ting Leung, Wilfred Ng (2012) Searching for Quality Microblog Posts: Filtering and Ranking Based on Content Analysis and Implicit Links. In DASFAA 1:397–413
36. Yang T, Lee D, Yan S (2013) Steeler nation, 12th man, and boo birds: classifying twitter user interests using time series. In: Advances in social networks analysis and mining (ASONAM), 2013 IEEE/ACM international conference on, pp 684–691
37. Yang Z, Wilson C, Wang X, Gao T, Zhao BY, Dai Y (2014) Uncovering social network sybils in the wild. ACM Trans Knowl Discov Data (TKDD) 8:2
38. Yang X et al (2015) Automatic visual concept learning for social event understanding. IEEE Trans Multimed 17(3):346–358
39. Zhang Z et al (2016) CyVOD: a novel trinity multimedia social network scheme (MTAP-D-16-01532). MTAP. Springer, New York
40. Zhang Z et al (2016) Social media security and trustworthiness: overview and new direction. Futur Gener Comput Syst 2016
41. Zuber M (2014) A survey of data mining techniques for social network analysis. Int J Res Comput Eng Electron 3(6):1–8

Muhammad Al-Qurishi is a Ph.D. candidate in the Information Systems Department in the College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia. He received his master's degree in information systems from King Saud University, Kingdom of Saudi Arabia. He has published several papers in refereed IEEE/ACM/Springer journals and conferences. His research interests include online social networks, social media analysis and mining, human-computer interaction, and health technology



Saad Alhuzami is a Masters student in the Information Systems Department in the College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia.

Majed AlRubaian is a Ph.D. candidate in the Information Systems Department in the College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia. He received his master's degree in information systems from King Saud University, Kingdom of Saudi Arabia. He has authored several papers in refereed IEEE/ ACM/ Springer journals and conferences. He is a student member of the ACM and the IEEE. His research interests include social media analysis, data analytics and mining, social computing, information credibility, and cyber security.

M. Shamim Hossain is a Professor at the King Saud University, Riyadh, KSA. Dr. Shamim Hossain received his Ph.D. in Electrical and Computer Engineering from the University of Ottawa, Canada. His research interests include serious games, social media, IoT, cloud and multimedia for healthcare, smart health, and resource provisioning for big data processing on media clouds. He has authored and coauthored around 150 publications including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, books, and book chapters. He has served as a member of the organizing and technical committees of several international conferences and workshops. He has served as co-chair, general chair, workshop chair, publication chair, and TPC for over 12 IEEE and ACM conferences and workshops. Currently, he serves as a co-chair of the IEEE ICME workshop on Multimedia Services and Tools for smart-health MUST-SH 2018. He is a recipient of a number of awards including, the Best Conference Paper Award, the 2016 ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) Nicolas D. Georganas Best Paper Award, and the Research in Excellence Award from King Saud University. He is on the editorial board of IEEE Access, Computers and Electrical Engineering (Elsevier), Games for Health Journal and International Journal of Multimedia Tools and Applications (Springer). Previously, he served as a guest editor of IEEE Transactions on Information Technology in Biomedicine (currently JBHI), International Journal of Multimedia Tools and Applications (Springer), Cluster Computing (Springer), Future Generation Computer Systems (Elsevier), Computers and Electrical Engineering (Elsevier), and International Journal of Distributed Sensor Networks. Currently, he serves as a lead guest editor of IEEE Communication Magazine, IEEE Transactions on Cloud Computing, IEEE Access and Sensors (MDPI). Dr. Shamim is a Senior Member of IEEE, a member of ACM and ACM SIGMM.

Atif Alamri is an Associate Professor in the Information Systems Department at the College of Computer and Information Sciences, King Saud University. Riyadh, Saudi Arabia. His research interests include multimedia-assisted health systems, ambient intelligence, and service-oriented architecture. Dr. Alamri was Guest Associate Editor of the IEEE Transactions on Instrumentation and Measurement, a co-chair of the first IEEE International Workshop on Multimedia Services and Technologies for E-health, a technical program co-chair for the 10th IEEE International Symposium on Haptic Audio Visual Environments and Games, and serves as a program committee member for many conferences in multimedia, virtual environments, and medical applications.

Md. Abdur Rahman is an Assistant Professor in the Department of Computer Science, Prince Muqrin University, Madinah Al Munawwarah, Kingdom of Saudi Arabia. Dr. Abdur Rahman received his Ph.D. degree in Electrical and Computer Engineering from the University of Ottawa, Canada in 2011. His research interests include serious games, cloud and multimedia for healthcare, multimedia big data, and next generation media. He has authored and co-authored around 85 publications including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, and book chapters. He has 6 US patent issued and couple of pending. He has served as a member of the organizing and technical committees of several international conferences and workshops. Recently, he received three best paper awards from ACM and IEEE Conferences. Dr. Abdur Rahman is a member of both IEEE and ACM.