Event Detection on Large Social Media Using Temporal Analysis

Abdulrahman Aldhaheri
School of Engineering
University of Bridgeport
Bridgeport, Connecticut 06604
Email: aaldhahe@my.bridgeport.edu

Jeongkyu Lee
School of Engineering
University of Bridgeport
Bridgeport, Connecticut 06604
Email: jelee@bridgeport.edu

***Abstract*—Social media networks are now considered as one of the major news channels that breaks news as they fold. The problem of event detection based on social media has attracted researchers' attention recently because of the enormous popularity of social media. Existing approaches focus on features that don't reflect full characteristics of the social network. For the purpose of this research, we define an event as an occurrence that has enough force and momentum that could create an observable change of the context of a social network. Such a definition provides us with a wider perspective through which we can view the big picture of the social network. In this research, we propose a novel framework for detecting events on social media. We introduce a temporal approach to detect structural change of the social network that reflects an occurrence of an event using machine learning algorithms. In this study, we show that processing temporal social networks captures the complete complexity of the social network, which results in a higher accuracy of event detection.**

***Index Terms*—Big data, data mining, Social media analysis, Event detection, Machine learning.**

1. Introduction

Social media networks had become very popular recently. Statistica, the online statistics portal, estimated that there are 2.22 billion active social network users by the end of the year 2016 [1]. The same source estimates that this number will increase to 2.72 billion active social network user around the globe by the end of year 2019. Publishing personal contents has never been easier with the wide availability of microblogging platforms such as Twitter [2]. This has enabled users to post their opinions swiftly [3]. Recent research shows that Twitter process more than 500 million tweets daily [4]. The number of tweets that has been sent since 2006 when Twitter was founded is more than 300 billion tweets [5].

Data generated by social media users is huge in volume, grows at a very high velocity, varies in its type, and varies in its quality. These characteristics, also called the four V's, i.e volume, velocity, variety, and veracity, are the main dimensions that characterize big data [6]. The availability of huge datasets representing more than a quarter of the world's population who are actively interacting creates an opportunity to uncover patterns that could explain a lot of social phenomena [7], [8]. Meanwhile, the availability of these datasets introduces many challenges for researchers who are trying to analyze and process such data [7], [9].

1.1. Research Problems

Social media networks are now considered as one of the major news channels. Mainstream media tend to monitor social media networks by looking for breaking news and interesting event. Furthermore, government entities are also relying on social media for the purpose of collecting security related intelligence. On the other hand, social network analysis focuses on individual users and their networks.

The problem, i.e. the event detection on social media, attracted researchers attention recently because of the enormous popularity of social media. Existing approaches focus on features that doesn't reflect the characteristic of the social network. Therefore, it fails to detect events in the context of the social network as a whole, which result in lower accuracy in detecting events. To address the problem, the temporal approach for processing a social network as we can detect an event from multiple temporal images. We define an event as an occurrence that has enough force and momentum that could create an observable change the shape of social network. We can measure such change by comparing the shape of data as time goes by. Therefore, if the shape of data at time t_1 is different from the shape of data at time t_2 we can conclude that there was a certain event that has an impact on the data and changed its shape.

In this study, we show that processing temporal social networks graphs captures the complete complexity of the social network, which results in a higher accuracy of event detection model. We propose a temporal social network graphs event detection framework based on which we propose a novel social network transformation approach that transforms social media streams into temporal images. This allows for building a better event detection predictive model. We validate the proposed approach by performing experiments on streamed social media data collected for the purpose of this research. The ground truth collected data is

978-1-5090-4228-9/17/$31.00©2017 IEEE

extracted from mainstream media and labeled the dataset to create training and testing data.

We achieve an accuracy rate in detecting events that surpasses existing approaches. We evaluate our proposed approach by using commonly used model evaluation metrics. Accuracy alone could be deceiving especially when data is imbalance. We calculated and compared precision, recall, and F1-score. We also used precision-recall and ROC curves to evaluate the performance of our proposed approach.

1.2. Motivation Behind the Research

Recent studies show that there are more than 30% of the worlds population active across different social media platforms. Indeed, the largest sample of the global population available to any researcher. This fact attracted much attention recently to social media. Furthermore, the way in which social media users communicate along with the content they post messages made social media based event detection systems desired more than ever. The demand of an event detection system based on social media comes from different entities for many purposes. For example, detecting event based on social media improves security intelligence, social studies, news gathering, and many others.

However, the performed literature review shows that theres still a need for an event detection system that looks at the social network as a whole and consider the big picture rather than focusing on certain hand-picked features. Moreover, the development of new artificial intelligence and machine leaning approaches that are supported by specialized hardware architecture, such as graphical processing unit (GPU) and tensor processing unit (TPU), also played an important role in motivating us. We are intrigued by all this to attempt to approach social media based event detection problem from a different perspective.

2. Related Work

The problem of event detections has been approached in different ways depending on how the researcher define events. Some researchers define events as activities that take place at a specific place on a certain time [10]. Such a definition of events has driven researcher to focus on the locality and geo-spatial features of the messages [11]. Others combined locality with other features as in [12], [13], [14], [15], [16]. Others researchers focus on the actual conversation between the users, which lead to another approach in the development of event detection algorithm that utilizes text mining algorithms. In this approach, researchers look at events detection as a pure text mining problem as in [13], [17], [18], [19], [20], [21].

2.1. Location Based and Geo-Spatial Features

One research where event detection is done based on the locality of tweeting users is proposed by [11]. The authors collect geo-tagged tweets related to a target region using a custom built system to overcome restrictions imposed by Twitter [22]. After collecting a large amount of data, the authors partition the target region into sub-regions. This enables the authors to estimate regular pattern for each sub-region, which depends on the number of users, the number of posted tweets, and crowd movements. The authors can detect if there is an event is occurring at a certain sub-region by comparing the current pattern with the previously decided as a regular pattern for this region. This approach might be useful if you look at event as a certain occurrence at a specific time and place. However, given the popularity of social media network, users are interacting outside geographical boundaries. This is a limitation in all prosed event detection algorithms that tend to be influenced by the location of the users.

2.2. Textual Features

Another approach dominated the development of event detection algorithm is the utilization of text data mining. In this approach, researchers regard at events detection as a pure text mining problem. Many researchers are utilizing clustering algorithms in their proposed approach for detecting event [13], [18], [19], [20], [21]. Clustering is the process of grouping objects together according to their similarity in a way that results in having multiple clusters where object within a cluster is similar to each other and could be distinguished from objects in any other clusters [17]. In this approach, researchers start by selecting features upon which the clustering algorithm groups tweets into different clusters. The selected features are usually hashtags, keyword, and contents of the tweets.

As explain in the previous section, TwitterStand [13], is an event detection system that focuses on news related topics in tweets. It uses a weighted list of keywords as features vector. The proposed algorithm detect events based on textual features. Users location features are only used for graphical user interface purposes.

One event detection algorithm based on clustering is proposed in [18]. The authors used Twitter streaming API to collect tweets based on hashtags and keywords. Then used Apache Lucene to create an index for each tweet based on the content of the tweet. The authors applied Term-Frequency Inverse Document Frequency (TF-IDF) to define similarity between different tweets.

The use of textual features in event detection introduces many challenges. For example, users miss spell a word, use abbreviated word, or sometimes different languages. Figure 1 Shows how users connected through the same social networks are using different languages. Therefore, developing a system that relies only on textual features would result in inconsistency. Specifically, if the developed event detection system is running in a bilingual society.

2.3. Hybrid Features - Location and Textual

Some researcher implemented a hybrid approach that combines both of the location and the textual features of

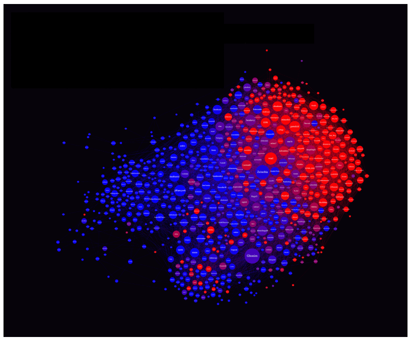

Figure 1: Egypt influence network by @kovasboguta. The size represents the influence of the user across the entire social network. The colors of the node represent the language used by the users; English in blue, Arabic in red

the tweets such as [12], [13].

One approach proposed by [12] implements geo-spatial along with text search. The proposed system implements support vector machine (SVM)classifier on some features of the tweet, such as tweet text and the number of words in each tweet. The tweets are collected every second using Twitter API by querying for keywords such as earthquake and shaking. This enabled them to prepare an events probabilistic spatiotemporal model. The proposed system will detect an event if the probability is more than a manually set threshold, i.e. (0.95). The location of the users is detected using geo-location attribute in the tweet. If its not enabled, the authors implemented two widely used filters to estimates the location of the users, Kalman and particle filters [12]. The tweet is ignored if the location couldnt be specified in either of the two ways. Although the authors used a manually set threshold, this approach is suitable for a specific event at a specific location.

A similar approach that uses both text and location is proposed by [13]. In their proposed processing system, the systems start by a set of seeders. Those seeders are 2000 manually identified active users that are known to publish news frequently, which include many mainstream media account. The system will always consider tweets from a seeder account as news. Other tweets collected using Twitter API are classified into news or junk using Naive Bayes classifier. The system, then, applies a modified version of leader-follower clustering algorithm to group tweets into clusters, where each cluster represents a different topic. The authors applied a geo-tagging technique to identify the location of each event represented as a cluster. For all tweets in each cluster, the resulted geo-tags are then aggregated into a geo-tagging range that represent the location of the event. The detected location is used for the user interface of the system. The proposed system depends on the handpicked list of users that is called seeders. Choosing users for this list would highly impact the performance of the system.

There are many other event detection algorithms proposed in the literature that either detect events either based on the locality of the users or based on a hybrid approach that uses the locality of the users along with the contents of the message. The shortcoming in these approaches lies in the fact that only a very small percentage of social media users enable geo-tagging for privacy concerns. This will result in excluding non-geo-tagged data from the collected dataset as in [12], which would result in missing an opportunity of detecting real events.

The understanding of events has influenced the way researchers extract features from social networks. Generally, an event is something that happen at a certain time [23]. It's also defined by Sakaki et al. as an arbitrary classification of a space/time region [12]. Troncy et al. define events as real-world occurrences that unfold over space time [24]. For the purpose of this research, we define an event as an occurrence that has enough force and momentum that could create an observable change the shape of social network. We can measure such an event by comparing the shape of data as time goes by. So, if the shape of data at t_1 is different from the shape of data at t_2 we can conclude that there was a certain event that has an impact on the data and changed its shape. This definition allows us to see the big picture of the social network being studied.

3. Temporal Event Detection

Modeling social networks as a graph enables researchers to perform a wide range of social networks analysis. However, processing huge social networks imposes many challenges. We are proposing a novel approach to process social networks. By taking advantage of the adjacency matrix representation of the social network graph. We convert the social network graph into an $n \times n$ adjacency matrix, where n is the number of nodes in the social network. Then, we convert the adjacency matrix to a gray-scale image with an $n \times n$ pixels as Figure 2 shows. The gray-scale image could be scaled down to reduce the size of the data.

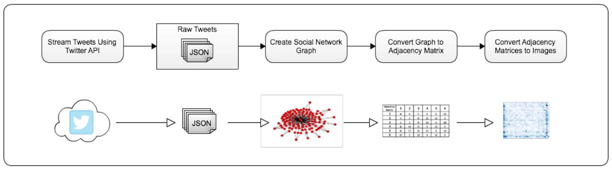
Stream Tweets Using Twitter API
Raw Tweets
JSON
Create Social Network Graph
Convert Graph to Adjacency Matrix
Convert Adjacency Matrices to Images
JSON

Figure 2: Converting raw data into an image

We scaled down the images using a well-known down sampling technique, i.e. max pooling. Max pooling is a down sampling technique that results in reducing the size of the image drastically. In the max pooling, a defined region with a dimension $p \times q$ returns a single value, which is the maximum value in the region.

We implement the system on actual dataset collected from Twitter live using Twitter public streaming API. A complete description of the dataset is illustrated in section 3.1. We use 10-folds cross validation for training and testing the model. 10-folds cross validation is a method used to validate the consistency of the predictive model and to resolve model overfitting. In this method, the dataset is divided into 10 equal size partitions. The method, then holds

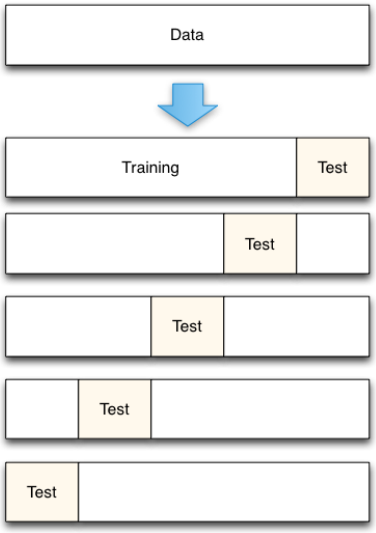

Figure 3: 5-folds cross validation

the k^{th} partition as a testing sample and train the model using the rest of the data. The process is repeated for k-rounds with a different testing partition. Figure 3 shows how k-folds cross validation works on a dataset when $k = 5$, however, im our implementation $k = 10$. The final score is the average score for all rounds. Figure 4 shows how we've plotted the precision-recall curve and calculated the area under the curve (AUC) for each fold, and then computed the overall AUC score.

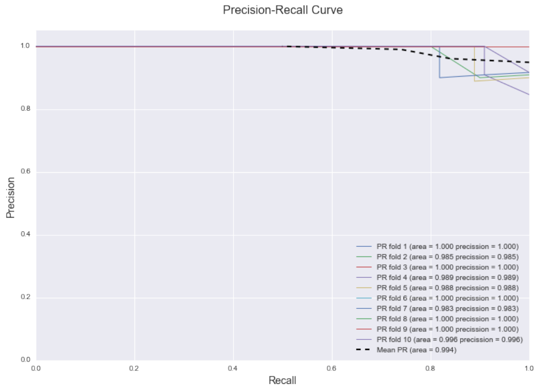

Figure 4: Precision-recall curve

3.1. Dataset

Twitter provides its registered developers with access data generated by users, known as tweets. These tweets entities are stored in JSON format. We retrieve posts generated by users over three weeks time period starting from Sep. 17, 2014 until Nov. 20, 2014. We have filtered these data by collecting only 'NFL' related posts using twitter Streaming API. The amount of tweets is more than 4.4 million tweets. The size of the dataset is 17 Gb. We started our analysis by classifying the dataset based on each message type. Messages on Twitter false into two types, tweets and retweets. Tweets are message created by the posting users, while retweets are messages posted by users. Therefore, classify each tweet within our dataset into one of the two described status. As shown in Figure 5, 68.5% of the 4,410,717 collected data are original tweets created by the tweeting user. The rest, 31.5% are actually retweets. This means that only 31.5% of the collected tweets are deemed worthy by the users to be reposted again.

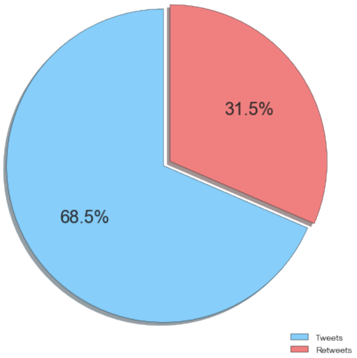

Figure 5: Tweets Type

3.2. Evaluation Metrics

We evaluate the proposed model by applying well-known metrics in data mining, document retrieval, and machine learning algorithms. We select model evaluation metric that covers a wide range of special cases and scenarios. In particular, we use the following metric to evaluate the proposed framework:

- *Confusion matrix*: provides a detailed description of the model prediction, which is required for computing other performance evaluation metrics. The confusion matrix is a table that compares between the actual classes and the predicted classes. The confusion matrix considers only two classes, negative and positive. Table 1 shows how a confusion matrix is organized:

		Actual		
		Positive	Negative	
Predicted	Positive	True Positive (**TP**)	False Positive (**FP**)	↔ PPV
	Negative	False Negative (**FN**)	True Negative (**TN**)	↔ NPV
		↕ se	↕ sp	

TABLE 1: A typical confusion matrix for a binary classification problem.

- *Accuracy*: measures how well a model performs for all of the predicted classes. It shows the proportion of the predicted labels that has been successfully detected. Model accuracy is computed by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- *Precision, also know as Positive Predictive Value (PPV)*: measures the proportion of the positive predictions that are correct. Model precision is computed using the following formula:

$$Precision, PPV = \frac{TP}{TP + FP} \quad (2)$$

- *Recall, also know as True Positive Rate (TPR), Sensitivity, or Hit Rate* : measures the model performance in detecting positive labels. Its calculated using the following formula:

$$Recall, TPR = \frac{TP}{TP + FN} \quad (3)$$

- *F1-Score*: combines both of the precision and recall scores. Could be looked at as the harmonic average of the precision and recall scores. Its bias toward the smaller value of them, which means that it penalizes the model for having low precision or recall. F1-Score is calculated using the following formula:

$$F1 - Score = 2 \times \frac{PPV \times TPR}{PPV + TPR} \quad (4)$$

The selected model evaluation metrics covers a wide range of special cases and scenarios. For example, accuracy itself is a good indicator for the overall model performance, however, it fails in providing reasonable meaning if the dataset is imbalanced. Consider a dataset that has 99 instances labeled as true, and only 1 instance labeled as false. A prediction model that predict every instance to be true will have an accuracy of 99%. In this case, precession and recall provide better measures for the model performance.

4. Experimental Results and Discussion

We implemented the proposed framework on an actual dataset that we collected using Twitter public API. The dataset contains more than 4.4 million raw tweets with a size that exceeds 17Gb. Using the proposed framework, we extracted temporal social network graphs from the dataset and then transform them into temporal snapshots. We examine the impact of the applied dimensionality reduction technique on the performance of the event detection model. We applied a multilayers neural network on every dataset and recorded the model performance's results in Table 2.

As Table 2 shows, the performance of the model remains consistence even after reducing the dimension of the data drastically. The observed recorded fluctuation in the model performance measures are irrelevant to the implemented dimensionality reduction technique. In fact, these fluctuations are due to two factors; the architecture of the multi-layer neural network, and the randomization of the initialization variable, such as weights and biases.

Region	***Accuracy***	***Precision***	***Recall***	***F1-Score***
100 × 100	0.970	0.980	0.980	0.980
110 × 110	0.985	1	0.981	0.990
120 × 120	0.978	0.991	0.981	0.986
130 × 130	0.949	0.981	0.953	0.967
140 × 140	0.963	0.982	0.973	0.978
150 × 150	0.993	0.991	1	0.995
160 × 160	0.956	0.990	0.953	0.971
170 × 170	0.978	0.973	1	0.986
180 × 180	0.949	0.991	0.948	0.969
190 × 190	0.978	0.991	0.983	0.987
200 × 200	0.971	0.982	0.982	0.982

TABLE 2: Impact of image down sampling on model performance

Figure 6 shows how the number of features is reduced drastically as the dimension of the region increase. Lowering the number of features without compromising the quality of the data allows the proposed approach to be scalable. In this experiment, we can process social media data for huge network efficiently.

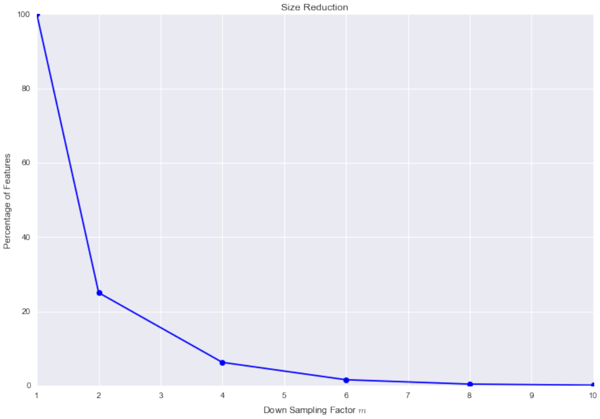

Figure 6: A drastic reduction in the number of features.

The cost of using machine learning algorithms have is high because of the high computing complexity. This makes them hard to scale for large datasets. Our proposed approach allows for the utilization of such algorithms because it achieved high dimensionality reduction, while maintaining the main features of the dataset. The proposed approach is scalable for large social networks. We show empirically that we can maintain high accuracy for event detection even after drastically reducing the size feature space.

5. Conclusion

In conclusion, the event detection of social media is attractive because of the popularity of social media. Existing approaches falls into one of three categories: (i) event detection based on users location, (ii) event detection based on users messages, and (iii) event detection based on a hybrid approach that combines location and textual features of the message. Constraining the detection algorithm to such features results in a detection system that lacks the depth required to capture the full complexity of the network.

We introduce a temporal approach to detect structural changes in the social network that reflects an occurrence of an event using machine learning algorithms. In this study, we show that processing temporal social networks graphs captures the complete complexity of the social network, which results in a higher event detection model accuracy. We proposed a Temporal Social Network Graphs Event Detection framework based on a novel social network transformation approach that transforms social media streams into temporal images, which allows for building a better event detection predictive model.

We validate the proposed approach by performing experiments on streamed social media data collected for the purpose of this research. The ground truth collected data is extracted from mainstream media and appended to the dataset to create training and testing data. We evaluated the proposed model by applying well-known metrics used to evaluate data mining, document retrieval, and machine learning algorithms. We selected model evaluation metrics that covers a wide range of special cases and scenarios. We have also shown that the proposed framework detects events with very high accuracy.

References

[1] S. T. S. Portal, "Number of global social network users 2010-2018," Nov. 2014. [Online]. Available: http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] A. Acerbi, V. Lampos, P. Garnett, and R. A. Bentley, "The expression of emotions in 20th century books," *PLoS ONE*, vol. 8, no. 3, p. e59030, 03 2013. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0059030

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 2008, pp. 183–194.

[4] F. Zhao, J. Liu, J. Zhou, H. Jin, and L. T. Yang, "Ls-ams: An adaptive indexing structure for realtime search on microblogs," *IEEE Transactions on Big Data*, vol. 1, no. 4, pp. 125–137, 2015.

[5] R. Agrawal, B. Golshan, and E. Papalexakis, "Whither social networks for web search?" in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1661–1670.

[6] Q. Yang, "Introduction to the ieee transactions on big data," *IEEE Transactions on Big Data*, vol. 1, no. 1, pp. 2–15, 2015.

[7] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, Jan 2010. [Online]. Available: http://dx.doi.org/10.1016/j.bushor.2009.09.003

[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[9] D. Maynard, K. Bontcheva, and D. Rout, "Challenges in developing opinion mining tools for social media," *Proceedings of@ NLP can u tag# usergeneratedcontent*, 2012.

[10] S. M. Alqhtani, S. Luo, and B. Regan, "Fusing text and image for event detection in twitter," *arXiv preprint arXiv:1503.03920*, 2015.

[11] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*. ACM, 2010, pp. 1–10.

[12] 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.

[13] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2009, pp. 42–51.

[14] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 2012, pp. 1273–1276.

[15] M. Walther and M. Kaisser, "Geo-spatial event detection in the twitter stream," in *European Conference on Information Retrieval*. Springer, 2013, pp. 356–367.

[16] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 389–398.

[17] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011, pp. 443–446.

[18] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in twitter," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 3. IEEE, 2010, pp. 120–123.

[19] R. Long, H. Wang, Y. Chen, O. Jin, and Y. Yu, "Towards effective event detection, tracking and summarization on microblog data," in *International Conference on Web-Age Information Management*. Springer, 2011, pp. 652–663.

[20] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 181–189.

[21] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter." *ICWSM*, vol. 11, pp. 438–441, 2011.

[22] T. Fujisaka, R. Lee, and K. Sumiya, "Discovery of user behavior patterns from geo-tagged micro-blogs," in *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*. ACM, 2010, p. 36.

[23] H. Becker, M. Naaman, and L. Gravano, "Event identification in social media." in *WebDB*, 2009.

[24] R. Troncy, B. Malocha, and A. T. Fialho, "Linking events with media," in *Proceedings of the 6th International Conference on Semantic Systems*. ACM, 2010, p. 42.