2018 IEEE International Conference on Data Mining Workshops (ICDMW)

An Empirical Study on Sentiments in Twitter Communities

Noha Alduaiji*†, Amitava Datta*
*Department of Computer Science and Software Engineering
The University of Western Australia
Perth, 6009, Australia
Amitava.Datta@uwa.edu.au
†Majmaah University
11952, Saudi Arabia
N.Alduaiji@mu.edu.sa

***Abstract*—Sentiment analysis and community detection are two popular research subjects in data mining. Lots of research have been published in recent years that aim to enhance the mining of text using sentiment analysis tools and to mine network structure to find cohesive and important communities in social networks. However, there is a lack of knowledge of the importance of understanding the sentiment and its changes on the community lifetime. In this paper, we aim to study the sentiments and its impact on user behaviour and the evolution of social network communities. To do that, we collect three Twitter datasets, two of which are based on the communications between people who share following links and the third dataset is based on people who talked about world cup subject. Next, we analyse the sentiments of communications to positive, negative or neutral. After that, we detect communities using k-core. Later, we track changes of sentiments in communities for an extended period of time. Our results showed that the positive sentiment is contagious because members of the communities increasingly share positive tweets more than the negative ones over time. Also, we found a strong correlation between positive sentiments and the size of the community in all three datasets. These results lay shed on the existence of like-minded users within the communities which attract social network companies for their viral marketing and recommendation systems.**

***Keywords*-Sentiment; Twitter Communities; Like-Minded; Sentiment Evolution**

I. Introduction

With the growth of social networks, such as Twitter and Facebook, a massive amount of data become accessible. Therefore, a lot of researchers' attention has been shifted to mine social networks including detecting communities and analysing text for sentiments and opinions. A community is defined as a group of people who tend to have dense connections with each other compared to the rest of the social networks. Example of community detection methods are Louvain [1], Infomap [2] Label Propagation [3], Newman's Leading Eignvector [4], and k-core [5]. In addition, the evolution of communities has been studied in the literature, which classifies the life cycle of the communities according to different events such as birth, merge, split and dissolve. These events are based on the changes of community structure over time gap. The changes in the community structure include establishing links, engaging and deleting links [6].

Another interesting way to mine opinions and emotions is by automating the sentiment analysis of user tweets. Sentiment analysis is a process of classifying opinions to positive negative or neutral opinion. Many techniques and algorithms have been developed to identify whether online text is subjective or objective and whether any opinion expressed is positive or negative [7].

There is a lack of knowledge in regards to the changes of sentiments within the social network communities. We argue that understanding the changes in sentiment in communities helps predict future links and better recommendation systems. For example, a group of people who interact with each other regularly and their interactions carry positive sentiments are likely to communicate more in the future and share similar opinions. On the contrary, a group of people who engage with negative opinions are likely to stop engaging in the future and probably may not share the same opinions on many things.

This paper aims to study the changes in the sentiments within the communities and its impact on community behaviour. The research questions include:

- Do people tend to share positive or negative opinions within communities?
- Does sentiment contagion exist in social communities?
- What are the impacts of sentiments on the evolution of communities?

We verify our study using Twitter data. Twitter is one of the well known social networks at the moment. It has more than 330 million monthly active users and more than 500 million tweets a day [8]. We collected our datasets using Twitter API [9] and since our paper is not focusing on developing a new method to detect communities, we use a k-core decomposition method. It is a well-known method that has been effectively used in the literature to capture the users' engagement and cohesive communities in the social networks [10], [11] and to capture the influence of a community [12]. Moreover, it has been used in the gaming networks to find the most stable communities over time [13]. We analyse the tweets' sentiments and classify

2375-9259/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDMW.2018.00167

them as negative -1, positive +1 and neutral 0 using *TextBlob* which is a natural language processing tool [14].

We study the sentiment evolution in Twitter communities in both types of graphs, static graphs and dynamic graphs. Static graphs are based on the following-follower relationships which are stable and last for a long time. Dynamic graphs are based on interactions such as the use of retweet and mentions. We track the interactions of static communities through their following-follower links on a weekly basis for two and a half years. While for the dynamic communities, we used a popular hashtag and tracked the changes in sentiment on an hourly basis for two days. We analyse both types of communities and then present the results.

The remainder of this paper is presented as follows: in section II, we presents the related work. Section III we introduce the design specifications and explains framework details of our data collecting and filtering system. Section IV we present the experiment and the results of the data collection and analysis of three different Twitter datasets. Section V we discuss our results, and in the last part, section VI, we conclude the work by summarising this paper and suggesting future research directions.

II. Related Work

Since our study is related to sentiment analysis and community detection and evolution, this section summarises some of the associated works in both these fields.

A. Sentiment Analysis

The literature suggests that users are more likely to form following-follower links if they have the same opinion about a particular topic [15]–[17]. Therefore, applications such as recommendation engines and links prediction, have been developed based on analysis of users' sentiments. Also, Choo et al. [18] used the sentiment of user relations to identifies the spammer groups. Their approach includes sentiment analysis of the interactions between users and extracts the common positive sentiment of the relationship. They suggested that the communities with strong positive opinions in the Amazon dataset are more likely to be spammer communities. Lin et al. [19] studied the role of strong connections like close friends on the outcome of emotions such as happiness and envy on Facebook. Their experiment showed that the positive emotions are more pronounced than negative emotions and the strength of connections control the feelings of happiness or envy after receiving posts on Facebook. However, their study is limited to specific countries and two specific products only on the reaction of happiness and envy.

Sentiment analysis has been used to study user behaviour with online friends and whether it is possible to influence someone emotions online. For example, Karmer [20] showed that the emotional contagion is possible over social networks in text communications (post/tweet) through social network links. These links can spread these emotions to direct and indirect connections. That indicates expressing emotion affects the other ties within the expresser community and therefore, the whole community might share the same feeling and get in the same mood as the source of that feeling. Also, Kramer et al. [21] showed that emotions are contagion, so they transfer to other leading people, and the same community shares the same emotion without much awareness. They also showed that the emotion contagion transfers not only to direct connection but also to indirect connections. Their results are based on Facebook. In this paper, we focus on more diverse and broader communities and track these communities over a long time.

B. Community Detections and Evolutions

There are many methods on detecting communities that have cohesive structures such as, Louvain [1], Infomap [2], Clique percolation method [22] and k-core [5]. These methods have shown good results for detecting cohesive and dense communities in social networks with variation in efficiency. However, much attention has shifted recently to include the properties of the graph with community detection such as, incorporating the topic of discussion within the community using LDA topic model [23] or Group-Topic model [24], incorporating Geo-location of users to detect Geo-Social communities [25], or incorporating features such as interaction (mentions, retweets) to discover active communities [26].

Community evolution has been studied in the literature to show different states of a community in regards to the structure [6]. The states are: birth, grow, continue, merge split, death. However, all these studies are limited to the structure of the community. There is a limited study in the literature that focuses on showing the change of the community sentiment over time and the impact of expressing sentiment on the future interactions and the lifetime of the community. One study by Thelwall et al. [27] studied the evolution of topic change in online discussions and used it to predict future Amazon sales. Their work used a manual selection of online blogs on about 340 topics and a time series data that mention a topic on a daily basis.

III. Framework

In this section, we introduce our framework which includes a description of Twitter dataset and the preprocessing steps for collecting and filtering the datasets.

A. Dataset

Twitter social network was the logical choice for our study for three reasons. First, it is publicly available and free to collect data using Twitter API from Twitter Inc. [9]. Second, it is rich in active users around 330 million monthly active users and about 500 million tweets per day. Third, it is

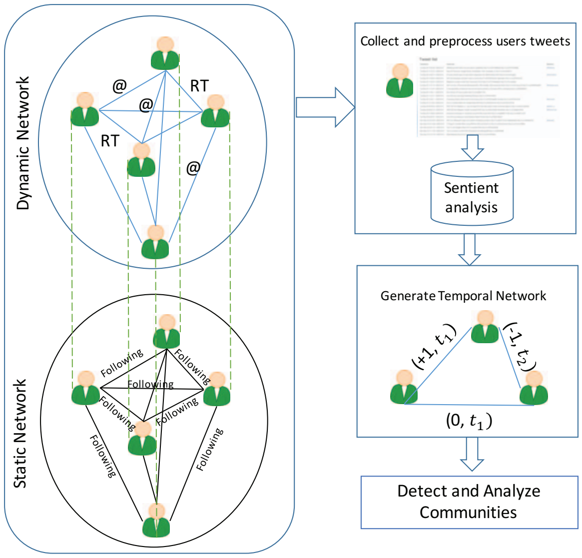

Figure 1. Framework

rich in user attributes and networks features which allow us to construct two different types of networks: static and dynamic. The static network consists of users and following-follower links and can be modelled as a graph where the users are represented as nodes and links are presented as edges connected these nodes. It is categorised as static because the links are stable for a long time unless a user closed his/her profile or decided to unfollow the other user. The dynamic network consists of users and their communications between others using twitter features, which are mentions '@' and retweet 'RT'. In the next section, we show how to construct both graphs and detect the communities.

B. Preprocessing

The framework of our study is presented in Fig. 1. The first step is to construct the static network using the following-follower relationships in Twitter. We randomly select a seed user and crawl the users following-follower links and go to the other users in the following list using the Breadth-First Search method to create the static network. Second, for each user in the static network, we crawl their tweets, analyse them to construct their dynamic network which is based on the actual communication between users at different times. We represent the networks as a graph where the nodes represent users, and the edges represent the following relationship in the static network or communication in the dynamic network. After that, we analyse the communications that occurred in the dynamic network to get the time and the sentiment of all interactions (Algorithm 1).

IV. Experiment

In this section, we show the statistics of our datasets, the analysis and results of our study.

Algorithm 1 Preprocessing

Require: G(V,E), M scale of time(days, weeks, hours), tweets
  $T \leftarrow [\,], S \leftarrow [\,]$ $\{T$ is a list of all the timestamps, S is list for all the sentiments$\}$
  for $\forall\ v \in V$ **do**
    User Tweet $\leftarrow$ Get (v, tweets)
    $T[t] \leftarrow$ Get (User Tweets, M)
    $S[s] \leftarrow$ Analyse Sentiment (User Tweets)
  end for
  $update\overline{G} \leftarrow V, E, S, T$
  return $\overline{G}(V, E, S, T)$ $\{$ $\overline{G}$ is a dynamic graph$\}$

Datasets	Nodes	Edges	Max(k)-core	Dates	Timestamps
Twitter_1	25,237	33,534	11	2013-2015	163
Twitter_2	24,352	45,839	27	2013-2015	145
#WorldCup	208,910	265,356	6	(3-5)July-2018	30

Table I
Overview of the datasets used in the experiment

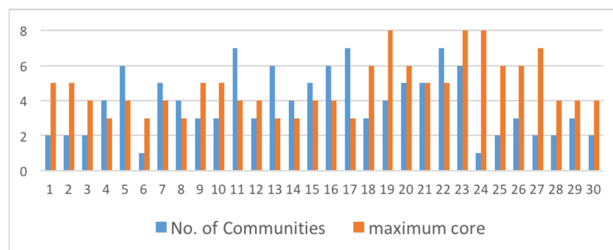

Figure 2. #WorldCup communities numbers and $Max(k)$-core numbers at each timestamps

A. Experiment Setup and Datasets

This experiment was carried out on Windows 7, a 64-bit operating system, IntelCore i7 CPU 3.4GHs with 16GB RAM. We extracted three different datasets from Twitter. The statistics of these datasets is shown in Table I. Twitter_1 and Twitter_2 have been collected in June 2015, and the M parameter is set to a weekly basis for Twitter_1 and Twitter_2. Thus, each timestamp represents a week-long of communications. The edges in both Twitter_1 and Twitter_2 represent interactions at different times with different sentiments between users with a following-follower relationship. That means users in these datasets have following-follower links and they also communicate with each other. The reason for choosing these datasets is to track them for an extended period and observe the changes in the sentiment they share over time. Note that $\text{Twitter}_1$and Twitter_2 do not overlap and they vary in size. The third dataset has been collected based on communications between users who were involved the hashtag #WorldCup in their tweets over two days where M is set to the hourly basis which means each timestamp is an hour long of communications. The purpose of having this dataset is to compare the evolution of sentiments in a different type of network and a different time scale as well.

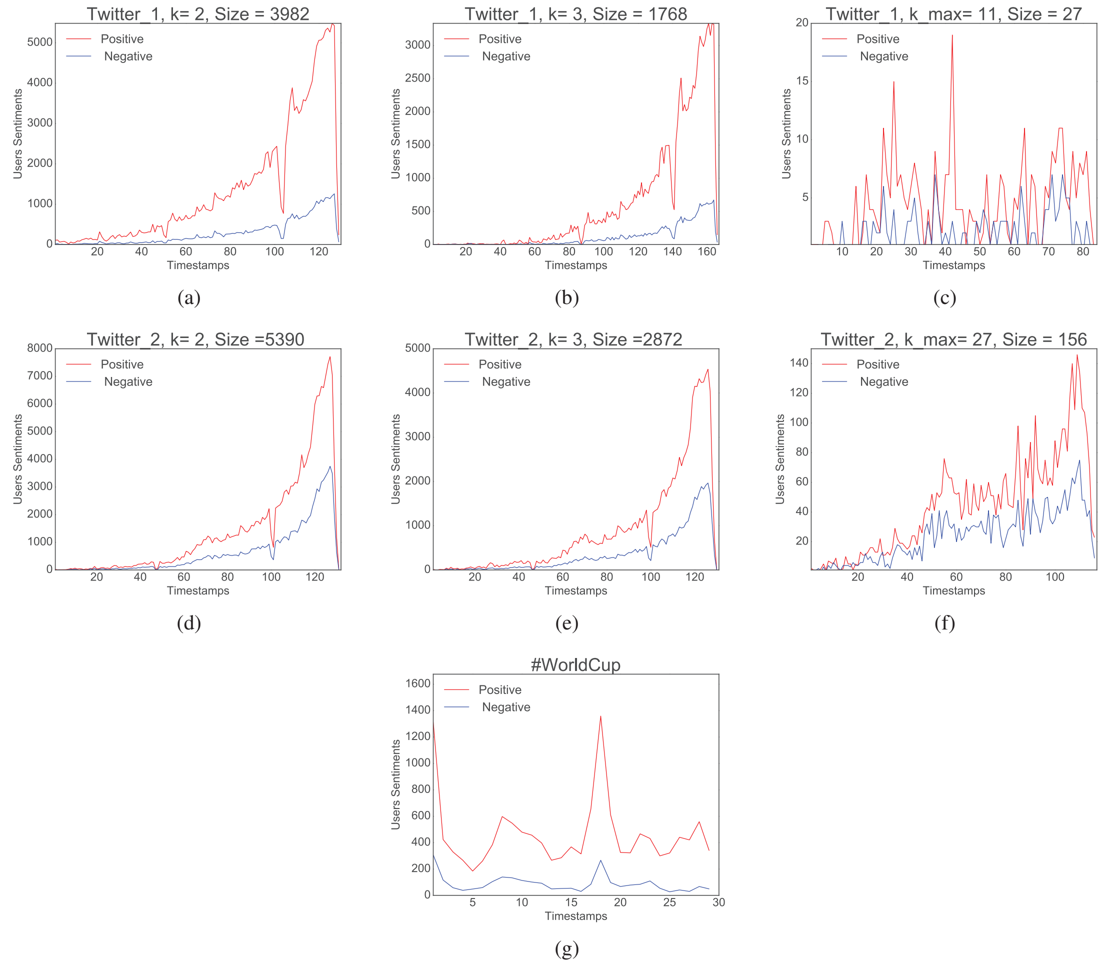

Figure 3. Sentiment Evolution in Twitter Communities

Datasets	k	Size
Twitter$_1$	2	3982
	3	1768
	Max= 11	27
Twitter$_2$	2	5390
	3	2872
	Max= 27	156

Table II
OVERVIEW OF THE COMMUNITIES CORES AND SIZE FOR TWITTER DATASETS BASED ON FOLLOWING RELATIONSHIPS

B. Analysis and Results

We analysed sentiment evolution for all communities in datasets Twitter$_1$, Twitter$_2$ and #WorldCup dataset. Table II shows some of the detected communities core numbers, and their sizes in Twitter$_1$ and Twitter$_2$. Because the number of communities and the Max(k)-core number for #WorlsCup is unstable, we show these changes at each timestamp in Fig. 2.

From Fig. 3 we noticed two things: first, the positive sentiment increase over time and this indicates that users with the following relationship tend to share positive tweets with each other. The second thing we noticed is that the positive sentiment is shared more than the negative sentiment in all communities, both static and dynamic. These two facts indicate that positive sentiment is more likely to get reposted and get a response more than negative opinions because of the increase of positive sentiment over time and that positive sentiment is contiguous in twitter communities. For the dynamic communities, the situation is a little different. The positive sentiment does not increase over time because the users of the community may not interact again after few hours. However, the positive posts are shared more than the negative ones.

Our results also showed an evolution of states for the sentiments in Twitter communities, (Fig. 3), and can be categorised as follows:

- **Birth** of a community happens when the community establishes a communication that carries a sentiment (+, - or 0).
- **Continuous** state indicates that the members of a community communicate in continuous timestamps.
- **Growth** of the community occurs when the sentiment and interactions increase which affect the increase of

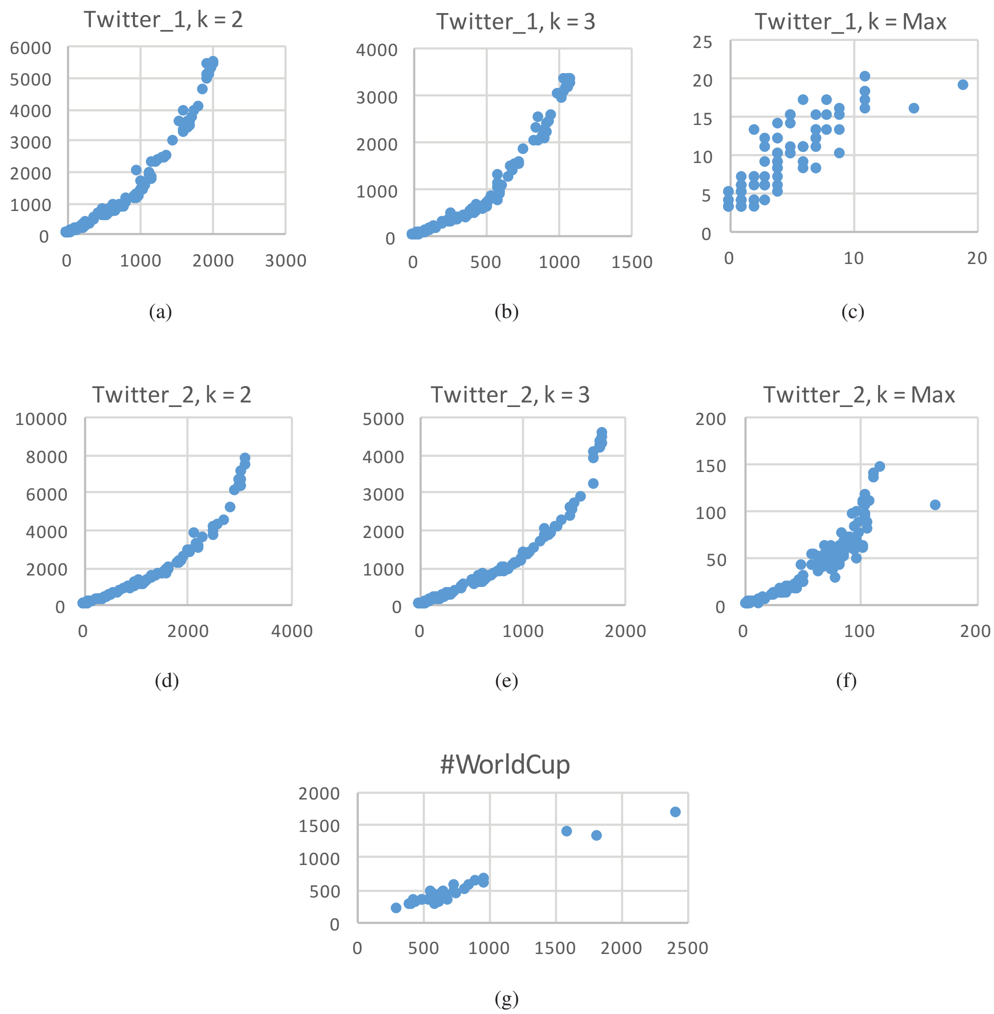

Figure 4. Correlation between Postive sentimetnt and Community Size

the positive sentiments.

- **Split** state happens when a community splits into multiple communities; we noticed that it is more likely to have a large size community with mixed sentiments and small communities with positive sentiment only.
- **Merge** state occurs when the split communities merge into one community with mixed sentiments in the next timestamp.
- **Dissolve** state happens when the community stops interacting and sharing any positive or negative sentiment.

Fig. 4 shows the correlation between the positive sentiment and the community size. There is a significant correlation between the size and positive sentiment, the bigger the size, the more positive sentiment.

V. Discussion

Our results indicate that the majority of communities tend to share positive tweets than negative tweets and the positive ones are more likely to get replies and retweeted. It has been shown in the literature as well that positive sentiments are shared more than the negative ones [28]–[30].

We also found a correlation between the size of the community and the number of the shared positive sentiments within the community. This indicated that the bigger the size, the more positive is the sentiment. We noticed that the bigger the community we detect, the more positive sentiment is being shared and retweeted. This result is useful because significant communities attract social network companies for their viral marketing and recommendation systems.

In the literature, there is an indication that the users with maximum core numbers are more likely to be influential in communities [31]. Influential users may influence others opinions and emotions as well. Therefore, we looked at the community with the maximum core number, Fig. 3(c) and Fig. 3(f). We noticed that they shared more positive sentiments than the negative sentiments as well. This might be one of the reasons that the other members of the community have been influenced by them, and shared more positive sentiments with each other.

It appears that the sentiment is contagious within a community because positive sentiment increases over time. The opinions and the emotions of one user spread to their following links, i.e. both direct and indirect links. This observation has also been indicated in the literature by Kramer et. al [21]. Moreover, we noticed that users tend to increase their interactions over time with their following-follower relationships which have more positive sentiments than negative ones. This fact indicates that there is a homophily relationship which is why users have the agreement with their following-follower relationships which makes their communication easier [32], [33]. We call these types of communities like-minded communities because they share similar sentiments and their interactions increase over time. Like-minded communities are useful for targeted marketing and recommendation systems. For example, Facebook social network discovers like-minded groups by using their friendship links and their likes on different pages to recommend new pages that they most probably will like them too [10]. Also, like-minded communities attract lots of companies to market and recommend their product/services because having someone opinion from this group regarding their product is more likely to influence others.

VI. Conclusion

In this paper, we showed the evolutions of sentiments in Twitter communities and its impact on the community behaviour, birth, growth, continue, split, merge and dissolve. We also showed that positive sentiment is shared more than negative sentiment within k-core communities and there is a strong correlation between the positive sentiment and the size of a community. In the future, we plan to develop a method to discover like-minded communities in social networks based on their sentiments.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful and valuable comments.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[2] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[3] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[4] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[5] V. Batagelj and M. Zaversnik, "An o (m) algorithm for cores decomposition of networks," *arXiv preprint cs/0310049*, 2003.

[6] P. Bródka, S. Saganowski, and P. Kazienko, "Community evolution," *CoRR*, vol. abs/1605.00069, 2016. [Online]. Available: http://arxiv.org/abs/1605.00069

[7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 12, pp. 1–135, 2008. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[8] S. Aslam, "Twitter by the Numbers: Stats, Demographics and Fun Facts." [Online]. Available: https://www.omnicoreagency.com/twitter-statistics

[9] Twitter Developers,"Twitter Developer Platform." [Online]. Available: https://developer.twitter.com/en.html

[10] F. Zhang, Y. Zhang, L. Qin, W. Zhang, and X. Lin, "When engagement meets similarity: efficient (k, r)-core computation on social networks," *Proceedings of the VLDB Endowment*, vol. 10, no. 10, pp. 998–1009, 2017.

[11] F. Zhang, Y. Zhang, L. Qin, W. Zhang, and X. Lin, "Finding critical users for social network engagement: The collapsed k-core problem." in *AAAI*, 2017, pp. 245–251.

[12] R.-H. Li, L. Qin, J. X. Yu, and R. Mao, "Influential community search in large networks," *Proceedings of the VLDB Endowment*, vol. 8, no. 5, pp. 509–520, 2015.

[13] K.-T. Chen and C.-L. Lei, "Network game design: Hints and implications of player interaction," in *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*. ACM, 2006, p. 17.

[14] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey *et al.*, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.

[15] G. Yuan, P. K. Murukannaiah, Z. Zhang, and M. P. Singh, "Exploiting sentiment homophily for link prediction," in *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 2014, pp. 17–24.

[16] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization," *Future Generation Computer Systems*, vol. 78, pp. 430 – 439, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17304077

[17] D. M. Romero, C. Tan, and J. Ugander, "On the interplay between social and topical structure," in *ICWSM*, 2013.

[18] E. Choo, T. Yu, and M. Chi, "Detecting opinion spammer groups through community discovery and sentiment analysis," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2015, pp. 170–187.

[19] R. Lin and S. Utz, "The emotional responses of browsing facebook: Happiness, envy, and the role of tie strength," *Computers in human behavior*, vol. 52, pp. 29–38, 2015.

[20] A. D. Kramer, "The spread of emotion via facebook," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 767–770.

[21] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[22] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," vol. 435, no. 7043, pp. 814–818, 2005.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 248–256.

[24] X. Wang, N. Mohanty, and A. McCallum, "Group and topic discovery from relations and text," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 28–35.

[25] H. Yin, Z. Hu, X. Zhou, H. Wang, K. Zheng, Q. V. H. Nguyen, and S. Sadiq, "Discovering interpretable geo-social communities for user behavior prediction," in *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*. IEEE, 2016, pp. 942–953.

[26] N. Alduaiji, A. Datta, and J. Li, "Influence propagation model for clique-based community detection in social networks," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 563–575, June 2018.

[27] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 406–418, 2011.

[28] A. L. F. Alves, C. de Souza Baptista, A. A. Firmino, M. G. de Oliveira, and H. F. de Figueirêdo, "Temporal analysis of sentiment in tweets: a case study with fifa confederations cup in brazil," in *International Conference on Database and Expert Systems Applications*. Springer, 2014, pp. 81–88.

[29] A. Gruzd, S. Doiron, and P. Mai, "Is happiness contagious online? a case of twitter and the 2010 winter olympics," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 2011, pp. 1–9.

[30] C. Byun, H. Lee, Y. Kim, and K. Ko Kim, "Twitter data collecting tool with rule-based filtering and analysis module," *International Journal of Web Information Systems*, vol. 9, no. 3, pp. 184–203, 2013.

[31] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

[32] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[33] H. Chen, H. Yin, X. Li, M. Wang, W. Chen, and T. Chen, "People opinion topic model: Opinion based user clustering in social networks," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 1353–1359.