Next Topic Recommendation for Influencers on Social Media

Masafumi Iwanaga
Kyoto University
Kyoto, Japan
iwanaga@dl.soc.i.kyoto-u.ac.jp

Keishi Tajima
Kyoto University
Kyoto, Japan
tajima@i.kyoto-u.ac.jp

Yoko Yamakata
The University of Tokyo
Tokyo, Japan
yamakata@mi.u-tokyo.ac.jp

***Abstract*—To maintain popularity on social media over the long term, many users need to shift to a new topic instead of sticking to one topic. When selecting a new topic, a user needs to consider both its popularity on the entire social media and its popularity among the current followers. The former affects the expected number of new followers, and the latter affects the ratio of the current followers the user can retain. The timing is also important. The user should change to a new topic before losing the current followers. If the user change the topic after losing the followers, it is more difficult to obtain new followers. In this paper, we introduce a new task based on these observations: recommending appropriate new topics for currently popular social media users at appropriate timing. As an example of opportunities in the research on this task, we also propose a method of predicting the popularity a given user would gain after shifting to a given new topic. Our method predicts it based on the similarity between the user's current topic and the given new topic. In our experiment with data collected from Twitter, our method improves the prediction accuracy compared with the baseline method.**

***Index Terms*—social network, Twitter, topic shift, topic selection, popularity prediction, retweet prediction**

I. Introduction

On today's social media, such as Twitter, there are many individuals who gain popularity by disseminating information on specific topics, such as video games and movie series. In order for such individuals to maintain long-term popularity, they need to periodically switch to new topics, instead of sticking to the same topic, e.g., a specific video game. It is because: (1) most topics become less popular over time, and (2) the more number of fans they have acquired, the less number of potential fans remain.

When such users change their topics, they want to choose a topic that is likely to gain high popularity, but simply selecting a topic that is currently popular on the social media is not necessarily the best strategy for users who currently have many followers. If the topic is popular on the entire social media, the user can expect many new followers, but if the topic is not popular among the current followers, the user will lose many of them, and need to gain a completely new set of followers from scratch. On the contrary, if the topic is popular among the current followers, the user can retain many of them, but if it is not very popular on the entire social media, it is difficult to obtain new followers.

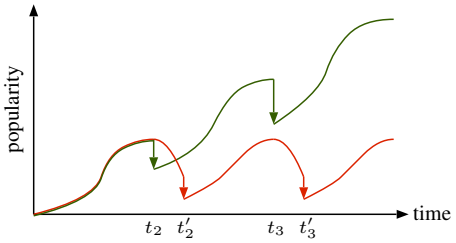

Fig. 1. Change of a user's popularity over time when switching to new topics at the ideal timings (green line) and when switching too late (red line).

Of course, if there is a topic that is popular on the entire social media and also popular among the current followers, the user can simply select it, but such a topic does not always exist, and users often have the trade-off between these two factors. In such cases, if we can provide with some good recommendation system, it would be very helpful for the users.

When to change to a new topic is also an important and difficult decision. On social media, rich-get-richer effect exists. It is difficult for a newcomer to obtain many followers, but once a user has obtained some popularity, it becomes easier to obtain more followers. Therefore, once a user has become popular by posting information on some topic, it is basically a good strategy to keep focusing on that topic. However, many topics become less popular over time. When the current topic has become less popular, the user starts to lose the current followers. The user should change to a new topic before losing many followers. Once the user has lost most followers, it becomes very difficult to obtain many followers again. If the user changes to a new topic before losing the current followers, the user have more chances to have the posts on the new topic viewed by many followers. If some followers forward the posts, the user also has chances to obtain new followers.

Figure 1 illustrates the case where a user changes to new topics at the ideal timings (green line), and the case where changes to new topics are too late (red line). The x-axis is time and the y-axis is popularity, e.g., the number of followers. The user first starts to post on some topic and obtains popularity. The rate of the popularity growth first increases, but when the topic has become less popular, the growth stops. In the green line, the user soon changes to a new topic at t_2. Because of the change of the topic, the user loses some of the followers (the

green down arrow to the left), but still has many followers, and starts to obtain popularity in the new topic. When the second topic has also become less popular, the user soon changes to the third topic at t_3.

By contrast, in the red line, the user changes to the second topic at t_2' after the first topic has become even less popular and the user has lost many followers. The user then loses some more followers because of the topic change (the red down arrow to the left). The user then obtains new followers in the second topic, but the growth is slower than that of the green line because the user has to start from a small number of followers. In addition, the user has shorter time until when the second topic becomes less popular because the user started later. Because of these two reasons, the second peak of the red line is far below the second peak of the green line at t_3. A similar process is repeated for the third topic.

How to smoothly change to a new topic is yet another problem. A user does not need to completely stop posting on the old topic and fully switch to the new topic. A user should rather posts on the old and new topics in parallel for some period so that the user has more chances to have posts on the new topic viewed by the current followers. If the user posts many posts on the new topic before losing the current followers, the user has more chances, but if the user posts only on the new topic, the user will lose the current followers faster.

In summary, new topic recommendation for social media users who are currently popular in some topic is an interesting research topic including (at least) the following research problems.

1) The selection of the best new topic. We should consider both the popularity on the entire social media, and the popularity among the current followers. The former affects the expected number of new followers, and the latter affects how many of the current followers can be retained.
2) The selection of the best timing for topic changes. If it is too late, the user will lose many of the current followers and it becomes difficult to obtain popularity again.
3) The best strategy for smoothly shifting to the new topic. The goal is to maximize chances to have posts on the new topic viewed by the current followers.

Possible approaches to these research problems includes but not limited to those listed below.

- Comparison of topics that have been popular on the entire social media and among the current followers.
- Comparison of topics that were popular in the past and the candidate new topics.
- Analysis of temporal changes of popularity of the current and newly introduced topics.
- Comparison of current followers in favor of the newly introduced topic and those who are not.

II. Opportunities: new topic popularity prediction

As an example of the opportunities in the research on the topic explained above, we show a simple method for predicting the popularity the given user would gain if the user switches to the given new topic. We assume X (Twitter) as the target social media, and measure the popularity by the number of favorites (likes) and retweets. Therefore, our goal is to develop a method for predicting the number of favorites and retweets the user's tweet on the new topic would gain.

When the current followers have already exposed to the new topic sufficiently, it is not difficult to measure its popularity among them. However, when a topic is relatively new, it is not evident how to predict its popularity among the current followers. Our method is based on the following two hypotheses:

- The more popular a user were on the old topics, the more popular will be the user on the new topic.
- The more similar the old and new topics are, the more popular will be the user on the new topic.

The first hypothesis is based on two observations. First, users whose tweets on the old topics were popular must be good at writing interesting tweets, and therefore their tweets on the new topic have the potential to become popular. Second, when a user with many followers starts to tweet on a new topic, the user's tweets are viewed by many followers, and the user has greater chances of having them interested in the new topic. In addition, if some followers retweet some tweets, the user have chances to attract new followers.

The reason of the second hypothesis is as follows. When we focus on a user who has gained popularity by tweeting on a specific topic, most followers follow the user because they are interested in the topic. Therefore, if the new topic is similar to the old topic, the current followers are more likely to be interested in it.

After tweeting on a new topic for the first time, the user may tweet on the old and the new topics in some ratio as explained before. We, therefore, divide the process of topic change into two phases: a phase where a user tweets on a new topic for the first time, and a phase where the user continues to tweet both on the old and the new topics for a while. In this paper, we focus on the former. Our method predicts the popularity gained by the tweets on the new topic during the short period immediately after the user has tweeted on the new topic for the first time.

In order to evaluate the proposed method, we conducted experiments using data collected from Twitter. We collected users who recently shifted to new topics, and predicted the popularity (the number of favorites and retweets) of their tweets on the new topics by using our method and a baseline method. We then evaluated the prediction accuracy by comparing the prediction results with the popularity the tweets had gained in reality.

We run our experiment on two data sets, and on both data sets, our method outperformed the baseline method when users have sufficient popularity. On the other hand, our method is sometimes outperformed by the baseline for users without many favorites and retweets. This is an expected result. When a user is not very popular, the followers are mostly personal friends. As a result, the popularity gained by the tweets on the new topic does not depend much on the topic. However, our

main target is users who currently have some popularity and want to change the topic in order to maintain it. Our method outperforms the baseline method for such users.

III. Related Work

There have been much research on topic recommendation for "consumers" on social media, i.e., users who are looking for interesting topics to receive information. However, to the best of our knowledge, there have been no academic research on topic recommendation for "influencers", i.e., users who post information on a specific topic and have obtained popularity.

Several studies have proposed methods for detecting trending topics on Twitter [1]–[3]. When a user changes to a new topic, the user should consider both its popularity on the entire social media and its popularity among the current followers as explained before. These methods can be used for estimating the former.

Personalized topic recommendation has also been extensively studied [4], [5]. For example, Cataldi et al. [4] proposed a method for personalized topic recommendation for a specific user or community. Their method generates keywords that can search for topics that are most relevant to the given users by analyzing the content of the users' own tweets. Cataldi et al. [5] extended this approach by analyzing the topology of the Twitter user graph. These methods, however, can only detect topics that are emerging within the given community, and cannot predict the popularity of a topic that is completely new within the given target users. Our purpose is to predict the popularity of a topic that exists somewhere else in Twitter but has not been introduced into the target community.

There have also been studies on predicting the popularity of a single tweet [6]–[8]. These methods, however, also assume tweets on some existing topics, and cannot be applied to the prediction of the popularity of newly introduced topics.

Several studies have proposed methods to predict the final number of retweets of a tweet using the number of retweets observed within a short period after it is tweeted [9]–[11]. These methods predict the popularity of a tweet only after it has been tweeted, and cannot be used to choose a new topic before tweeting on it. Can et al. [12] showed a method to predict the total number of retweets without using post-tweet data. It uses image features in addition to the ordinary features such as the number of followers. However, their method can achieve high accuracy only for tweets with images.

There have been a few studies on the prediction of other types of users' responses to a new topic on Twitter. Ren and Ye [13] proposed a method to predict a user's opinions on new topics by using the user's follow/follower structure and the similarity of the topic with the old topics tweeted by the user. Their method is similar to ours in that they use topic similarity to predict users' responses to a new topic, but they focus on users' opinions, not the popularity. Wu et al. [14] developed a system for visualizing the flow of opinion propagation, which can also be used to predict the information diffusion pattern for a new topic. However, their system does not predict the popularity of a new topic.

IV. Proposed Method for Popularity Prediction

In this section, we describe the details of the method we explained in Section II. We first explain how we collect the data used by the method, and introduce some notations.

A. Data Collection Step

We collected data on users who have recently shifted to a new topic through the following steps. Let N_j $(1 \leq j \leq m)$ denote m new topics. (Selection of N_j used in our experiment will be explained in Section V.)

1). We collect tweets on each N_j by using the main hashtag used by Twitter users for representing N_j.
2). We obtain U_j, a set of users who have tweeted at least one of the tweets collected above for N_j.
3). For each user in U_j, we collect their recent tweets starting from the most recent ones. Let T_j denote the set of all tweets collected from all users in U_j.
4). From T_j, we extract T_j^N, a set of tweets related to the new topic N_j, by using the corresponding hashtag. Note that we extract T_j^N only for N_j, and we do not extract N_k $(k \neq j)$ from T_j. The remaining tweets $T_{*,j}^O = T_j \backslash T_j^N$ are regarded as tweets on some of the old topics by users in U_j. We use the subscript $*$ because we later divide $T_{*,j}^O$ into $T_{1,j}^O, \ldots, T_{n,j}^O$ corresponding to n old topics.
5). We repeat the steps above for all $j = 1, \ldots, m$, and let $T_{*,*}^O = \bigcup_j T_{*,j}^O$. The obtained $T_{*,*}^O$ is the set of tweets on the old topics by all users in any of $U_1, \ldots, U_m$.

We then exclude the following two types of data.

- Because we target users shifting from old topics to a new topic, we exclude users for whom we could extract no old topic. That is, if the oldest u's tweet in T_j is also in T_j^N, we exclude u from the dataset.
- We also exclude tweets posted by a user u later than the u's first tweet on N_j and not in T_j^N. We predict the popularity of tweets on the new topic by using the popularity of the u's earlier tweets. Therefore, tweets satisfying this condition are neither the prediction target nor used for the prediction.

Next, we extract old topics of all users in $\bigcup_j U_j$ by clustering the tweets in $T_{*,*}^O$. We first calculate TF-IDF vectors of the tweets. We calculate $w(t, d)$, TF-IDF value for a term t in a tweet d, by the formula below:

$$w(t,d) = tf_{t,d} \cdot (\log \frac{|T_{*,*}^O|}{n_t} + 1)$$

where $tf_{t,d}$ is the number of occurrences of t in d, and n_t is the number of tweets in $T_{*,*}^O$ including t. We then cluster $T_{*,*}^O$ into n clusters by applying k-means algorithm to the TF-IDF vectors defined above. We regard the obtained clusters $T_{1,*}^O \ldots T_{n,*}^O$ as representing n old topics $O_1, \ldots, O_n$ for all users. Let $T_{i,j}^O = T_{i,*}^O \cap T_{*,j}^O$. $T_{i,j}^O$ is the set of tweets on the old topic O_i tweeted by a user who later tweeted on the new topic N_j.

TABLE I
SUMMARY OF FEATURES

Structure-Based	#follower	(log scale)
	#friend	(log scale)
	followerFriendRatio	
	#status	
Tweet-Based	#hashtag	
	#URL	
	#mention	
	tweetLength	
	expected favorite count $E(f)$	(log scale)
	expected retweet count $E(r)$	(log scale)

B. Topic Similarity

We next define $Sim_{i,j}$, an asymmetric similarity of an old topic O_i and a new topic N_j, as follows:

$$Sim_{i,j} = \frac{|U_{i,*} \cap U_j|}{|U_{i,*}|}$$

where $U_{i,*}$ is the set of users who tweeted tweets in $T^O_{i,*}$. U_j is the set of users who tweeted on N_j as defined before. $Sim_{i,j}$ is the proportion of the users who tweeted on N_j among the users who tweeted on O_i. We use $Sim_{i,j}$ to approximate the probability that a current follower of a user in U_j is also interested in the new topic N_j. It is based on the assumption that a user who tweeted on a topic is interested in that topic.

In our definition, even if two topics are not conceptually similar, we regard them as similar if the user sets interested in them are similar. In addition, if the new topic N_j becomes popular, the similarity between N_j and the old topics tend to increase. These properties are suitable for us. Also note that our similarity is time-dependent because we use the information on recent tweet data.

C. Linear Regression

We use standard linear regression for predicting the number of favorites and retweets gained by a given user's tweet on a given new topic. We use the following two types of features for the regression (see Table I).

Structure-Based Features: The features #follower and #friend are the number of followers and friends (i.e., followees), respectively. We transform their values into the log scale because they follow power law distributions. The feature followerFriendRatio is the ratio of the number of followers to the number of friends, and #status is the total number of tweets posted by the user before the first tweet on the new topic.

Tweet-Based Features: The features #hashtag, #URL, #mention, and tweetLength are the mean of the number of hashtags, URLs, mentions, and characters, respectively, in the user's tweets on the new topic. The expected favorite count $E(f_j)$ is a feature we propose and is calculated by:

$$E(f_j) = \sum_{i=1}^{n} \left(\frac{|T^O_{i,j}|}{|T^O_{*,j}|} \cdot Sim_{i,j} \cdot f_{i,j} \right) \quad (j = 1, \ldots, m)$$

where $f_{i,j}$ is the mean of the number of favorites to the tweets in $T^O_{i,j}$. $E(f_j)$ is the weighted average of the product of $f_{i,j}$ and $Sim_{i,j}$ over the n old topics, where the weights are the proportion of each old topic O_i in $T^O_{*,j}$. We use the product of $f_{i,j}$ and $Sim_{i,j}$ because we hypothesized that the popularity of tweets on the new topic is proportional to each of these two factors. We define the expected retweet count $E(r_j)$ similarly:

$$E(r_j) = \sum_{i=1}^{n} \left(\frac{|T^O_{i,j}|}{|T^O_{*,j}|} \cdot Sim_{i,j} \cdot r_{i,j} \right) \quad (j = 1, \ldots, m)$$

where $r_{i,j}$ is the mean number of retweets of the tweets in $T^O_{i,j}$. Because $f_{i,j}$ and $r_{i,j}$ also follow power law distributions, we transform $E(f_j)$ and $E(r_j)$ into the log scale. Note that tweet-based features are related to tweets on the new topic, but their values are known before we post them.

The target variables are f_j and r_j, which are the mean number of favorites and retweets for tweets on the new topic N_j. These values are also transformed into the log scale.

V. EXPERIMENTS

We conducted experiments using two datasets collected from Twitter. In the first dataset, 11 video games released after November 2019 were used as the new topics. We collected 10,176 users who tweeted on any of these games, and collected 1,000 tweets (or less when the user has less than 1,000 tweets) of each user. In the second dataset, 15 animated TV programs broadcast after January 2020 were used as the new topics. We collected 9,028 users who tweeted on them, and collected up to 1,000 tweets of each user. The number of old topics, n, was set to 20. We generate TF-IDF vectors of tweets by using top 1,500 frequent nouns.

The baseline is linear regression where $E(f_j)$ and $E(r_j)$ are replaced with the mean number of favorites and retweets of the user's past tweets. We evaluate the prediction accuracy by RMSE.

Figure 2 shows the results with the first dataset. We predict f_j and r_j, the mean number of favorites and retweets for tweets on each user's new topic N_j. The vertical axis of the graph represents RMSE, and the horizontal axis, LowerLimit, is the lower limit of the f_j values of the users we include in the evaluation. For example, LowerLimit = 5 means we exclude the users whose f_j was less than 5 when estimating the regression parameters and evaluating the prediction accuracy. The blue and orange lines show RMSE of the prediction by the baseline method and the proposed method, respectively.

For both f_j and r_j, the proposed method is superior to the baseline except when the LowerLimit is very small. When it is very small, many users without many followers are included in the evaluation. The followers of such users are mostly personal friends, and the number of favorites and retweets by them does not depend on the similarity between the old and new topics. As a result, our method does not outperform the baseline.

Figures 3 shows the result for the second dataset. In both graphs, the proposed method outperforms the baseline when LowerLimit is sufficiently large, although the margin is small for f_j on this dataset compared with the other three cases.

Our target is the users who have some popularity and want to change topics to maintain it. On both datasets, our method

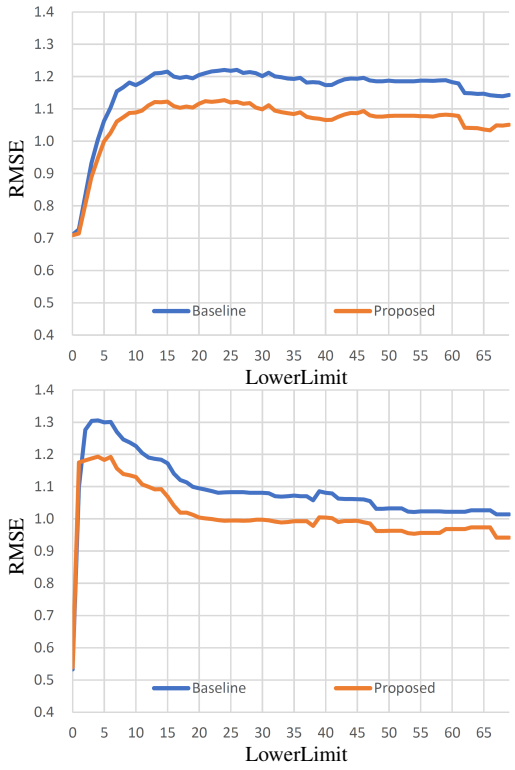

Fig. 2. RMSE of f_j (left) and r_j (right) prediction for video games.

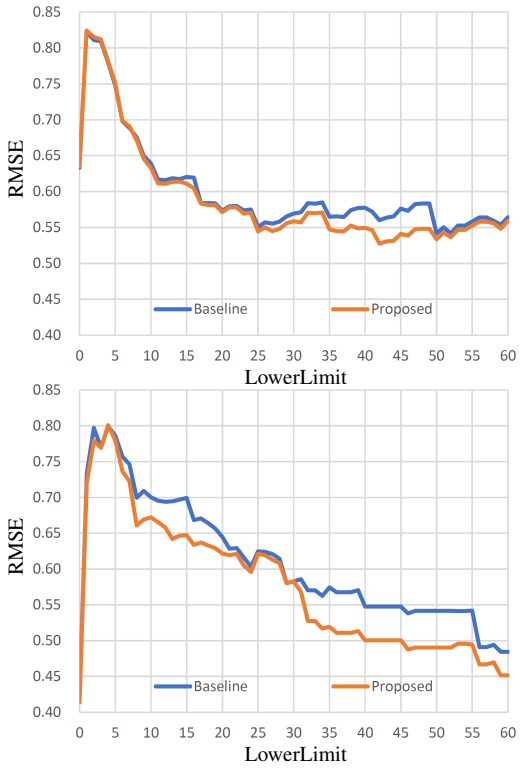

Fig. 3. RMSE of f_j (left) and r_j (right) prediction for animated TV programs.

outperformed the baseline method for users who had more than 5 to 10 favorites or retweets for one tweet on average.

VI. Conclusions

In this paper, we introduce a new task of recommending the next topics to users who has gained popularity on social media by posting information on some specific topics. It includes the following interesting research problems: selection of the best next topic balancing the popularity on the entire social media and the popularity among the current followers, selection of the best timing to switch to new topics, and the best strategy for smoothly switching to new topics.

As an example of the opportunities in the research on that task, we also proposed a method of predicting the popularity gained by a user after the user switches to a new topic. It predicts the number of favorites and retweets that a tweet on the new topic will gain. Our method uses the topic similarity between the old topic and the new topic. We compared the accuracy of this method with a baseline method which is simply based on the users' popularity on the old topics. The result of our experiment with two datasets collected from Twitter shows that our method outperforms the baseline when applied to sufficiently popular users. This result suggests that our two hypotheses explained in Section I are valid for users who have sufficient popularity.

References

[1] M. S. C. Sapul, T. H. Aung, and R. Jiamthapthaksin, "Trending topic discovery of Twitter tweets using clustering and topic modeling algorithms," in *Proc. of JCSSE*. IEEE, 2017, pp. 1–6.

[2] J. Benhardus and J. Kalita, "Streaming trend detection in Twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122–139, 2013.

[3] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang, "Topicsketch: Real-time bursty topic detection from Twitter," *IEEE Tansaction on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2216–2229, 2016.

[4] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *Proc. of International Workshop on Multimedia Data Mining*, 2010, pp. 1–10.

[5] M. Cataldi, L. D. Caro, and C. Schifanella, "Personalized emerging topic detection based on a term aging model," *ACM TIST*, vol. 5, no. 1, pp. 1–27, 2014.

[6] Y. Liu, J. Zhao, and Y. Xiao, "C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network," *Neurocomputing*, vol. 275, pp. 733–746, 2018.

[7] X. Tang, Q. Miao, Y. Quan, J. Tang, and K. Deng, "Predicting individual retweet behavior by user similarity: A multi-task learning approach," *Knowledge-Based Systems*, vol. 89, pp. 681–688, 2015.

[8] D. Huang, J. Zhou, D. Mu, and F. Yang, "Retweet behavior prediction in twitter," in *Proc. of International Symposium on Computational Intelligence and Design*, vol. 2. IEEE, 2014, pp. 30–33.

[9] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proc. of CIKM*, 2012, pp. 2335–2338.

[10] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. of WWW*, 2014, pp. 925–936.

[11] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "Seismic: A self-exciting point process model for predicting tweet popularity," in *Proc. of KDD*, 2015, pp. 1513–1522.

[12] E. F. Can, H. Oktay, and R. Manmatha, "Predicting retweet count using visual cues," in *Proc. of CIKM*, 2013, pp. 1481–1484.

[13] F. Ren and Y. Wu, "Predicting user-topic opinions in twitter with social and topical context," *IEEE Transaction on Affective Computing*, vol. 4, no. 4, pp. 412–424, 2013.

[14] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, "Opinionflow: Visual analysis of opinion diffusion on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1763–1772, 2014.