Activity-Based Sampling of Twitter Users for Temporal Prediction Models

Somayyeh Aghababaei, Eren Gultepe, Iuliia Chepurna, Masoud Makrehchi
Department of Electrical, Computer, and Software Engineering
University of Ontario Institute of Technology, Oshawa, ON, Canada
{somayyeh.aghababaei, eren.gultepe, iuliia.chepurna, masoud.makrehchi}@uoit.ca

***Abstract*—Increasingly more applications rely on crowd-sourced data from social media. Some of these applications are concerned with real-time data streams, while others are more focused on acquiring temporal footprints from historical timelines of users. Nevertheless, determining the subset of "credible" users is crucial. While the majority of sampling approaches focus on individuals' static networks, dynamic user activity over time is usually not considered, which may result in activity gaps in the collected data. Models based on noisy and missing data can significantly degrade in performance. In this study, we demonstrate how to sample Twitter users in order to produce more credible data for temporal prediction models. We present an activity-based sampling approach where users are selected based on their historical activities in Twitter. The predictability of the collected content from activity-based and random sampling is compared in a user-centric temporal model. The results indicate the importance of an activity-oriented sampling method for the acquisition of more credible content for temporal models.**

I. Introduction

Twitter's public and open nature provides great opportunities for its users to actively participate in sharing their opinions and produce high quality content that is reflective of their tendencies and preferences in their day-to-day life [1]. This vast amount of publicly available user-generated content is applied to many applications ranging from tracking human social behavior [2, 3] to detecting events of interest [4]. These studies are either concerned with pulling Twitter and aggregating tweets as bulk or tracking content over time in order to find meaningful patterns for targeted events. The main challenge of the former studies is the limitation of the Twitter API in accessing only 1% of all existing tweets. However, despite this limitation, the latter studies are concerned with retrieving historical timelines of users.

To tackle the above issues of retrieving more tweets beyond the 1% threshold and obtaining historical timelines, topic-based sampling and the REST API are both shown to be more effective [5, 6]. In topic-based sampling [7], a set of specific keywords or hashtags are applied to collect tweets through the search API. In fact, using more specific parameters in sampling, such as keywords, provides us with a comparable amount of data from firehouse [8]. A very substantial problem with this group of sampling is that it is limited to the studies around the content of shared topics which is not scalable to many applications. In the case of the REST API, a set of Twitter users are needed in order to retrieve historical tweets. However, the issue of selecting a credible subset of users still remains. Nevertheless, many network-based sampling approaches were studied which focus on sampling a subset of users from their networks [9] or sampling users based on their popularity [10]. The drawback behind network-based sampling is that, a set of users are sampled from a static network while ignoring the availability of their posts over time. In fact, there is no guarantee that sampled users are active on a daily basis which is necessary for temporal models.

In this study, we sample Twitter, whereby we propose an activity-based sampling method to retrieve a selection of users for the REST API. In activity-based sampling, we leverage users profiles to extract their historical activities. The most active users are designated as "credible" users for employing in a temporal prediction model. We address two main characteristics in our sampling model: (a) obtaining the most active users, (b) avoiding missing content or activity gaps over time. The term active users does not refer to celebrities, news agencies, or major companies whose corporate accounts in social media are normally managed by a group of employees.

We gathered two samples of Twitter users; active users using our proposed sampling approach and random users using the Steaming API. Since the Streaming API is widely used approach in many topical and user-based models [11, 12], it is important to understand the effectiveness of the activity-based sampling proposed in this study compared with random sampling. The selected users from both approaches are employed in the REST API to collect their historical tweets. We compare the content of users, selected from both sampling approaches in different aspects, including statistical properties and predictability in a temporal model.

We employ the collected historical content in a temporal user-centric model, which aims to discover conclusions from user-generated data. In user-centric, the content of a set of selected users is aggregated based on user timelines, to extract meaningful patterns with respect to the events of interest. This approach is considered to be a temporal model which suffers from activity gaps or missing data. Therefore, we can evaluate the effectiveness of our proposed sampling compared with random approach in providing more credible content while mitigating the effect of missing content.

II. Related Works

With the increasing number of Twitter users, the amount of aggregated tweets have become overwhelming, thus the selection of a relevant subset of tweets or users is crucial. Many sampling techniques have been previously used ranging from topical [7] to user-based approaches [9]. As dicusssd before, topic-based sampling limits the study around the content of shared topics. The second type focuses on sampling a subset of users from their networks [13]. The drawback of the latter approach is that the availability of users' posts over time is not considered. In fact, there is no guarantee that sampled users are active on a daily basis which is necessary for temporal models where content or user timelines are aggregated considering their timestamps [14, 15]. Therefore, the selection of a proper sampling approach is the primary key for temporal prediction models.

The most common sampling approach is random sampling using the Streaming API, which allows the retrieval of 1% of the real-time data with some specific parameters. There have been many empirical studies dealing with the evaluation of the data sampled from random sampling with other approaches, including random vs firehouse [8]. They discusses the situations in which random sampling has less coverage compared to firehouse. However, when there are more specific parameters such as keywords, random sampling can provide a comparable amount of data as firehouse. In another study [7], the Streaming API was compared with Expert sampling, in which the expert users are those with a high number of followers. In their study, the tweets of expert users were compared with random users in terms of the trustworthiness of their content. It was shown that expert content contains divers and more popular topics, and includes less spam. This has applications in many topical extraction models, such as breaking news detection. We can conclude that expert sampling is rich in content and are more valuable for content-centric models, such as topical models. In fact, the Streaming API preserves the statistics of the sample size as the whole representative sample, but for content-centric models, which can benefit from the context, expert sampling is more superior. Hence, using the Streaming API highly depends on the type of coverage and targeted problems.

Many empirical studies evaluated the effectiveness of expert sampling in many dimensions such as trustworthiness, diversity of discussion topics, or sentiment. However, compared to random users, there are many challenges in utilizing the content of experts, whose corporate accounts in social media are normally managed by a group of employees. In many applications, ranging from content-centric [16] to user-centric [14], crowd opinion collectively provides predictive signals for the prediction models. In fact, by selecting the experts we ignore the valuable content coming from a crowd and we neglect the vast amount of information contributed by citizens.

A vast amount of research prefers network sampling rather than selection of experts based on popularity. In network sampling, a subset of users are chosen from the entire network of collected users. Different techniques have been applied in recent years, of which Random Walk and Breadth-First Search (BFS)[17] are well-known. However, the major problem with these techniques is that they are mostly biased toward high degree nodes similar to expert sampling. A solution to this problem is the traditional Monte Carlo Markov Chain (MCMC), which was proposed by White et al. [9]. They applied a technique based on MCMC and Coupling From The Past (CFTP) to have better convergence in sampling. These methods ignore the activity of users over time, whereas in temporal models, the presence of users over time is mostly needed. In our problem, we can not compare the activity-based sampling with the network sampling. While we are looking for independent opinions, the network sampling (users and their networks) is biased toward the same opinions.

In temporal models, such as detecting targeted events [18], discovering spatio-temporal topics [19, 20], or tracking users behavior over time [21], user activity or content is tracked to extract meaningful signals. Therefore, if there is an activity gap or missing opinions, the performance of both content-centric and user-centric models can be significantly degraded. Although many sampling approaches are presented to select a subset of users or content in static mode, there is still a significant need for a sampling approach to address the temporal aspect of the data. In this study, we leverage user profiles to estimate their activities in the past for the selection of the most active users as opposed to experts users.

III. Twitter Sampling

The objective of this study is to present a sampling approach to collect the best representative users for the REST API. Given a set of users, the REST API provides access to historical timelines, with the limitation of at most 3,200 recent tweets for a single user. The main challenge is how to sample Twitter users to avoid the absence of data in historical tweets. Nevertheless, absent data could be inevitable, users do not necessarily share posts on a daily basis. However, as far as possible, to avoid missing opinions in historical tweets, we address some characteristics for the selection of users. In this method, the interest is to find a set of the most active users while showing no bias toward individuals with a high or low number of tweets. We collect users selected by two different sampling strategies; a random approach using the Streaming API and activity-based sampling, which is based on the historical activity of a user.

A. Random Sampling

As previously discussed, random sampling is the most common approach to access data streams. In order to obtain random users, we gather 1% of tweets using the Streaming API. The historical timelines of the randomly selected Twitter users are later retrieved using the REST API.

B. Activity-based Sampling

In this method, the interest is to find a set of active users while showing no bias toward individuals with a very high or low number of tweets. In our sampling approach, two factors are considered: the period of time a user is active and the daily number of tweets. Since these specifications are not available, we retrieve them from user profiles. In this regard, we applied the Streaming API to access real-time stream of tweets. For each tweet, the user profile of its author is retrieved, which includes some specific elements, such as: $status\_count$, $created\_at$, and $followers\_count$. For each user, two main parameters are calculated as follows:

(a) The number of days a user is active ($days$). In order to identify for how many days a user is active, we calculate the number of days the user's profile was generated ($created\_at$) until the current time ($time\_now$) as follows:

$$days = time\_now - created\_at \tag{1}$$

A long period of activity is a primary criteria for the selection. As we track the content of users over time, users who recently became members are ignored.

(b) The average number of tweets per day ($tweets\_day$): As this parameter is irretrievable, we leverage the total number of tweets and the number of days a user is active

$$tweets\_day = total\_tweets/days \tag{2}$$

Users are considered active if they have a high number of active days ($days$) as well as tweets per day ($tweets\_day$). Active users are classified using $followers\_count$ to filter out accounts belonging to celebrities, news agencies, or major companies .

IV. Temporal User-centric Model

A user-centric prediction model is proposed to evaluate the credibility of the content retrieved from the selected user in order to predict a trend of interest. The user-centric model is inspired from content-based approaches for trend prediction [15] whereby historical tweets, posted earlier, are leveraged to predict the targeted problem. However, in contrast to content-based models, the content of the selected users is aggregated based on user timelines rather than considering the content as bulk.

In this study, our targeted problem is crime trend prediction by leveraging Twitter data. Historical tweets have been shown to be successful in predicting the directions of crime rates [15, 14]. The problem of trend prediction is converted to a binary classification problem where the objective is to detect the directions of the targeted trends (in our case, crime trends). For this classification problem, a set of documents along with their labels is generated.

Let $T_u = \{\left(p_1^{\langle u\rangle}, t_1^{\langle u\rangle}\right), \left(p_2^{\langle u\rangle}, t_2^{\langle u\rangle}\right), ..., \left(p_J^{\langle u\rangle}, t_J^{\langle u\rangle}\right)\}$ denotes the timelines of a user u, where tuple $\left(p_j^{\langle u\rangle}, t_j^{\langle u\rangle}\right)$ represents user u's post j along with its timestamps: $t_1^{\langle u\rangle} < t_2^{\langle u\rangle} < ... < t_J^{\langle u\rangle}$. Post $p_j^{\langle u\rangle} = \{w_1^{\langle u\rangle}, w_2^{\langle u\rangle}, ..., w_K^{\langle u\rangle}\}$, is comprised of tokens $w_k^{\langle u\rangle} \in V$, where $V = \bigcup_k w_k$ is global vocabulary and $k \in [1, K]$. In order to aggregate tweets based on users timelines, we assume an aggregation window in which the timelines are concatenated as follows:

$$\begin{aligned} x_m &= \frac{1}{q}\sum_{j=1}^{q} p_{j-q+1}^{\langle u\rangle}, \quad q = [1, n] \\ z_i^{\langle q\rangle} &= (x_{i,1}^{\langle q\rangle}, x_{i,2}^{\langle q\rangle}, ..., x_{i,M}^{\langle q\rangle}) \\ Z &= \bigcup_{i=1}^{n} z_i^{\langle q\rangle} \end{aligned} \tag{3}$$

where q is the size of the aggregation window and M is the total number of users. Therefore, x_m is the timeline of a user after aggregation and z_i is a document that consists of a series of user timelines.

Let $Y = \{y_1, y_2, ..., y_n\}$ be the targeted time series whose future values are to be predicted. The time series Y is sampled in time steps $t(i)$, $1 \le i \le n$. To convert regression-based prediction into a classification problem, the continuous signal Y has to be mapped into a categorical set, which is defined as a set of labels. There are several techniques to infer the labels from a continuous variable such as quantization or the direction of changes in rates. Due to the nature of the research, we adopt trend analysis of the continuous rates for labeling:

$$l_i = sgn(y_{i+d} - y_i), \quad if \begin{cases} d > 0 : \; lag \\ d \le 0 : \; lead \end{cases}, L = \bigcup_{i=1}^{n-d} l_i \tag{4}$$

where d is the lag from the current state (z_i) and the target label, l_i is the label at $t(i)$ and L is the sequence of the labels in $n - d$ consecutive time steps. In the case of crime rate prediction, the label of the document z_i is the change in crime rate with different lags, $l_i = \{-1, 1\}$. After inferring the labels, a set of annotated documents is generated by associating high dimensional temporal data to one dimensional target labels inferred from the time series of interest, $\forall z_i \in Z, z_i \rightarrow l_i$, where $n - d$ training examples of the form $\{(z_1, l_1), ..., (z_{n-d}, \; l_{n-d})\}$ are generated. Therefore, features vectors are represented as follows:

$$Z^{\langle q\rangle} = \bigcup_{i=1}^{n-d} z_i^{\langle q\rangle} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,M} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n-d,1} & u_{n-d,2} & \cdots & s_{n-d,M} \end{pmatrix} \tag{5}$$

where $s_{i,m}$ is a sentiment of the user m which belongs to document i. Since the idea of this model considers a sample of users representative of the emotion of the all collective users, we selected LIWC [22] to derive sentiments as user-dependent features. We extract the positiveness and negativeness scores. Each user is defined by the normalized mentioned scores.

V. Experimental Results

In this section, we evaluate the feasibility of our sampling approach compared with random sampling in retrieving historical tweets. We begin with comparing statistical characteristics of data collected from both approaches. The intention is to understand how well data are distributed over time. We then evaluate the credibility of the datasets in the proposed temporal model.

Using both random and activity-based sampling, we collected an equal number of users. Chicago was targeted due to its importance as the third populous city in U.S as well as its data portal, which is a rich resource providing all reported incidents on a daily basis. The selected users are fed to the REST API to collect their historical timelines with the limitation of at most 3,200 recent tweets of a user (REST API's limitation). The historical timelines of the selected users were retrieved and restricted to January 1, 2014- August 23, 2015. Crime rates were also retrieved from Chicago Data Portal [1]. Each incident was gathered with its timestamp, its type, and exact location. Three different crime types (Narcotics, Prostitution, and Public violation) as well as the overall crime rates were targeted due to their high frequencies.

A. Comparing Timelines

We compare the number of posts (see Figure 1a) and users (see Figure 1b) observed on a daily basis from both datasets. The historical tweets obtained from active and random users are mapped between our consideration period (the previous 600 days) using their timestamps; we did not include tweets from more than 600 days because of low level activity . Figure 1a presents that the daily number of tweets from active users are higher than tweets of random users. This can be an asset for content-centric models where content is aggregated on a daily basis for a temporal model. Figure 1b shows the daily number of unique users, defined as those who post at least once per day. From Figure 1b, we can observe that the daily number of active users obtained from activity-based is higher than the number of users from random sampling. In user-centric models, the number of available active users plays an important role. In general, activity-based sampling, compared to random, has better coverage in terms of the number of tweets and users (see Table I).

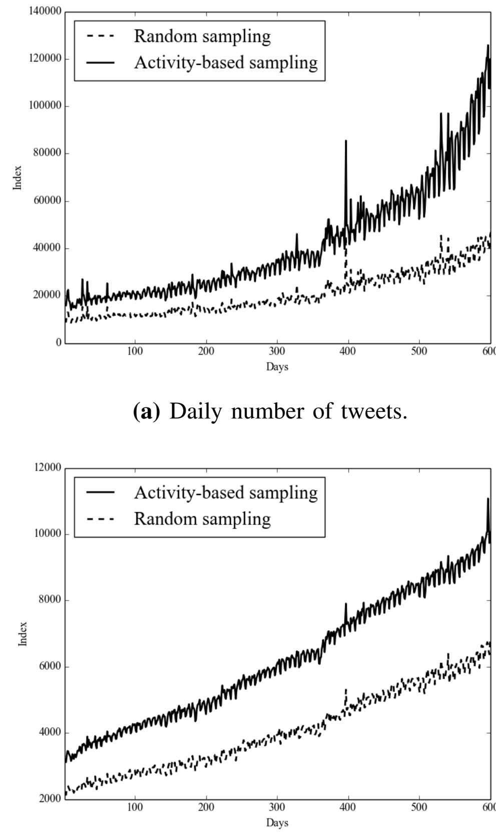
TABLE I: Statistics on the size of users and posts observed on a daily basis for both sampling approaches.

	Daily users		Daily tweets	
	Activity-based	Streaming	Activity-based	Streaming
MIN	3,116	2,128	13,952	8,555
STD	1,987.284	1,326.04	24,061.123	9,135.875
AVG	6,328.71	4,160.23	41,568.725	20,591.25
MAX	11,077	7,131	125,782	45,352

B. Comparing Activity Gap

We also investigate the presence of user activity over time, which is the key element in user-centric approaches. Models directly working with user streams are prone to vast amounts of missing opinions. Although activity gaps are inevitable, it is crucial to retrieve the most active users while avoiding activity gaps in their timelines. Figure 2 shows the daily activity of the top 100 users (the most active) during 600 days. The black

[1]City of Chicago Data Portal: https://data.cityofchicago.org



(a) Daily number of tweets.

(b) Daily number of users.

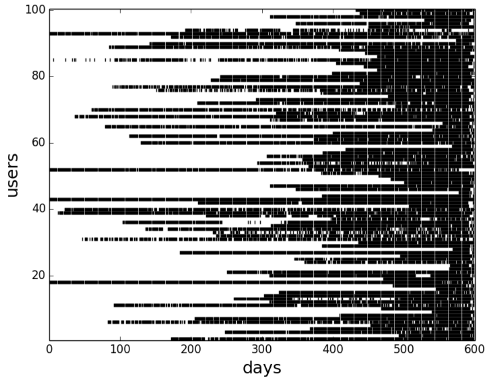
Fig. 1: Daily number of tweets (a) and users (b) captured from activity-based and random users.

vertical bar indicates that a user had at least one post on that day. Ideal data would resemble a black square. It can be observed that users selected by activity-based sampling are more active over time compared to random users.

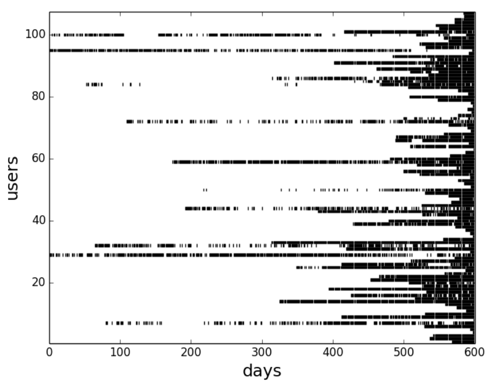
C. Comparing Credibility

The predictability of content captured from both methods is compared in our proposed temporal prediction model. The classifier is linearSVC, which is the implementation of liblinear [23]. LinearSVC is faster compared with LinearSVM, since kernel transforms are not used and it scales better for large datasets in a linear classification problem. The evaluation is processed by calculating the Macro-averaged F1-score and using rolling origin [24] as the common method for training and evaluating the performance of the model for series observations. In this approach, the training set is the first i and it is tested on the $i+1th$ document. In the second iteration, the training set is moved one document forward (the first $i+1$), and it is tested on the $i+2th$ document. This process is continued until all the test data is classified.

The predictability of the proposed prediction model with different lags is examined. In this regard, document z_i which has been generated at time t_i, is labeled with crime trend l_i



(a) Activity-based.



(b) Random.

Fig. 2: Rastergram of daily activity by the top 100 users captured by both activity-based and random sampling.

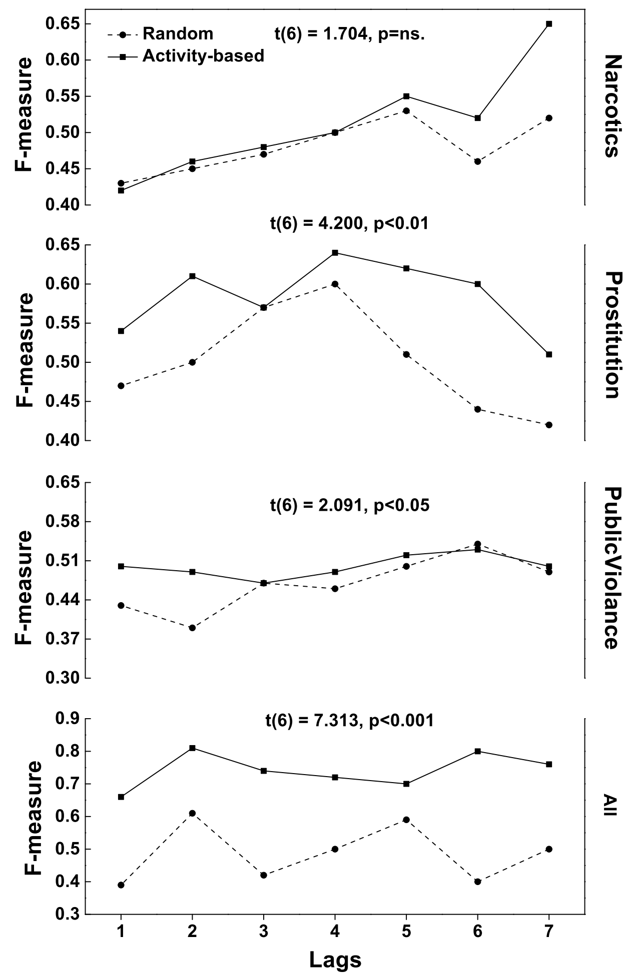


Fig. 3: Macro-averaged F1-score for different lags. "All" is the overall crime rates.

(see Equation 4). Figure 3 illustrates the Macro-averaged F1-score and t-Test of different crime types over different lags ($d = 7, q = 1$). The lag does not represent a particular day or week; it is a window of time in which crime rate directions are captured. As an example, if lag = 1 ($d = 1$), each document is labeled with the direction of crime rate in a day later. In each lag, the classifier is separately fed with the generated training data. The results in Figure 3 indicate that in most cases (lags), content obtained from active users has higher predictability compared with random sampling. In the best case that is the overall crime rate with $lag = 6$, the activity-based sampling achieved F-measure up to 0.80, which is 20% higher than random sampling (0.60). Although there are some lags which both datasets achieved the same results, in the most cases, the predictability of content captured from active users is higher than content of random users. Overall, content of active users was shown to be more credible for the proposed user-centric model. This can be the result of having fewer activity gaps compared with random sampling.

VI. Conclusion

In this work, we focused on sampling Twitter users to retrieve their historical tweets for a temporal user-centric prediction model. In this regard, we presented an activity-based approach, which leverages user profiles to estimate historical activities for the selection of the most active users as opposed to experts users. Two sets of users were sampled using activity-based and random sampling. The historical timelines of the selected users were retrieved using the REST API. We compared the primary statistical differences between two datasets in terms of user activity and historical timelines. We provided a prediction model to compare the credibility of the content gathered from the selected users. The findings indicated that the activity-based approach has more coverage in terms of historical tweets and user activity compared to random approach. The activity-based approach identified users who are more historically active, whereas in random sampling high activity gaps were observed. In addition, we also studied the credibility of the content captured from active users compared with random users. The findings indicated that content of active users is more credible in predicting the trend of interest. However, in the future, the study will be expanded to other crime trends as well. In addition to the timeline properties and credibility, in the future, we would like to investigate the quality of content in terms of topic of discussion and sentiments. We are also interested in determining the credibility of the content obtained by activity-based sampling in a content-centric model to examine the effectiveness of the proposed sampling approach for other temporal calcification problems.

References

[1] A. E. Marwick *et al.*, "I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience," *New media & society*, vol. 13, no. 1, pp. 114–133, 2011.

[2] J. Weng and B.-S. Lee, "Event detection in twitter." *ICWSM*, vol. 11, pp. 401–408, 2011.

[3] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[5] M. Bošnjak, E. Oliveira, J. Martins, E. Mendes Rodrigues, and L. Sarmento, "Twitterecho: a distributed focused crawler to support open research with twitter data," pp. 1233–1240, 2012.

[6] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 379–388.

[7] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi, "On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 1739–1744.

[8] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," *arXiv preprint arXiv:1306.5204*, 2013.

[9] K. White, G. Li, and N. Japkowicz, "Sampling online social networks using coupling from the past," in *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 266–272.

[10] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, "Cognos: crowdsourcing search for topic experts in microblogs," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 575–590.

[11] K. Joseph, P. M. Landwehr, and K. M. Carley, "Two 1% s dont make a whole: Comparing simultaneous samples from twitters streaming api," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 2014, pp. 75–83.

[12] M. Sofean and M. Smith, "A real-time architecture for detection of diseases using social networks: design, implementation and evaluation," in *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 2012, pp. 309–310.

[13] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 631–636.

[14] I. Chepurna, S. Aghababaei, and M. Makrehchi, "How to predict social trends by mining user sentiments," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 2015, pp. 270–275.

[15] S. Aghababaei and M. Makrehchi, "Temporal topic inference for trend prediction," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 877–884.

[16] M. Gaurav, A. Srivastava, A. Kumar, and S. Miller, "Leveraging candidate popularity on twitter to predict election outcome," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013, p. 7.

[17] M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of bfs (breadth first search)," in *Teletraffic Congress (ITC), 2010 22nd International*. IEEE, 2010, pp. 1–8.

[18] D. Wang, P. Hom, R. Griffeth, and J. K. Sager, "âĂIJnothing endures but changeâĂİ: Investigating temporal dynamics within a turnover model," in *Academy of Management Proceedings*, vol. 2015, no. 1. Academy of Management, 2015, p. 15237.

[19] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: discover spatio-temporal topics for twitter users," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 605–613.

[20] L. Li, M. F. Goodchild, and B. Xu, "Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr," *Cartography and Geographic Information Science*, vol. 40, no. 2, pp. 61–77, 2013.

[21] L. Hong, A. S. Doumith, and B. D. Davison, "Co-factorization machines: modeling user interests and predicting individual decisions in twitter," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 557–566.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, p. 2001, 2001.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] E. Zivot and J. Wang, "Rolling analysis of time series," in *Modeling Financial Time Series with S-Plus®*. Springer, 2003, pp. 299–346.