Social Media Analysis using Optimized K-Means Clustering

Ahmed Alsayat
Department of Computer Science
Bowie State University
asayat@ju.edu.sa

Hoda El-Sayed
Department of Computer Science
Bowie State University
helsayed@bowiestate.edu

***Abstract*—The increasing influence of social media and enormous participation of users creates new opportunities to study human social behavior along with the capability to analyze large amount of data streams. One of the interesting problems is to distinguish between different kinds of users, for example users who are leaders and introduce new issues and discussions on social media. Furthermore, positive or negative attitudes can also be inferred from those discussions. Such problems require a formal interpretation of social media logs and unit of information that can spread from person to person through the social network. Once the social media data such as user messages are parsed and network relationships are identified, data mining techniques can be applied to group different types of communities. However, the appropriate granularity of user communities and their behavior is hardly captured by existing methods. In this paper, we present a framework for the novel task of detecting communities by clustering messages from large streams of social data. Our framework uses K-Means clustering algorithm along with Genetic algorithm and Optimized Cluster Distance (OCD) method to cluster data. The goal of our proposed framework is *twofold* that is to overcome the problem of general K-Means for choosing best initial centroids using Genetic algorithm, as well as to maximize the distance between clusters by pairwise clustering using OCD to get an accurate clusters. We used various cluster validation metrics to evaluate the performance of our algorithm. The analysis shows that the proposed method gives better clustering results and provides a novel use-case of grouping user communities based on their activities. Our approach is optimized and scalable for real-time clustering of social media data.**

Keywords: K-Means, Genetic Algorithm, Clustering, Social Media Analysis, DataMining.

I. Introduction

Social media serves as a ubiquitous public platform which remains accessible to users as a multilayered group of internet applications. Within the applications, the user creates individual and unique expression for content exchange [1]. What remains valuable and fascinating is the level of social media data influence as the platform remains an intense portal of human interaction and behavior. What remains dynamic about social media is the level of opportunity that influences individuals, groups and society. The study of this data by industry specialists seeking new and inventive tactics to collect data for analysis remains important to the future of social media [2]. There are many social media network sites, Twitter and Facebook remain the most well-known but other forms being used are blogs, wikis and platforms with unfiltered text and information which remains of true focus for users. Slashdot [3] is a website that focuses on the publication of technology news where the stories could be written by editors, or posted and commented by users.

What makes Slashdot different from other sites is the ability for the user to rate those within the community based upon a positive or negative rating. Industry researchers remain focused on three areas for social media application: business, bioscience and social science. What has been found is that the social media is extremely valuable to statistical study of information technology, social behavior [4] within quantitative attributes for e-learning process and simulation design for further data mining [5].

The rate of digital interaction and exchange of data amongst users is at an all time high. These platforms like Facebook and Twitter remain go-to sites for information, user expression of opinions and the place many are getting his or her news [6] [7]. These are platforms of free speech and protests [8] [9], organization and sharing of common interests [10] and ways to keep family and friends in the loop of everyday life. Still this presents a challenge to industry specialists that collect data from these platforms and social communities. Such data mining actions require the ability to interpret the massive amount of content continuously produced on online social media and manual labeling is infeasible on a large scale. Such activities also support a myriad of muddling through a vast varying degree of content, some of which may not be valuable to the data mining task. Textual content equals a unit of information and this can also be coded to represent a particular trait of the individual posting the content. Each piece of content also represents a users score point and this makes the process of assessing information from the standpoint of learning methods as these act as individual means of identifying group behaviors.

In this paper, we combine text mining, network analytics and data mining together to provide a better description of each user in a Slashdot forum in terms of leadership and sentiment. For each individual user, the authority score, hub score, and rating are calculated by using the network analytics. High authority scores differentiate between users with a high level of leadership where as high hub score represents users with a high follower behavior. The attitude of each individual user in the forum is measured by rating. So, by using the attitude

978-1-5090-0809-4/16/$31.00 copyright 2016 IEEE
SERA 2016, June 8-10, 2016, Baltimore, USA

measure, we could identify positive, neutral, and negative users among a wide spectrum of positivity and negativity. When examining the attitude measure, we look to the authority/follower model for scoring which helps us in furthering focusing upon the scatter plot of user outcomes for attitude within the authority/follower framework. The process continues as we apply optimized K-Means clustering algorithm to find different groups and explain the overall community. Our method uses Genetic algorithm along with the novel method called Optimized Cluster Distance (OCD) to improve the performance of ordinary K-Means algorithm. Empirical results show that the combination of Genetic algorithm and OCD method shows better performance and can be used in several clustering applications. Using the proposed method, we are able to quickly and effectively create new understanding of the social media segmentation with higher confidence.

The rest of this paper is organized as follows. Section II presents a review of existing social media analysis. Section III describes the K-Means, Genetic algorithms, and also provides an overview of our optimized K-Means algorithm where OCD is applied. Section IV provides a description of the dataset along with the experimental setup, and Section V provides a discussion of the results with a use-case application. Finally, Section VI concludes this paper along with plans for future work.

II. Related Work

Much can be learned about the retail and finance behaviors of users by studying social media analysis. It is nothing new that retail companies market via social networks to discover what consumers think about branding, customer relationship management, and other strategies including risk prevention. For finance, studies have been looking at how users rate sentiment and how this influences trading activities. A good example [11] is the found correlation of data on Twitter with industry market behavior and sentiment posted by users. Wolfram [12] used Twitter data to develop machine learning model using Support Vector Regression and predicted prices of individual stocks and found significant advantage of using social media data for forecasting future prices.

Such data can be used to track health issues like smoking and obesity for bio-scientific study. Researchers at Penn State University [13] found innovative systems and techniques to track the spread of infectious diseases because the data social media reflects about users within these groups.

As far as computational social science applications are concerned, it includes monitoring public responses to announcements, speeches and events with emphasis on political comments and initiatives. It also gives insights about community behavior, social media polling within groups and early detection of emerging events. For example, [14] by using the computational linguistics, the automatic prediction impact of news on the public perception of political candidates was implemented. Yessenov and Misailovic [15] use reviewing comments of movie to learn much about various approaches techniques used for extracting features of text based on the accuracy of several machine learning methods such as Nave Bayes, Decision Trees, Maximum Entropy and K-means Clustering. In addition, Karabulut [16] found that Facebook also exhibits and captures major public events in its data.

Social network analysis has a well-defined relation and background in sociology [17]. With the rapid growth of the web forums and blogs, the users participation on content creation led to a huge amount of dataset. Hence, the advancement of data mining techniques are required. An overall discussion of one news forum called Slashdot can be found in [3] [18]. The focus of our work is to develop a clustering framework using optimized K-Means algorithm that is more accurate than existing methods. Clustering is used as an exploratory analysis tool that aims at categorising objects into categories, so the association degree between the objects is maximal when belonging to the same categories. Clustering structures the data into a collection of objects that are similar or dissimilar and is considered an *unsupervised learning*. The application of our method is mainly on finding user groups based on leadership, follower, and attitude features as suggested in the authority/follower model.

III. Methods

In this section, we describe K-Means clustering algorithm, Genetic algorithm and our proposed Optimized K-Means algorithm in detail.

A. K-Means Clustering

K-Means is an unsupervised clustering algorithm which is used to find groups within the data [19] [20]. Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional vector, k-means clustering aims to partition the n observations into a set of k clusters $(\leq n)$ such as $S = S_1, S_2, \ldots, S_k$ so as to minimize the within-cluster sum of squares (WSS) which is defined as sum of distance functions of each point in the cluster to the k centers. The objective function of K-Means is to find

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x_i - c_i\|^2 \tag{1}$$

where c_i is the centroid of points in S_i.

B. Genetic Algorithm

Use of genetic algorithms fall under a larger standard of algorithms called evolutionary algorithms or EA. These EAs work to promote problem solving toward optimizing techniques which promote significant variance in genetic features like inheritance, mutation, selection and crossover of attributes [21]. Much of what takes place happens as random with the population of user/individuals and this can be labeled as a generational group. To evaluate, each group generation is assessed for fitness within the optimization for serving to promote problem solving. The more up to the challenge the individual proves to be within the population, also signifies the rate of individual genome mutation which leads to creating the next generation of genetic characteristics. Thus, the algorithm continues to provide solutions within the generational testing. With the maximum amount of generations produced during the process, the algorithm finalizes and the fitness level standard is reached.

C. Proposed Algorithm

We propose an optimized K-Means clustering algorithm which combines the power of Genetic algorithm to find best possible centroids for K-Means algorithm. Further, the solution is more enhanced by Optimized Cluster Distance (OCD) method which involves re-clustering of data with new centers and try to increase the between cluster distances (BSS) and decrease within cluster distances (WSS).

Figure 1 shows the flow of our algorithm. Genetic algorithm starts by generating initial population based on k provided as a parameter for number of clusters. Then, it computes the fitness of initial population Mean Square Error (MSE). The MSE calculates the distance between the cluster centers and remaining data points.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 \qquad (2)$$

where c is the center from generated population.

$$Fitness = \frac{1}{MSE} \qquad (3)$$

The algorithm then goes to the next generation or the next iteration if we want to get the better fitness than the fitness from previous generation. In that case, genetic algorithm performs selection, crossover, and mutation to generate new population. Process will stop if the best fitness value seems not to be changed in next generation.

Once Genetic algorithm outputs the best centroids, we run K-Means algorithm on those centroids and get the clustering results. At this point, we introduce our novel technique called Optimized Cluster Distance (OCD) which further process the K-Means clusters and try to minimize the within cluster variance and maximize distances between clusters. This is done by first calculating a pairwise distance between all clusters. Then the algorithm picks all pairs of clusters which have distances smaller than overall average of cluster distances. Once the pair of clusters are selected, the algorithm re-clusters them using the same K-Means and GA algorithms. Since at this point we are clustering a small subset of data, the algorithm easily finds the partitions which are more distant to each other as compared to the partitions before OCD process. The pseudocode of our algorithm is stated in Algorithm 1.

IV. Dataset Description

This section explains the dataset and experimental setup used in this paper. We use a public social media dataset from Slashdot where the users can post, read, and comment on the published news. There are many user communities that can serve as participant pools and it was found that an active community can consist of more than 200 participants with active responses. Most of the users are registered and

Algorithm 1 Optimized K-Means-GA-OCD Algorithm

Input: Input dataset D, Number of clusters k, Number of iterations N
Output: Output dataset with k cluster labels

```
procedure K–MEANS–GA–OCD
    $i$=1
    while $i \leq N$ do
        fitness = GA(population(i), D)
        if $fitness(i) > fitness(i+1)$ then
            centers = GA-Centers(D, $k$)
        end if
    end while
    clusters = K-Means(D, centers)
    dist = PairwiseDistance(centers)
    for each $d_{xy} \in dist \leq average(dist)$ do
        while $max\ iterations$ do
            centers = GA-Centers(D, 2)
            newclusters = K-Means(D, centers)
            if $wss_{new} \leq wss_{xy}$ then
                $cluster_x = cluster_{new_x}$
                $cluster_y = cluster_{new_y}$
            end if
        end while
    end for
end procedure
```

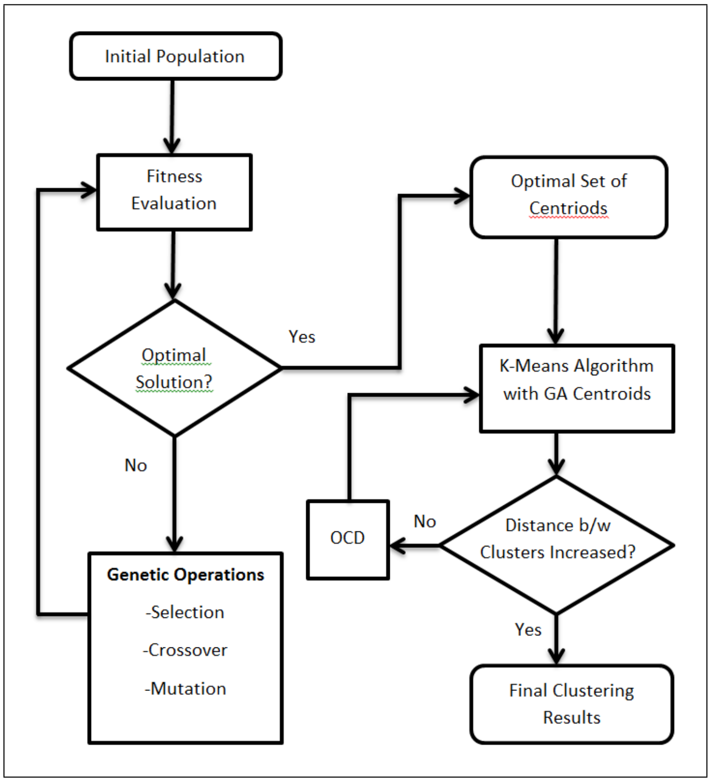

Fig. 1. Optimized K-Means Clustering Flow Diagram.
The figure shows the process of Optimized K-Means Clustering algorithm (a) Genetic algorithm to get optimal centroids (b) K-Means algorithm to get clusters (c) Optimized Cluster Distance (OCD) to minimize sum of square distances.

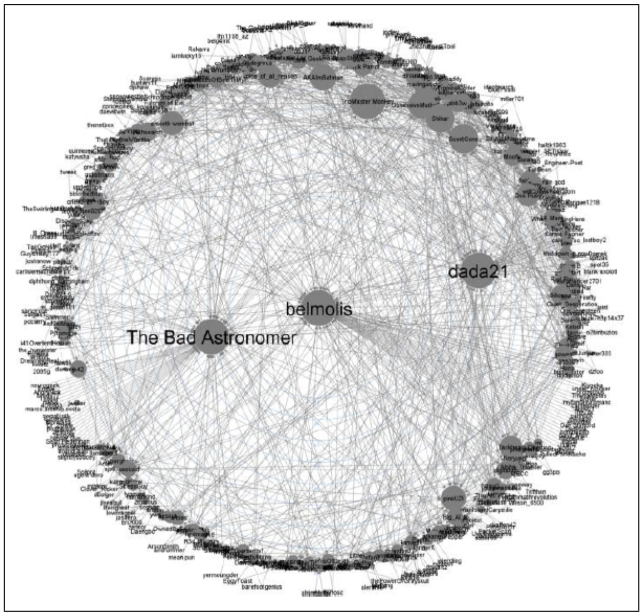

Fig. 2. Network Diagram for Slashdot Dataset.
Slashdot networking diagrams on the topics of NASA.

leave comments by their nickname, although some participate anonymously. For example, Figure 2 shows the networking diagram of slash data on the topic of NASA where users who are active participants are shown using big circles [22]. In our experiments, we used a subset of the Slashdot data provided by Barcelona Media. The content of this dataset has about $140,000$ comments from 496 articles subjected in politics where the total number of users is about $24,000$ users.

In order to make it useful for our Optimized K-Means algorithm, we preprocessed this data using KNIME tool [23]. For each community user, an Authority Score and a Hub Score is calculated by detecting activity of users on web pages. The positive and negative attitude of each user called rating. This rating is also calculated by extracting bag of words across all of the published posts using sentiment analysis. The description of each individual community user is defined by means of an authority score (leadership), a hub score (follower), and the level for the attitude of positivity or negativity. The opinion of the leaders spread quickly over a large number of users; hence, they represent an important category [23]. What has been found is that followers do not create influence over other users as predicted. In other words, their attitude does not influence other users.

V. Results and Discussions

A set of experiments evaluating the clustering performance of our proposed K-Means-GA-OCD algorithm was performed. Additionally an evaluation of the algorithm with a different number of clusters was made and compared with the existing methods in order to measure their effectiveness. Figure 3 shows the visualization of clustering output using K-Means-GA-OCD algorithm when a number of clusters (k) is set to 10. Principal Component Analysis (PCA) was used on a multi-dimensional data in order to visualize the results in two-dimension [24]. Section V-A is a comparison between other methods while Section V-B gives an insight into the use-case and applications of the proposed method:

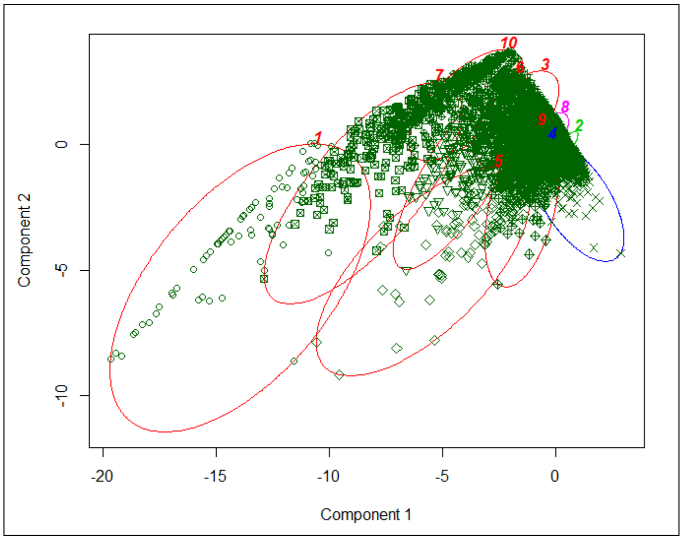

Fig. 3. Cluster Representation using K-Means-GA-OCD algorithm.
The figure shows the visualization of clustering output using K-Means-GA-OCD algorithm when number of clusters are set to 10. We use Principal Component Analysis (PCA) to visualize the results in two-dimension.

A. Comparison to other methods

Figure 4 shows the sum of square distances (WSS) for K-Means, K-Means-GA and K-Means-GA-OCD algorithm with varying number of clusters. From this graph we can conclude that our proposed algorithm is performing better than other algorithms by minimizing the WSS across all the number of clusters. Furthermore, k=10 this gives a breaking knee point for our data which shows that using 10 clusters would give us better understanding of the data. Table I shows the actual within sum of square distances between K-Means and the proposed algorithm.

When evaluating the proposed approach with other algorithms the properties and attributes of the social media data being clustered were considered. An experiment was also designed to compare Optimized Cluster Distance method (OCD) step of our proposed algorithm. This experiment, evaluates if introducing a novel OCD method actually improves the clustering results. Table II shows that our proposed method helps K-Means-GA algorithm to further improve its clustering results by minimizing WSS.

B. Use-Case and Application

In order to provide a use-case as an application to our proposed method, we processed our clustering results to gain some insights when k is set to 10. Figure 5 shows a bubble chart of our clustering result. In this chart, center of a cluster represents the authority and hub score, whereas size of the bubble represents the number of users within that cluster [25]. Cluster 5 has mostly moderate users (medium values for the

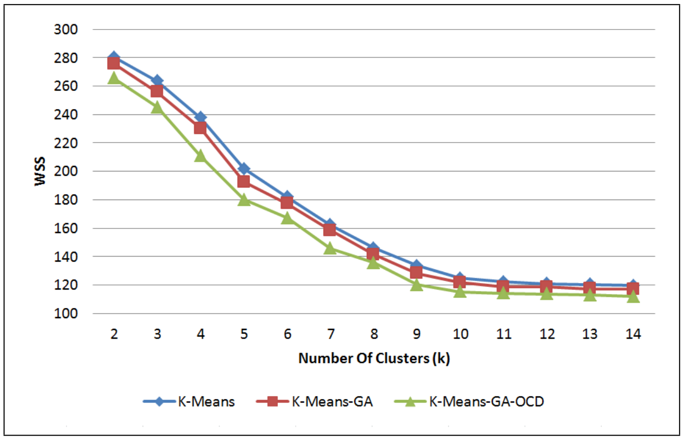

Fig. 4. Comparison Of Proposed Algorithm with other methods.
The figure shows comparison of Optimized K-Means Clustering algorithm (K-Means-GA-OCD) with simple K-Means and K-Means-GA algorithm. The comparison is based on within sum of square distances (WSS).

TABLE I
WSS COMPARISON OF PROPOSED ALGORITHM WITH K-MEANS

	Within Sum of Square Distance (WSS)	
# of clusters (k)	K-Means	K-Means-GA-OCD
2	280.18	265.50
3	263.52	245.15
4	237.78	200.74
5	201.77	180.2
6	181.86	167.32
7	162.33	145.86
8	145.94	135.66
9	133.69	120.27
10	125.02	115.41
11	122.18	114.31
12	121.10	113.78
13	120.45	112.94
14	119.74	112.05

The number shows the within sum of square distance for each algorithm against number of clusters.

TABLE II
WSS COMPARISON OF K-MEANS-GA AND K-MEANS-GA-OCD

	Within Sum of Square Distance (WSS)	
# of clusters (k)	K-Means-GA	K-Means-GA-OCD
2	275.52	265.50
3	255.76	245.15
4	230.08	200.74
5	192.45	180.2
6	177.27	167.32
7	158.48	145.86
8	141.45	135.66
9	128.25	120.27
10	123.82	115.41
11	118.79	114.31
12	118.64	113.78
13	117.52	112.94
14	117.10	112.05

The number shows the within sum of square distance for each algorithm against number of clusters. Here we see that adding a new OCD method minimizes the WSS and improves clustering results.

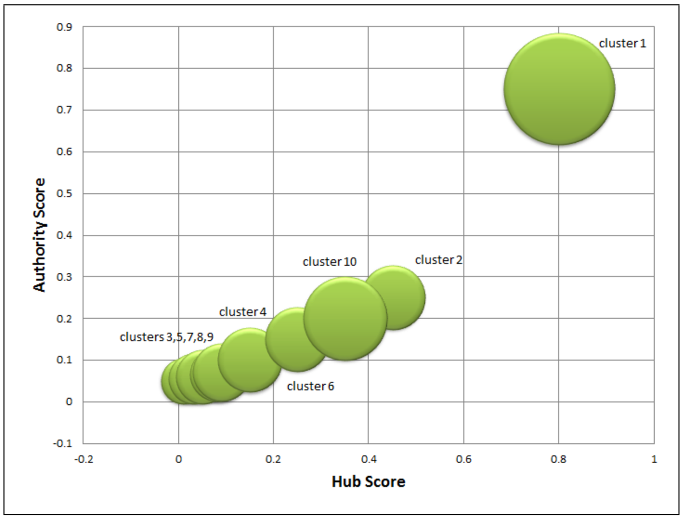

Fig. 5. Bubble Chart of the Clustering Result.
The figure represents a community of users on Slashdot based on our clustering results. x-axis represents Hub score and y-axis represents Authority score.

Authority Score and for the Hub Score) and slightly contains negative attitude (low score for Rating). On the other hand cluster 10 contains more active users (with higher values for the Authority and the Hub Score) and more negative attitude users. Three clusters, cluster 1, cluster 2 and cluster 4 have an attitude level close to 1, which indicates that they are positive attitude users. Especially cluster 1 and cluster 2 which contain very active users (very high values for the Hub and the Authority Score) and very positive attitude users. These are the teams you would like to call for help to spread news and explain new products to other users [24]. Cluster 4 also contains positive users, but they are not as active as cluster 1 and cluster 2. Furthermore, cluster 3 and cluster 6 have good attitude level but are less active as compared to previous three clusters. They have positive attitude but are moderate and less active than the previous ones. The remaining three clusters 7, 8 and 9 consist of users with neutral attitude [25]. These clusters also show very low activity scores (Authority and Hub Score), indicating that neutral users are also rarely involved in discussions.

In order to assign self-explanatory labels to each cluster, we can label clusters 1 and 2 as “Super Users” because of the highest activity and most positive attitude. It is possible to leverage their attitude to spread the news and talk about new issues. Cluster 10 is labelled as “Active Negative” as it contains very active and negative users and usually position complains on the forum. The remaining clusters can be labelled as “Neutral Inactive” which represents the majority of the users who can have any opinions such as negative, positive, or neutral about the discussion topic [26].

VI. Conclusion

This paper, propose a novel method to analyze social media data. Our method used K-Means algorithm along with Genetic algorithm and Optimized Cluster Distance method to cluster the social media community based on leadership, follower and attitude scores. With the empirical evaluation, the proposed algorithm outperforms other existing methods. It also presents a use-case of the method to further describe user community by getting more insights from clustering results and assigning self-explanatory labels to each cluster. For future work, the method is to be used with different domains such as Bioinformatics or Image Processing and comparing it with the state-of-the-art methods.

Acknowledgment

The authors would like to thank the Department of Computer Science at Bowie State University.

References

[1] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[2] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116, 2015.

[3] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750. ACM, 2009.

[4] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[5] Claudio Cioffi-Revilla. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271, 2010.

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[7] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[8] Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movement communication network. *PloS one*, 8(3):e55957, 2013.

[9] Michael D Conover, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. The digital evolution of occupy wall street. 2013.

[10] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM, 2011.

[11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[12] M Sebastian A Wolfram. Modelling the stock market using twitter. *School of Informatics*, page 74, 2010.

[13] Marcel Salathe, Linus Bengtsson, Todd J Bodnar, Devon D Brewer, John S Brownstein, Caroline Buckee, Ellsworth M Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L Mabry, et al. Digital epidemiology. *PLoS Comput Biol*, 8(7):e1002616, 2012.

[14] Kevin Lerman, Ari Gilder, Mark Dredze, and Fernando Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 473–480. Association for Computational Linguistics, 2008.

[15] Kuat Yessenov and Saša Misailovic. Sentiment analysis of movie review comments. *Methodology*, pages 1–17, 2009.

[16] Yigitcan Karabulut. Can facebook predict stock market activity? In *AFA 2013 San Diego Meetings Paper*, 2013.

[17] John Scott. *Social network analysis*. SAGE Publications Ltd, 2013.

[18] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645–654. ACM, 2008.

[19] Rousseeuw Peter J Kaufman, Leonard. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[20] John Burkardt. K-means clustering. *Virginia Tech, Advanced Research Computing, Interdisciplinary Center for Applied Mathematics*, 2009.

[21] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Genetic algorithm-based clustering technique. *Pattern recognition*, 33(9):1455–1465, 2000.

[22] Killian Thiel, Tobias Kötter, Michael Berthold, Rosaria Silipo, and Phil Winters. Creating usable customer intelligence from social media data: Network analytics meets text mining. *KNIME.org Report*, 2012.

[23] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31, 2009.

[24] Rosaria Silipo, Phil Winters, Killian Thiel, and Tobias Kötter. Creating usable customer intelligence from social media data: Clustering the social community. *KNIME.com Report*, 2012.

[25] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 548–555. IEEE, 2013.

[26] Lei Meng, Ah-Hwee Tan, and Donald C Wunsch. Adaptive scaling of cluster boundaries for large-scale social media data clustering. In *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14. IEEE, 2015.