Selection of Community Detection Features Influencing Negative Emotional Contagion on Twitter

Hatoon Al Sagri[a,b], Mourad Ykhlef [b]

[a]Information Systems Department College of Computer and Information Sciences Al Imam Mohammad Ibn Saud Islamic University(IMSIU) Riyadh, Saudi Arabia
[b]Information Systems Department College of Computer and Information Sciences King Saud University Riyadh, Saudi Arabia

Summary
Negative emotional contagion is spreading in social networks and is adversely affecting people; it can even lead to depression and suicide. By implementing the genetic algorithm along with community detection algorithms, this article aims to uncover the Twitter features that enhance the spread of negative emotions on the network. The novelty of this article is that it focuses on a combination of community detection features that enhance the diffusion of negative emotions in social networks. The genetic algorithm benefits the study as it uses the modularity of the community detection algorithms (Louvain and Label Propagation algorithms) as fitness values in order to find the most favorable values in cost-effective manner. While other studies have concentrated on singular Twitter features, applying the Louvain genetic algorithm and the Label Propagation genetic algorithm to negative emotion data from Twitter resulted in higher modularity from a combination of features than the results from single features. This demonstrates that a combination of features increases the diffusion of negative emotions in Twitter communities.

Key words:
Negative emotions, contagion, Twitter, feature selection, community detection, genetic algorithm.

1. Introduction

Ideas and information spread in social networks like a pathogen: each infected person affects his/her friends [2]. Studies have found that emotions, including happiness and depression, can be contagious both in person and online [3-5].

Recently, the Arab Spring and Occupy movements have demonstrated the powerful influence of social media [3, 6]. In a study that compared people's beliefs, Soliman et al. [6] reported that large networking sites such as Facebook and Twitter can have a significant psychological impact on our behavior, and people's opinions are currently changing for the worse [3]. Naveed et al.'s [7] analysis of tweets indicated that tweets containing negative emoticons are more likely to be retweeted than tweets with positive emoticons. Additionally, Sobkowicz and Sobkowicz [8] and Chmiel et al. [9] studied emotional expression in blogs and found that negative emotions drive interaction in users' communities [2]. In response, our study focused on the negative emotional contagion spreading in online social networks, which affect individuals who have negative emotions diffused within their communities.

The popularity of social networks increases the importance of community detection, which is useful for capturing valuable metadata about large-scale networks [2]. Community detection is important for the easy visualization of networks, including network structure and relationships [10, 11]. However, community detection is challenging, particularly because the number of communities is unknown and communities vary widely in size [3, 11]. Another challenge in community detection is evaluating the quality of community detection algorithms; many studies implement Newman's modularity to assess the quality of the results [12]. In this work, we use and compare two community detection algorithms, the Louvain algorithm (modularity-based) and the Label Propagation algorithm (diffusion-based), to emphasize communities in which negative emotions are spreading.

Twitter has a number of features, including retweets, replies, mentions, and URLs. In this study, we focus on retweets as a diffusion indicator [13] and on mentions, hashtags, follows, locations and languages as user features. These features have been specifically chosen to indicate link similarity, which is a better indicator of similarity between social network users than content similarity [10].

This article is interested in studying the emotions in tweets, specifically tweets that express negative emotions including depression and/or anger. Ekman's basic emotions are commonly used for emotion mining and classification [14]. Also, Parrott's emotion framework classifies emotions as primary, secondary, and tertiary [15]. Accordingly, for this study, we chose to gather data on the most common emotions on Twitter.

The number of features chosen for this study as well as the large amount of data crawled from Twitter required the use of a genetic algorithm to find the good (may be optimum) solution in polynomial time. In this article, we enhanced Louvain and Label Propagation algorithms by genetic approach yielding Louvain genetic algorithm (LGA) and the Label Propagation genetic algorithm (LPGA) which improved the community detection of negative emotions on Twitter. These algorithms helped find the features that enhance the diffusion of negative emotions in the network.

Manuscript received September 5, 2019
Manuscript revised September 20, 2019

The results of the study showed that the modularity was similar for both LGA and LPGA, but the Rand index showed higher accuracy in the LGA results than in the LPGA results. Both algorithms revealed that analyzing a combination of features produced higher results than the analysis of a single feature, demonstrating that social graphs with more than one feature influence the spread of negative emotions.

The contributions of this article can be summarized as follows:

1. We used the genetic algorithm along with community detection algorithms to enhance the computational time and reach a good solution. It also increases the efficiency of the proposed algorithm when used with huge number of features.
2. We tested a combination of Twitter features to determine the effect they had on the diffusion of negative emotions in the community. In contrast, previous studies on Twitter community detection concentrated on a single feature, such as retweets or mentions, as the relationship indicator.

The rest of the article is organized as follows: Section 2 reviews related works. Section 3 presents the Louvain and Label Propagation community detection algorithms used in the experiments. Section 4 explains the proposed community detection algorithms that are enhanced by the inclusion of multiple features. Section 5 describes and discusses the experiments and the results. Finally, section 6 outlines the conclusions of the study.

2. Related Work

Twitter and other social networks continue to gain popularity and have an increasing impact on users' emotions and the diffusion of these emotions. Happiness and other emotions have recently caught the attention of researchers in many fields, including psychology, economics, and neuroscience [4], [16-19]. Large-scale emotional contagion is occurring in online networks, and these emotions can spread worldwide in a single day [2, 19]. Lately, many researchers have attempted to detect and model emotions in social networks. Their findings have been made based on the observation that emotion can be propagated via online interactions [20, 21]. Councill et al [22] proposed a method to improve sentiment classification by detecting negation in sentences using a conditional random field. Cole et al. [23] presented a method of classifying novel variants of a diffusion model to predict the emotions of a large set of blog entries based on the emotions expressed in entries written by users' friends [23]. Moreover, Rosenquist et al. [24] studied depression in social networks and found its effects can extend up to three degrees of separation (to ones friends' friends' friends).

The findings of current viral marketing research studies show that emotions play a critical role in the spread of content online [3, 25]. Kanavos et al. [26] investigated the impact of a tweet's emotional content on its diffusion through retweets; they found that tweets containing negative emotions were more likely to be contagious.

Community detection "formalizes the strong social groups based on the social network properties." [27]. In an online social network, community detection is "based on analyzing the structure of the network and finding individuals who correlate more with each other than with other users." [28]. Lerman and Gosh's [29] study on Digg and Twitter showed that the structure of a network affects the dynamics of information spread. Also, Paranyushkin [30] and Ball [31] reported that community structure is essential for information contagion to propagate through the network. Moreover, Deitrick et al. [11] stated that combining sentiment analysis with community detection results helps to illuminate the sentiment that is expressed by different communities easier [2].

It is important to apply the appropriate tools in detecting and understanding the behaviors of network communities to be able to model the dynamism of the domain to which they belong [32, 33]. Different clustering techniques have been applied to detect communities in online social networks [27, 32, 34]. Clustering individuals into groups that share common characteristics helps to assess individual behaviors and what activities, goods, and services an individual is interested in [28]. Rani and Goyal [35] reported that the application of clustering techniques is used by many researchers to improve the performance of information retrieval. Furthermore, Kim et al. [36] found that dividing and clustering the Twitter network helps individuals to find a group of users with a similar inclination, which is called a "community" [2].

Dividing graphs into homogeneous, densely interconnected clusters of nodes with minimal connection between clusters has been a topic of community discovery [2, 27, 34, 37-40]. Diverse clustering techniques have been applied by different researchers to detect communities in social networks [27, 32, 34]. Shalizi and Thomas [37] thought that it was ideal to first establish the existence of these clusters, to note the memberships of each individual in the chosen model and to control this when looking for evidence of contagion or influence.

Rosenquist et al. [24] study of the factors of depression in social networks used clustering. The network was clustered according to the level of depression, such as moderately depressed or very depressed, and according to the relationships between individuals, such as friends, siblings or spouses. The study proved that emotion propagate from person to the other and that it depends on social ties between the nodes and where they are located in the network.

Zhu et al. [41] used Louvain algorithm [1] and CNM [42] algorithm to detect emotional communities where they

proofed that the emotional network is more suitable for detecting emotional communities. Also, Xiong et al. [21] indicated that users in a small community are more likely to change the mood for the influence of community leader.
Cha et al. [43] used indegree, retweets, and mentions as measures of influence on Twitter and concluded that users who are mentioned frequently are also retweeted frequently and vice versa. Kim et al. [44] indicated that tweets and retweets help diffuse information on Twitter and that the analysis of information diffusion help understand overall social behaviors among the users in a social network. Yang et al. [45] found Twitter interactions, specifically mentions, were strong predictors of information diffusion. In addition, Tareaf et al. [46] analyzed how quickly information spread on Twitter based on the influence of friend relationships.

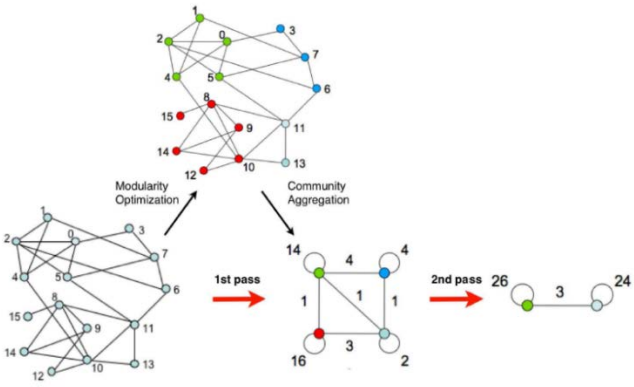

Fig. 1 Visualization of Louvain algorithm steps [1]

After a thorough review of related studies, we predicted the Twitter community structure influenced the diffusion of negative emotional contagion. Accordingly, we found that Twitter features (either one feature or a combination of features) affect the diffusion of tweets and increase the spread of negative emotions.

3. Community Detection

A widespread informal definition of community is a group of nodes that are densely interconnected [27, 47, 48]. Community detection is a key characteristic used to extract useful information from networks [49]. One of the greatest challenges in community detection is evaluating the quality of community detection algorithms, so many studies use Newman's modularity to assess the quality of the results [12]. On the other hand, the Rand Index is used to measure the similarity between two data clusters.

3.1 Community Detection Algorithms

There are numerous community detection algorithms, and they translate and combine aspects of cohesion and separation differently.
For this study, we selected community detection algorithms that assisted in the study of the diffusion of negative emotions. First, we chose the Label propagation algorithm, a diffusion-based community detection algorithm that detects communities by considering how information is propagated in a network. We also selected the Louvain algorithm, which is an efficient modularity-based community detection algorithm that assesses cohesion and separation through the number of intra-and intercommunity links, respectively [50, 51].
Louvain Algorithm: This method is an agglomerative hierarchical modularity-based community detection algorithm introduced by Blondel et al. [1] that relies on a greedy optimization process, and includes an additional aggregation step to improve processing in large networks [48]. This algorithm is fast, allowing for the analysis of huge networks with billions of edges [52], and produces significant partitions [53], which has made it extremely popular in recent years.
The algorithm starts with placing each node in its own community. The modularity gain is calculated for each node that is moved to its neighbors' community, where the node will be moved to the community with the largest gain or stay in its community if no gain is possible [48]. The procedure is repeated for all nodes until no further improvement is applicable, thus ending the first step. For the second step, a new network is built where nodes are the communities estimated during the first step, and community links are represented in the new network by weighted regular links and self-loops. Later, the first step is applied to the new network, and both steps are repeated until stable communities are reached [48] figure (1).
High modularity indicates strong ties between nodes in the community. This shows higher diffusion in negative emotions because it depends on the edge weights in calculating clusters, which serves our study since the edge weight reflects the diffusion of the negative emotions This algorithm has been proven to produce good community structures. Its complexity is expected to be O(n log n), but precise complexity analysis is still lacking due to the difficulty of describing the number of corrections in advance [54, 55].
Label Propagation Algorithm: This algorithm, by Raghavan et al. [56], uses the concept of the node's neighborhood and the diffusion of information in the network to identify communities [12]. Initially, each node in the network is assigned a unique label; then, through an iterative process, each node is updated with the label held by the majority of its neighbors [12, 57]. This process continues until one of several conditions is met, such as no label change. The

resulting communities are defined by the last assigned label values figure(2).

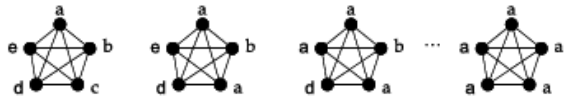

Fig. 2 Nodes are updated one by one as we move from left to right [56].

This algorithm is desirable because it is quick, effective, and easy to implement [48, 57]. The overall complexity of this method depends on the number of iterations (taking the linear time of O (m)); this number cannot be estimated but has been observed to stay relatively low in many examples [54, 55]. Papadopoulos et al. [55] pointed out that the computational efficiency and conceptual simplicity of the Label Propagation algorithm facilitates the development of method extensions or adaptations that cater to particular problems.

Similar to the Louvain algorithm, the Label Propagation algorithm is applied to data generated from the genetic algorithm and the modularity is calculated for the specific combination of features. High modularity indicates strong ties between the nodes in the community and shows a higher level of diffusion in the network.

3.2 Evaluation Metrics

Modularity: The most widely used and accepted metric designed specifically for the purpose of measuring the quality of a network's division into communities is modularity (Q) [38, 58]. Modularity compares the density of links inside communities to the links between communities [34, 50].

Thus, Q effectively measures the fraction of edges in the network that connect nodes in the same community minus the expected value of this quantity if the edges were placed at random. The value ranges from Q = 0, when the within-community edges are no better than random, to Q = 1 [58]. The modularity Q is defined as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \qquad (1)$$

Rand index: The Rand Index [59] corresponds to the proportion of node pairs for which both the estimated and reference community structures agree. A pair of nodes is considered correct if the nodes either share the same cluster in both clusterings or if they are in different clusters in both clusterings. The fraction of pairs of nodes that are correct is the Rand Index. Let a11 be the number of pairs of nodes that are in the same cluster in both clusterings. Let a00 be the number of pairs of nodes that are in different clusters in both clusterings. Let a01 and a10 be the nodes that are in the same cluster in one clustering and in different clusters in the other. Thus, then the Rand Index is as follows:

$$\frac{a_{11}+a_{00}}{a_{11}+a_{10}+a_{01}+a_{00}} \qquad (2)$$

The Rand Index ranges from 0 (the algorithm completely failed to estimate the community structure) to 1 (the algorithm perfectly estimated the community structure).

3.3 Testing

We used of a pair of publicly available Twitter datasets to test the community detection algorithms using evaluation metrics; this showed that the algorithm did not fit the data [60]. The datasets were chosen according to their similarity to our data and included retweet views. Four datasets, the Olympics, Political (UK), Political (IE), and Rugby, contained weighted, directed retweets of users in the specified domain.

- Olympics This dataset consisted of 464 users and covered athletes and organizations involved in the London 2012 Summer Olympics. The disjointed ground-truth communities corresponded to 28 different sports [60].
- Political (UK) This dataset included 419 Members of Parliament (MPs) in the United Kingdom. The ground-truth communities consisted of five groups corresponding to specific political parties [61].
- Political (IE) This dataset was a collection of Irish politicians and political organizations assigned to seven disjointed ground-truth groups by party affiliation [60].
- Rugby This dataset was a collection of 854 international Rugby Union players, clubs, and organizations that were active on Twitter [62]. The ground-truth groups consists of overlapping communities corresponding to 15 countries. The user accounts of rugby players can potentially be assigned to both their home nation and the nation in which they play rugby [60].
- Negative Emotions This dataset contained the data collected for our study and also contained retweets containing negative emotions that were passed between users. (Explained in detail in section 5.)

Table 1 shows that the modularity for the Political, Olympics, and Rugby datasets are higher with the Louvain algorithm than the Label Propagation algorithm, which indicates that the nodes in the communities are more connected and have stronger ties. On the other hand, the Rand index was higher for Label Propagation than Louvain, thus indicating a higher accuracy in the partitioning of communities. Although the Rand index showed the opposite, the modularity is considered a stronger evaluation metric for the formation of communities. On the other hand, the Negative emotions data showed high modularity in both the

Louvain and Label Propagation, but the Rand index was higher in the Louvain results, thus emphasizing more accuracy in the construction of communities. The number of nodes in each community of the Negative emotions dataset was small, while the number of communities was large, because the relations between the nodes depended on the diffusion of the emotions [63]. While the number of nodes in each community was high, the number of communities was small in the other datasets. As a result, the modularity of the Negative emotion data was higher than the modularity of the other datasets. These results guided us to use both the Louvain Community Detection algorithm and the Label Propagation algorithm with the Genetic Algorithm to the find an optimal solution, which, in our situation, is the combination of features with stronger relations and higher diffusion of negative emotions.

Table 1: Evaluation Metric Results for Community Detection Algorithms of the Different Datasets

Datasets	Evaluation Metrics	Community Detection Algorithms	
		Louvain algorithm	Label Propagation Algorithm
Political (UK)	Modularity Rand Index	0.4487 0.684	0.3459 0.995
Olympics	Modularity Rand Index	0.8534 0.642	0.582 0.9838
Political (IE)	Modularity Rand Index	0.5499 0.8032	0.5071 0.9892
Rugby	Modularity Rand Index	0.6203 0.6116	0.5235 0.9915

4. Proposed Feature Selection Community Detection Algorithms

4.1 Feature Selection

Twitter (http://twitter.com) is a popular social networking site that allows registered users to post and read short text messages called tweets [29]. Twitter has features, rules, and regulations to ensure a positive user experience [64]. A user can retweet or comment on another user's post, usually preceded by "RT @x," where x is the name of the user being retweeted. Twitter also allows users to designate other users they want to follow. Twitter

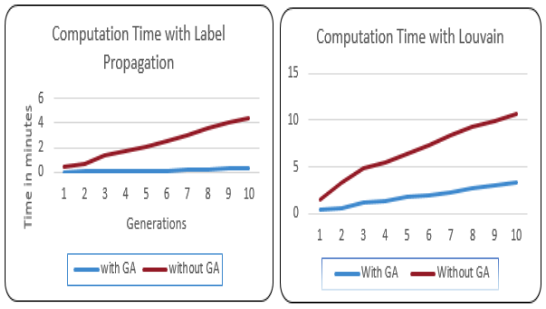

Fig. 3 The Computation Time of Data with and without Genetic Algorithm

restricts large-scale access to its data to a limited number of entities [29]. We used Twitter4J to retrieve tweets using keywords such as sad, upset, and angry to find tweets containing negative emotions; however, Twitter limits the retrieval of tweets to 3000 tweets every 15 minutes. For each tweet, we retrieved the account of the user who wrote the tweet and the account of users who retweeted it. Then, we crawled account features such as follows, mentions, language, location, and hashtags, and the data were stored in a database.

Twitter has many features such as replies, mentions, and followers that may affect the diffusion of negative emotions. Five commonly studied features proven to affect diffusion were selected for the study. Cha et al. [43] studied followers, mentions, and retweets and found retweets have the most influence on diffusion. Zhu et al. [65] also indicated that retweeting is the best way to spread information on Twitter. Therefore, our study depends on retweets for identifying the diffusion of negative emotions on Twitter. Del Vicario et al. [20] proved that active users shift more quickly to negativity than less active users. For this reason, we are concerned with interaction features that showed users' activity levels and their interactions in the network. We focused on an individual's potential to engage others in a certain act by concentrating on five popular activities on Twitter, and these activities were considered as features of the network communities. These features were chosen to show link similarity, which is a better indicator of similarity between users than content similarity [10] The five features are as follows:

- Mention: A mention is when a user comments on another user's tweet using the following tag: @username.
- Following: Another user account following the user who wrote the tweet.
- Hashtag: Hashtags are words or phrases preceded by a hash sign (#). Users can place this sign in tweets to identify messages on a specific topic
- Location: A user's location or place.
- Language: The first language a user employs when tweeting.

Community features that affect the diffusion of negative emotions in the network may be limited to one feature or a combination of features. According to the following calculation, number of combinations can be created from the initial features [40]:

$$C(n,r) = {}^{n}C_r = {}_{n}C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (3)$$

C(n,r) is a combination of features where n is the number of features and r is the number of initially chosen features. The community detection algorithm we ran had exponential computational time. The use of the genetic algorithm helped

us choose a useful set of features in polynomial time figure (3).

4.2 Genetic Algorithm

Since we aimed to find the features that affected emotional diffusion the most in polynomial time, we needed to decrease unnecessary features to reduce the computation time. The genetic algorithm is best used to find the most favourable solution in polynomial computation time; the solution might be optimal but it is not guaranteed. The algorithm is a randomized adaptive search and optimization method inspired by the natural gene selection process and is based on Darwin's principle of the survival of the fittest [66]. The genetic algorithm finds the most favourable solutions in a quick and inexpensive manner [67].

In this study, the genetic algorithm was used to select the best combination of features in polynomial computational time. The initial population of genetic algorithm is generated from the five features, and individuals with the highest fitness values are randomly combined to generate the next generation. Individuals are selected according to the fitness values, then crossover and mutation are performed. The crossover operator produces new individuals from two selected individual used as parents, by swapping segments to produce new individuals. The mutation operator is used to maintain diversity. During the mutation phase value of each segment in each selected individual is changed. Better solutions are created by repeating this process for many generations.

For feature selection, we used a string consisting of binary digits (i.e., 0s and 1s) , which denoted a solution [67]. The length of the string corresponded with the total number of features. In a binary string, the exclusion of a feature was indicated by a 0 in a particular position while a 1 in a particular position designated inclusion in the feature subset [67]. For example, a candidate solution with the binary representation 11001 indicates that there are a total of five features, and the feature subset contains three features (1, 2, and 5) Table 2. Data were later retrieved to generate graphs according to the selected features. Each graph represented a feature or combination of features and was used for the proposed community detection algorithms.

Table 2: Individuals (Features) Representation

Mention	Hashtag	Location	Language	Following
1	1	0	0	1

4.3 Louvain Genetic Algorithm (LGA) and Label Propagation Genetic Algorithm (LPGA)

In this section we will enhance Louvain algorithm and Label Propagation algorithm with genetic algorithm. The two enhanced algorithms are named LGA and LPGA. The LGA and the LPGA were used to calculate the fitness values from the modularity of the Louvain algorithm and Label Propagation algorithm, and the genetic algorithm was used to find the most favourable solution quickly and inexpensively [67].

The data used for the experiment has been crawled and organized in a database for easy extraction and generation of graphs. For both algorithms, the genetic algorithm generates a new individual that is consisted of combination of features (Table 2) in each population, crossover and mutation are performed on each individual to generate new individuals and accordingly the graph is extracted from the database according to the features of each individual. The graph is later used for the Louvain and Label

Propagation algorithms, and the modularity of each algorithm is calculated. Then, the next generation in the genetic algorithm is generated for a new combination of features, and each modularity is recalculated. The process continues until the condition is reached. The results shows the most optimal feature or combination of features with the highest modularity for both algorithms. The details of the LGA and LPGA steps are shown in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 LGA Pseudo-code

```
Generation = 0
Initialize population of individuals (features)
While Generation < MaxGeneration do
Evaluate Fitness of individuals
Repeat
    for each individual generate network G
    Repeat
        Put each node of G in its own community
        while some nodes are moved do
            for all node n of g do
                place n in its neighbouring community including its own which maximizes the modularity gain
            end for
        end while
        if the new modularity is higher than the initial then
            G = the network between communities of G
        end if
    until no further movements of nodes
Until BestFitness < MaxFitness
Select the best-fit individual for reproduction
Breed new individual through crossover and mutation operations
Replace least-fit population with new individuals
Increment Generation
End while
```

Algorithm 2 LPGA Pseudo-code

```
Generation = 0
Initialize population of individuals (features)
While Generation < MaxGeneration do
Evaluate Fitness of individuals
Repeat
    for each individual generate network G
    Initialize the labels for all nodes
        while not converged and num_iterations <
                max_iterations do
        Arrange the nodes in the network in random
    order
        Assign label with highest frequency among
    neighbors
     end while
Until BestFitness < MaxFitness
Select the best-fit individual for reproduction
Breed new individual through crossover and
    mutation operations
Replace least-fit population with new individuals
Increment Generation
End while
```

5. Experiments and Results

5.1 Data Description

The datasets used for this study consisted of data down-loaded directly from the Twitter API. The tweets containing negative emotions were crawled using Twitter4J. The negative emotion words were selected according to Parrot's emotions framework [15, 68] and Ekman's scale [14, 69]. Ekman's scale classified emotions to six basic emotions (Anger, Disgust, Fear, Joy, Sadness, and Surprise) that are commonly used for emotion mining and classification [14]. Also, Parrott emotion framework has classified emotions into primary, secondary and tertiary emotions [15]. Disappointment is secondary emotions that is classified from the primary emotion sadness, while depression is tertiary emotion classified from secondary emotion sadness. Accordingly, most commonly used emotions in Twitter have been chosen for crawling data needed for the study.

We used Twitter's search API to retrieve all English tweets containing words such as sad, upset, disappointed, angry, frustrated, and sad emoticons. We crawled 751 tweets and 529K users who have initiated, retweeted or mentioned the tweets. We also collected the retweets of each tweet, the hashtags and the mentions that each tweet contains, and other information of the users, such as followers, locations, and languages. Although tweets collected were in English only, some accounts showed that English was not their first language, 17 different languages were found for accounts who tweeted or retweeted English tweets. The search depended on the retweets as an indication of diffusion existence in the network. The data collected were uploaded in the database. The retrieval of the data from the database to generate a weighted and directed graph (G) was conducted according to the similarities between the users. Retrieval depended on 5 features.

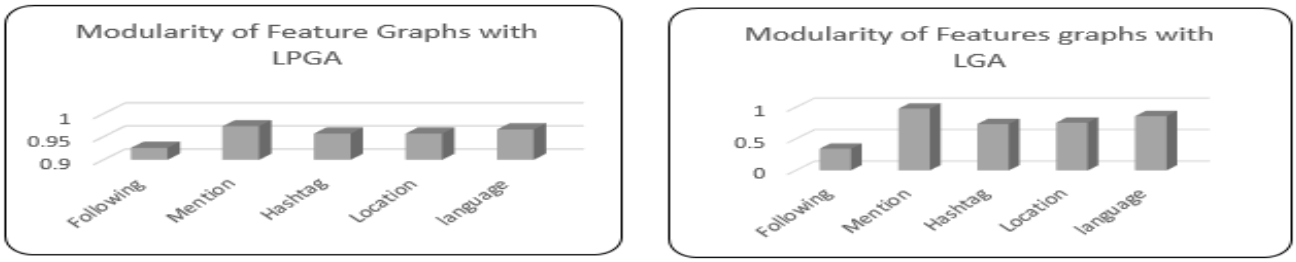

Fig. 4 Modularity of Feature Graphs with Community Detection Algorithms

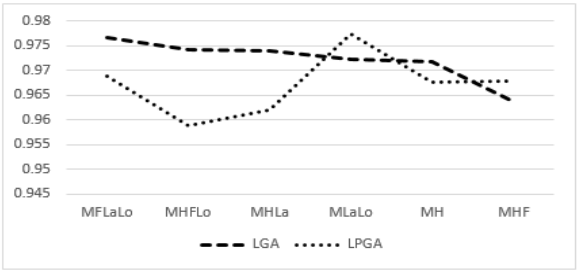

Fig. 5 Comparison of Modularity Measure

- Mention (M) Graph: Take a pair (u,v) where u,v ∈ G; Edge E exists between (u,v) if u mentions v. The edge is given a weight w, where w(u,v)= the number of times a node mentions the other.
- Hashtag (H) Graph: Take a pair (u,v) where u,v ∈ G; Edge E exists between (u,v) if u and v have the same hashtag. The edge is given a weight w, where w(u,v)= the number of times that nodes had similar hashtags.
- Location (Lo) Graph: Take a pair (u,v) where u,v ∈ G; Edge E exists between (u,v) if u and v have the same location. The edge is given a weight w, where w(u,v)= the number of times users retweeted each other while in the same location.
- Language (La) Graph: Take a pair (u,v) where u,v ∈ G; Edge E exists between (u,v) if u and v have the same language. The edge is given a weight w, where w(u,v)= the number of times that users retweeted each other when the accounts had the same first language.
- Following (F) Graph: Take a pair (u,v); u where v ∈ G; Edge E exists between (u,v) if u and v follow similar accounts. The edge is given a weight w, where w(u,v)= the number of followers that users have in common.

5.2 Experiment Results

In this section, we report the results of applying the LGA and LPGA to detect the features of communities that increase the diffusion of negative emotions in the network (Twitter). Applying the LGA to test Twitter's features (mentions, hashtags, locations, followers, and languages) found that mentions had the highest modularity (0.773) and language had the second-highest modularity (0.528) (Figure 4). However, combining mentions, followers, locations, and languages provided the highest modularity (0.976). If two accounts mention each other, follow similar accounts, are in the same location, and have the same first language, there is a higher possibility they retweet each other and negative emotions spread between them easily. The experiment results showed that the combination of mentions, hashtags, followers, and locations resulted in the second-highest modularity. This was followed by the combination of mentions, languages, and hashtags. On the other hand, the LPGA found the combination of mentions, locations, and languages had the highest modularity. The combination of mentions, followers, locations, and languages had the second-highest modularity followed by the combination of mentions, hashtags, and followers.

Fig. 6 Comparison of Rand Index Measure

The modularity of combinations was similar for both the LGA and the LPGA (Figure 5), but the Rand index showed higher accuracy in the results of the LGA than the results of the LPGA (Figure 6). The results of both algorithms revealed that combining features produced higher results than relying solely on one feature. This demonstrates that social graphs including more than one feature influence the spread of negative emotions. Overall, the study found that mentions have the most significant influence on the diffusion of negative emotions. Each combination of features that included mentions produced higher modularity results. Similar hashtags, locations, and languages also enhanced the diffusion of tweets containing negative emotions. High modularity indicated strong ties between the nodes in the community, which indicated the diffusion of negative emotions. These results support [42] the conclusion that users who are retweeted are usually mentioned in tweets as well. It also supports results of [45]; the conclusion that Twitter interactions, specifically mentions, were strong predictors of information diffusion on the network.

6. Conclusion

Social media such as Twitter heavily influences individuals daily lives. Large amounts of information are distributed, which affects users in different ways. Negative emotion diffusion on Twitter is enhanced by a communities features, such as mentions, hashtags, etc. This article has proven that the combination of features increases the diffusion of negative emotions. The LGA a less computationally expensive and quicker manner. This work can be enhanced by considering diffusion measures that can measure the extent of the diffusion and find the features that depend on that. Additionally, the work can be enhanced by finding the prominent actor and if he affects the diffusion of negative emotions contagion in social networks.

Acknowledgement

The authors would like to thank the Deanship of scientific research for funding and supporting this research through initiative of DSR Graduate Students Research Support (GSR).

References

[1] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E.: 'Fast unfolding of communities in large networks', Journal of statistical mechanics: theory and experiment, 2008, 2008, (10), pp. P10008

[2] AlSagri, H.S., and Ykhlef, M.: 'A framework for analyzing and detracting negative emotional contagion in online social networks', in Editor (Ed.)^(Eds.): 'Book A framework for analyzing and detracting negative emotional contagion in online social networks' (IEEE, 2016, edn.), pp. 115-120

[3] AlSagri, H., and Ykhlef, M.: 'Analyzing and Detracting Negative Emotion Contagion Influence in Online Social Networks-Position Paper', International Journal of Computing, Communication and Instrumentation Engineering (IJCCIE) 2017, 4, (1)

[4] Hill, A.L., Rand, D.G., Nowak, M.A., and Christakis, N.A.: 'Emotions as infectious diseases in a large social network: the SISa model', Proceedings of the Royal Society of London B: Biological Sciences, 2010, 277, (1701), pp. 3827-3835

[5] Joiner, T.E., and Katz, J.: 'Contagion of depressive symptoms and mood: meta-analytic review and explanations from cognitive, behavioral, and interpersonal viewpoints', Clinical Psychology: Science and Practice, 1999, 6, (2), pp. 149-164

[6] Soliman, M., Girgis, J., and Morgan, C.: 'Social Media and Health', in Editor (Ed.)^(Eds.): 'Book Social Media and Health' (Published research paper. Medlink Conference, 2013, edn.), pp.

[7] Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A.C.: 'Bad news travel fast: A content-based analysis of interestingness on twitter', in Editor (Ed.)^(Eds.): 'Book Bad news travel fast: A content-based analysis of interestingness on twitter' (ACM, 2011, edn.), pp. 8

[8] Sobkowicz, P., and Sobkowicz, A.: 'Dynamics of hate based Internet user networks', The European Physical Journal B-Condensed Matter and Complex Systems, 2010, 73, (4), pp. 633-643

[9] Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., and Hołyst, J.A.: 'Negative emotions boost user activity at BBC forum', Physica A: statistical mechanics and its applications, 2011, 390, (16), pp. 2936-2944

[10] Kewalramani, M.N.: 'Community detection in Twitter' (University of Maryland, Baltimore County, 2011. 2011)

[11] Deitrick, W., Valyou, B., Jones, W., Timian, J., and Hu, W.: 'Enhancing sentiment analysis on twitter using community detection', Communications and Network, 2013, 5, (03), pp. 192

[12] Orman, G.K., and Labatut, V.: 'A comparison of community detection algorithms on artificial networks', in Editor (Ed.)^(Eds.): 'Book A comparison of community detection algorithms on artificial networks' (Springer, 2009, edn.), pp. 242-256

[13] Riquelme, F., and González-Cantergiani, P.: 'Measuring user influence on Twitter: A survey', Information Processing & Management, 2016, 52, (5), pp. 949-975

[14] Bann, E.Y., and Bryson, J.J.: 'The conceptualisation of emotion qualia: Semantic clustering of emotional tweets', in Editor (Ed.)^(Eds.): 'Book The conceptualisation of emotion qualia: Semantic clustering of emotional tweets' (2013, edn.), pp. 249-263

[15] Murgia, A., Tourani, P., Adams, B., and Ortu, M.: 'Do developers feel emotions? an exploratory analysis of emotions in software artifacts', in Editor (Ed.)^(Eds.): 'Book Do developers feel emotions? an exploratory analysis of emotions in software artifacts' (ACM, 2014, edn.), pp. 262-271

[16] Hodas, N.O., and Lerman, K.: 'The simple rules of social contagion', Scientific reports, 2014, 4

[17] Potter, C.W.: 'Chronicle of influenza pandemics': 'Influenza ' (Oxford: Blackwell Science, 1998)

[18] Pinheiro, F.L., Santos, M.D., Santos, F.C., and Pacheco, J.M.: 'Origin of peer influence in social networks', Physical review letters, 2014, 112, (9), pp. 098702

[19] Coviello, L., Sohn, Y., Kramer, A.D., Marlow, C., Franceschetti, M., Christakis, N.A., and Fowler, J.H.: 'Detecting emotional contagion in massive social networks', PloS one, 2014, 9, (3), pp. e90315

[20] Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., and Quattrociocchi, W.: 'Echo chambers: Emotional contagion and group polarization on facebook', Scientific reports, 2016, 6, pp. 37825

[21] Xiong, X., Li, Y., Qiao, S., Han, N., Wu, Y., Peng, J., and Li, B.: 'An emotional contagion model for heterogeneous social media with multiple behaviors', Physica A: statistical mechanics and its applications, 2018, 490, pp. 185-202

[22] Councill, I.G., McDonald, R., and Velikovich, L.: 'What's great and what's not: learning to classify the scope of negation for improved sentiment analysis', in Editor (Ed.)^(Eds.): 'Book What's great and what's not: learning to classify the scope of negation for improved sentiment analysis' (Association for Computational Linguistics, 2010, edn.), pp. 51-59

[23] Cole, W.D.: 'An information diffusion approach for detecting emotional contagion in online social networks' (Arizona State University, 2011. 2011)

[24] Rosenquist, J.N., Fowler, J.H., and Christakis, N.A.: 'Social network determinants of depression', Molecular psychiatry, 2011, 16, (3), pp. 273-281

[25] Botha, E.M.: 'Contagious Communications: The role of emotion in viral marketing', KTH Royal Institute of Technology, 2014

[26] Kanavos, A., Perikos, I., Vikatos, P., Hatzilygeroudis, I., Makris, C., and Tsakalidis, A.: 'Modeling retweet diffusion using emotional content', in Editor (Ed.)^(Eds.): 'Book Modeling retweet diffusion using emotional content' (Springer, 2014, edn.), pp. 101-110

[27] Fortunato, S.: 'Community detection in graphs', Physics reports, 2010, 486, (3), pp. 75-174

[28] 2Nandi, G., and Das, A.: 'A survey on using data mining techniques for online social network analysis', Int. J. Comput. Sci. Issues (IJCSI), 2013, 10, (6), pp. 162-167

[29] Lerman, K., and Ghosh, R.: 'Information contagion: An empirical study of the spread of news on Digg and Twitter social networks', ICWSM, 2010, 10, pp. 90-97
[30] Paranyushkin, D.: 'Informational epidemics and synchronized viral contagion in social networks', in Editor (Ed.)^(Eds.): 'Book Informational epidemics and synchronized viral contagion in social networks' (Nodus Labs. Retrieved from ht tp://noduslabs. com/publications/text-polysingularity-network-analysis. pdf, 2012, edn.), pp.
[31] Ball, F., Mollison, D., and Scalia-Tomba, G.: 'Epidemics with two levels of mixing', The Annals of Applied Probability, 1997, pp. 46-89
[32] Adedoyin-Olowe, M., Gaber, M.M., and Stahl, F.: 'A survey of data mining techniques for social media analysis', arXiv preprint arXiv:1312.4617, 2013
[33] Aggarwal, C.C.: 'An introduction to social network data analytics', Social network data analytics, 2011, pp. 1-15
[34] Girvan, M., and Newman, M.E.: 'Community structure in social and biological networks', Proceedings of the national academy of sciences, 2002, 99, (12), pp. 7821-7826
[35] Rani, T., and Goyal, A.: 'Survey of Clustering Techniques for Information Retrieval in Data Mining', IJSETR
[36] Kim, Y.-H., Seo, S., Ha, Y.-H., Lim, S., and Yoon, Y.: 'Two applications of clustering techniques to twitter: Community detection and issue extraction', Discrete dynamics in nature and society, 2013, 2013
[37] Shalizi, C.R., and Thomas, A.C.: 'Homophily and contagion are generically confounded in observational social network studies', Sociological methods & research, 2011, 40, (2), pp. 211-239
[38] Newman, M.E., and Girvan, M.: 'Finding and evaluating community structure in networks', Physical review E, 2004, 69, (2), pp. 026113
[39] Bickel, P.J., and Chen, A.: 'A nonparametric view of network models and Newman–Girvan and other modularities', Proceedings of the national academy of sciences, 2009, 106, (50), pp. 21068-21073
[40] Porter, M.A., Onnela, J.-P., and Mucha, P.J.: 'Communities in networks', Notices of the AMS, 2009, 56, (9), pp. 1082-1097
[41] Zhu, J., Wang, B., Wu, B., and Zhang, W.: 'Emotional Community Detection in Social Network', IEICE Transactions on Information and Systems, 2017, E100.D, (10), pp. 2515-2525
[42] Clauset, A., Newman, M.E., and Moore, C.: 'Finding community structure in very large networks', Physical review E, 2004, 70, (6), pp. 066111
[43] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P.K.: 'Measuring user influence in twitter: The million follower fallacy', ICWSM, 2010, 10, (10-17), pp. 30
[44] Kim, K., Jung, J.-Y., and Park, J.: 'Discovery of Information Diffusion Process in Social Networks', IEICE Transactions on Information and Systems, 2012, E95.D, (5), pp. 1539-1542
[45] Yang, J., and Counts, S.: 'Predicting the Speed, Scale, and Range of Information Diffusion in Twitter', ICWSM, 2010, 10, (2010), pp. 355-358
[46] Tareaf, R.B., Berger, P., Hennig, P., Koall, S., Kohstall, J., and Meinel, C.: 'Information Propagation Speed and Patterns in Social Networks: A Case Study Analysis of German Tweets', Journal of Computers, 2018, 13, (7), pp. 761-771
[47] Caldarelli, G.: 'Large scale structure and dynamics of complex networks: from information technology to finance and natural science' (World Scientific, 2007. 2007)
[48] Orman, G.K., Labatut, V., and Cherifi, H.: 'Comparative evaluation of community detection algorithms: a topological approach', Journal of statistical mechanics: theory and experiment, 2012, 2012, (08), pp. P08001
[49] Khan, B.S., and Niazi, M.A.: 'Network community detection: a review and visual survey', arXiv preprint arXiv:1708.00977, 2017
[50] Newman, M.E.: 'Modularity and community structure in networks', Proceedings of the national academy of sciences, 2006, 103, (23), pp. 8577-8582
[51] Sohn, Y., Choi, M.-K., Ahn, Y.-Y., Lee, J., and Jeong, J.: 'Topological cluster analysis reveals the systemic organization of the Caenorhabditis elegans connectome', PLoS computational biology, 2011, 7, (5), pp. e1001139
[52] Haynes, J., and Perisic, I.: 'Mapping search relevance to social networks', in Editor (Ed.)^(Eds.): 'Book Mapping search relevance to social networks' (ACM, 2009, edn.), pp. 2
[53] Lancichinetti, A., and Fortunato, S.: 'Erratum: Community detection algorithms: A comparative analysis [Phys. Rev. E 80, 056117 (2009)]', Physical review E, 2014, 89, (4), pp. 049902
[54] Browet, A.: 'Algorithms for community and role detection in networks', Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2014
[55] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P.: 'Community detection in social media', Data Mining and Knowledge Discovery, 2012, 24, (3), pp. 515-554
[56] Raghavan, U.N., Albert, R., and Kumara, S.: 'Near linear time algorithm to detect community structures in large-scale networks', Physical review E, 2007, 76, (3), pp. 036106
[57] Deitrick, W., and Hu, W.: 'Mutually enhancing community detection and sentiment analysis on twitter networks', Journal of Data Analysis and Information Processing, 2013, 1, (03), pp. 19
[58] Steinhaeuser, K., and Chawla, N.V.: 'Identifying and evaluating community structure in complex networks', Pattern Recognition Letters, 2010, 31, (5), pp. 413-421
[59] Rand, W.M.: 'Objective criteria for the evaluation of clustering methods', Journal of the American Statistical association, 1971, 66, (336), pp. 846-850
[60] Greene, D., and Cunningham, P.: 'Producing a unified graph representation from multiple social network views', in Editor (Ed.)^(Eds.): 'Book Producing a unified graph representation from multiple social network views' (ACM, 2013, edn.), pp. 118-121
[61] Wang, S., Li, X., Ye, Y., Huang, X., and Li, Y.: 'Multi-attribute and relational learning via hypergraph regularized generative model', Neurocomputing, 2018, 274, pp. 115-124
[62] Bouguessa, M., Missaoui, R., and Talbi, M.: 'A Novel Approach for Detecting Community Structure in Networks', in Editor (Ed.)^(Eds.): 'Book A Novel Approach for Detecting Community Structure in Networks' (IEEE, 2014, edn.), pp. 469-477
[63] Jin, D., Liu, D., Yang, B., and Liu, J.: 'Fast complex network clustering algorithm using agents', in Editor (Ed.)^(Eds.): 'Book Fast complex network clustering algorithm using agents' (IEEE, 2009, edn.), pp. 615-619

[64] Hamed, A.A., Wu, X., and Rubin, A.: 'A twitter recruitment intelligent system: association rule mining for smoking cessation', Social Network Analysis and Mining, 2014, 4, (1), pp. 212
[65] Zhu, J., Xiong, F., Piao, D., Liu, Y., and Zhang, Y.: 'Statistically modeling the effectiveness of disaster information in social media', in Editor (Ed.)^(Eds.): 'Book Statistically modeling the effectiveness of disaster information in social media' (IEEE, 2011, edn.), pp. 431-436
[66] Gupta, D., and Ghafir, S.: 'An overview of methods maintaining diversity in genetic algorithms', International journal of emerging technology and advanced engineering, 2012, 2, (5), pp. 56-60
[67] Kashyap, H., Das, S., Bhattacharjee, J., Halder, R., and Goswami, S.: 'Multi-objective genetic algorithm setup for feature subset selection in clustering', in Editor (Ed.)^(Eds.): 'Book Multi-objective genetic algorithm setup for feature subset selection in clustering' (IEEE, 2016, edn.), pp. 243-247
[68] Parrott, W.G.: 'Emotions in social psychology: Essential readings' (Psychology Press, 2001. 2001)
[69] Eckman, P.: 'Emotions revealed: Understanding faces and feelings', Weidenfeld & Nicolson, London, England, 2003

Hatoon S AlSagri received the Master's degree in information systems from the Department of Information Systems, College of Computer and Information Sciences, King Saud University, where she is currently pursuing the Ph.D. degree in information systems. She is currently a Lecturer with the Department of Information Systems, College of Computer and Information Sciences, Al-Imam Mohammad bin Saud Islamic University. During her graduate studies, she has had the opportunity to participate in various conferences and has published various journal articles. Her main research interests lie in the field of data mining, information diffusion, and social analysis

Mourad Ykhlef received the B.Eng. degree in computer science from Constantine University, Algeria, the M.Sc. degree was in artificial intelligence from University Paris 13, France, and the Ph.D. degree in computer science from University Bordeaux 1, France. He is currently a Professor with the Department of Information Systems, College of Computer and Information Sciences, King Saud University, Saudi Arabia. His main research interests include data mining, data warehouse, XML and bio-inspired computing.