International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Event Detection Model for Facebook News Posts

Wafa Zubair AL-Dyani, Farzana Kabir Ahmad, Siti Sakira Kamaruddin

Abstract*: Event detection has wide application especially in the area of news streams analyzing where there is a need to monitor what events are emerging and affecting people's lives. This is crucial for public administrations and policy makers to learn from their previous mistakes to make better decisions in the future. Different researchers have introduced several event detection models for Facebook news posts in. However, majority of these models have not provided adequate information about the discovered news events such as location, people and activity. In addition, existing models have ignored the problem of high dimensional feature space which affects the overall detection performance of the models. This research presents a conceptual event detection model for mining events from large volume of short text Facebook news posts and summarize their valuable information. This is crucial for public administrations and policy makers to learn from their previous mistakes to make better decisions in the future. The proposed model includes pre-processing, feature selection, event detection and summarization phases. The pre-processing phase involves several steps to convert unstructured text news posts into structure data. Feature selection phase to select the optimal feature subset. Meanwhile, event detection phase uses these features to construct undirected weighted graph and apply dynamic graph technique to identify the clusters from the graph and then annotate each cluster to its corresponding event. At the end of this paper, several unresolved problems in the construction of event detection model from Facebook news posts are reported to be used as future work for the current study.

Keywords: Event detection, Facebook, Feature selection, graph, news, text clustering.

I. INTRODUCTION

In the last decade, a large volume of data which has created on different social network sites has attracted many researchers from various fields to collect and analysis this data [1],[2]. According to the literature, textual data represents about 80% of the total data on the web [3]. In fact, textual data comes in two styles (e.g., formal and informal) and generated from various sources such as news websites, news feeds, Weblogs, Forums, Emails, Facebook and Twitter [4]. Recently, Facebook has emerged as a powerful source to get knowledge about different real-world events [5].

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Wafa Zubair Al-Dyani, Department of Computer Science, Information Technology College, Hadramout University, Hadramout, Yemen.
Farzana Kabir Ahmad*, School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia.
Siti Sakira Kamaruddin, School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia.

On top of that, it has recognized as the most popular platform for consuming, sharing and commenting on news items about events in comparison with Twitter and YouTube [6]. For such reasons, almost all news media nowadays, use Facebook to quickly broadcast their breaking news in the form of Facebook posts [7]. However, the huge volume of news posts which are published on a daily basis by different Facebook news pages has caused the problem of information overload about various real-world events [8]. The wealth of this information is acknowledged as a significant for police-makers in different disciplines as well as it remains unexploited by machine learning applications. The reasons for such worthiness include assisting news analysts to perform an accurate comparison between different Facebook news pages based on the type of events covered and hence, measure how well these pages are known by news readers around the world. In addition, supporting managers of news channels in recognizing news readers' reactions towards different real-world events through utilizing the meta-data associated with the Facebook news posts e.g., number of likes, comments, sharing, engagement, etc. Consequently, improve the strategies for selecting the type of news to be published in the future by the news channels.

All together has motivated many researchers from Event Detection (ED) field to introduce several ED models [9]. However, several challenges for building high accuracy ED model for Facebook news posts are identified by different researchers [4]. One important challenge is the high dimensional feature space that contains various kind of features such as redundant, irrelevant and noisy features [10]. These features cause the problem of overfitting for the given ED model, mislead the detection method and eventually decrease the overall accuracy of ED model [11]. The other crucial challenge is designing an unsupervised detection method which is capable to detect and summarize different real-world events from large volume of short text Facebook news posts [4]. Therefore, phases such as pre-processing phase, feature selection phase and event detection phase have become vital phases in building ED model since the model's accuracy is highly influenced by the methods used under those phases [4].

ED also known as Retrospective Event Detection (RED) has wide applications in discovering important patterns and has been extensively studied since a long time ago. However, it remains as an active research area until present time [4], [12]–[16]. One major example of RED applications is in the identification of real-world events from Facebook news posts. Recently, several RED models for Facebook news posts have been introduced in literature [9].

However, almost all existing models have not provided sufficient summary about the discovered events in terms of answering questions like what has happened, when, where and who was involved [6], [9]. On top of that, all current models apply only single FS technique, which is proved to be insufficient to remove all unnecessary features and select the optimal ones [10]. In addition, majority of models use traditional ED methods, which are reported to be unsuitable for short text Facebook news posts as they do not provide sufficient statistical information for calculating the similarity between posts [17]. Moreover, they do not provide adequate information about an event such as the location, people and activity to create a summary about the discovered events [18]. For such reasons, this study proposes a conceptual ED model for detecting and summarizing events from Facebook news posts through applying improved pre-processing, feature selection and event detection methods. Section II discusses the general view of the model followed with an explanation of the techniques used under each phase of the proposed model. The paper ends with a conclusion in Section III.

II. CONCEPTUAL EVENT DETECTION MODEL

Fig. 1 illustrates the conceptual ED model for detecting and summarizing events from Facebook news posts. It includes phases like pre-processing, feature selection, event detection and summarization phases. The study has collected sample of news posts from eight popular news pages on Facebook as dataset for this research using Facebook built-in application called Netvizz.

A. Pre-processing Phase

Pre-processing phase involves several steps such as filtering noisy and unrelated posts, remove URLs, remove special characters, tokenization, Named Entity Recognition (NER), remove stop words, stemming and data representation. These steps except NER are used to eliminate all noisy contents and convert the unstructured data into a structured format which is suitable for the subsequent phases. On the other hand, NER activity is used to identify the different entities like persons, locations, organizations. As a result, this help in summarizing the information related to the discovered events later in the summarization phase. Also, pre-processing phase includes data representation step in which Vector Space Model (VSM) is used to represent news posts as feature vectors weighted by Term Frequency-Inverse Document Frequency (TFIDF). The results from this phase is used as input for the next feature selection phase.

B. Feature Selection Phase

Feature selection is an important process of ED model, whereby this process is used to remove all unnecessary features. Existing ED models for Facebook news posts have utilized only single feature technique such as TFIDF [19], Term Frequency (TF) [6]. However, single FS technique is not capable to remove all noisy, redundant and irrelevant features [10]. Therefore, this study proposes improved hybrid filter-wrapper methods based on one popular meta-heuristic swarm intelligence algorithm. Recently, many researchers from text mining field have employed such methods to select optimal feature subset due to their good balance between the computational efficiency of a filter technique and the high accuracy performance of the wrapper technique [20]. These methods have shown promising results in comparison with other feature selection methods [21]. At the end of this phase, the optimal feature subset is selected, which is used as input by the subsequent phase.

C. Event Detection Phase

This phase consists of several processes. Starting with constructing undirected weighted graph from the optimal feature subset obtained from the previous phase. The graph is denoted as Gt (Vt, Et, Wt) and represents the relations among the features. whereby, Vt, Et, Wt are sets of nodes, edges and weights, respectively. $F \in Vt$ is an n-dimensional optimal features F = {F1, F2, …., Fn}, which have obtained from previous phase and represent a set of nodes within the time interval t.

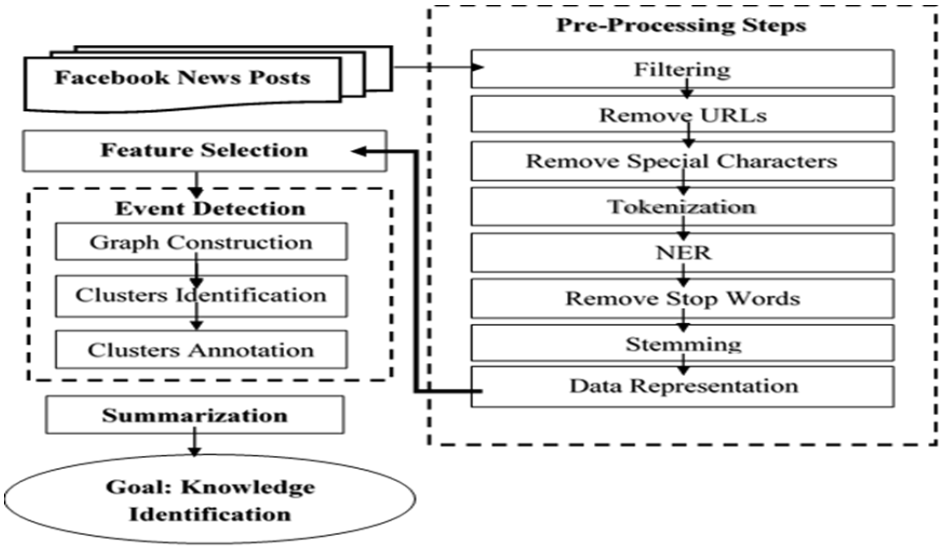

Fig. 1.Conceptual event detection model

Subsequently, a weight w ∈ Wt will be assigned to the edge e (F1, F2) ∈ Et. The weight is calculated using the following adapted Equation which was proposed in [22]:

$$w_{F1,F2} = sim(F_1, F_2) = \frac{\sum_{k=1}^{n}(F_{1k} \times F_{2k})}{\sqrt{\sum_{k=1}^{n} F_{1k}^2} \times \sqrt{\sum_{k=1}^{n} F_{2k}^2}} \quad (1)$$

In this study a graph-based method is selected as a detection method because existing traditional ED methods fail to identify the events that do not have high volume of documents. Besides that, they ignore the semantic relationships among the features [23],[24]. After the graph is constructed, the process of detecting clusters (i.e., events) is carried out to identify the structures of clusters within the graph. In literature, many graph-based techniques have been employed to detect events from different types of graphs like Girvan and Newman algorithm, cut-off method, Agglomerative Hieratical Clustering (AHC) algorithm, Louvain clustering algorithm, Markov Clustering Algorithm (MCL), voltage clustering algorithm [25]. However, among all these techniques, dynamic graph-based techniques have shown better performance in detecting clusters from either weighted or unweighted graphs that are built based on data taken from Facebook, Twitter and LinkedIn [26]. Additionally, they generate good clustering results and has the ability to deal well with noisy contents [27]. On top of all that, they does not require to specify the number of clusters in advance and this is ideal as events often occur without prior warning or knowledge [22]. Therefore, a dynamic graph-based technique is used by this study to identify the clusters from the undirected weighted graph. Subsequently, most active posts of each cluster are identified using the number of engagements which is associated with Facebook news posts. In fact, the number of engagements is the total number of post's meta-data such as number of likes, shares and comments. After that, a modified TFIDF will be applied over them in order to extract the events. Finally, each cluster will be assigned to its corresponding event as it can be seen from Fig. 2.

Summarization

In this phase, important information of the detected events such as where and when the event has happened and who was involved are extracted. To illustrate, current study uses similar method which has proposed in [28]. First, the study generates a gazetteer base for names and locations using the latest version of DBpedia in order to compare words in the cluster with gazetteer entries. Later, a TFIDF is run on the extracted names and locations to rank them and select the most top from them. On the other hand, to identify the time at which the event has occurred the study uses timestamps associated with the news posts that are related to an event and detect the minimum time among all timestamps. This time is considered as the starting time for that event. All together produces a summary about each detected event.

III. CONCLUSION AND FUTURE WORK

In this paper a conceptual ED model is presented which will be the foundation for this research in the future. It highlights the major tasks included in this research towards the achievement of the goal; identifying the significant past events and summarize their valuable information from large volume of short text Facebook news posts. The outcome of this study can be seen as an effort to monitor what events are emerging and affecting people's lives. Whereby, this is crucial for public administrations and policy makers to learn from their previous mistakes in order to make better decisions in the future. In the same context, ED model can help news analysts to carry out an accurate comparison between different Facebook news pages based on the type of events covered. Consequently, find out how well these pages are popular by analyzing readers interaction toward the published news. Thus, improve their strategies for selecting the type of news to be published in the future. Finally, the proposed model can help in organizing the news posts into various events so news readers can search for their desired news posts easily and effectively.

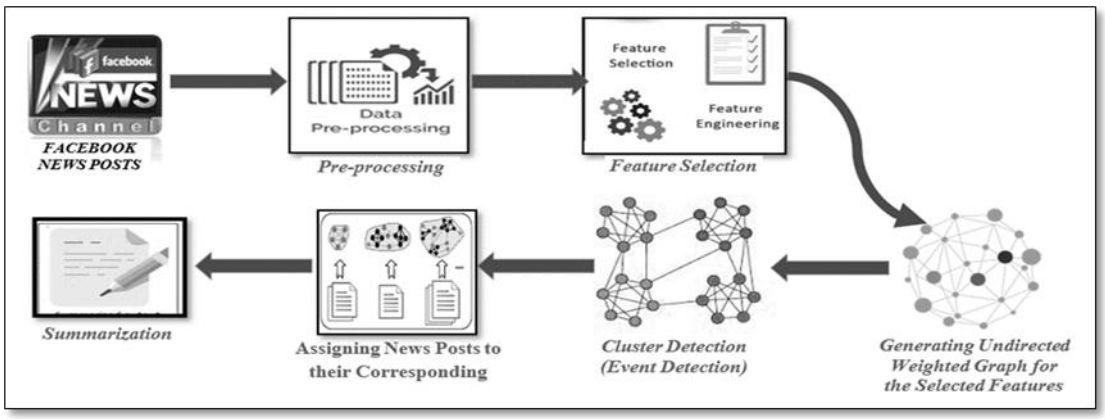

Fig.2. Main processes of the conceptual event detection model

Despite the existence of several ED models for Facebook news posts, however, there are still number of unsolved problems which need further research. Firstly, applying only single feature technique is not enough to remove all noisy, redundant and irrelevant features. Therefore, new or improved feature selection method is required to remove all unnecessary features and select optimal ones. Recently, hybrid filter-wrapper based on popular meta-heuristic swarm intelligence algorithm have used as feature selection method and have shown promising performance in solving feature selection problems in different areas. Yet. majority of them have applied over just few numbers of instances that have up to only thousands features. Thus, more improvements should be done to fill up the scalability gap and make them more fit to work on high dimensional datasets with millions of features. Secondly, ED models that use dynamic graph-based methods as detection methods have applied on static and relatively very small-scale graphs. Therefore, more improvements are required to be done fit with large scale dynamic graphs like graphs constructed from Facebook and Twitter data. Finally, current annotation techniques for ED models have not taken advantage of associated meta-data. Whereby, employing text contents with meta-data confirmed to achieve more accurate annotation of clusters to their correct events. For such reasons, an automatic labelling technique is needed which employ meta-data for accurate assigning of clusters to their corresponding events as well as consumes minimum human intervention. Given all these limitations in the existing ED models for Facebook news posts, our future work will focus on proposing an enhanced ED model for Facebook news posts. In particular, our focus will be on improving the methods used for feature selection and event detection phases in order to resolve the given above problems.

REFERENCES

1. F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim, "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams," pp. 1–15, 2014.
2. F. Abdullah, K. R. Ku-Mahamud, F. Ahmad, N. F. A Ghani, and M. M Kasim, "Relative efficiency assessment of projects using data envelopment analysis: A case study," Int. J. Digit. Content Technol. its Appl., vol. 6, no. 9, pp. 310–318, 2012.
3. M. H. Singh, "Clustering of text documents by implementation of K-means algorithms," Streamed Info-Ocean, vol. 1, no. 1, pp. 53–63, 2016.
4. A. Goswami and A. Kumar, "A survey of event detection techniques in online social networks," Soc. Netw. Anal. Min., vol. 6, no. 1, p. 107, 2016.
5. N. Sormanen, J. Rohila, E. Lauk, T. Uskali, J. Jouhki, and M. Penttinen, "Chances and challenges of computational data gathering and analysis: The case of issue-attention cycles on Facebook," Digit. Journal., vol. 4, no. 1, pp. 55–74, 2016.
6. L. C. Passaro, A. Bondielli, and A. Lenci, "FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis," in Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016., 2016.
7. S. A. Salloum, C. Mhamdi, M. Al-Emran, and K. Shaalan, "Analysis and Classification of Arabic Newspapers' Facebook Pages using Text Mining Techniques," Int. J. Inf. Technol., vol. 1, no. 2, pp. 8–17, 2017.
8. G. Leban, B. Fortuna, and M. Grobelnik, "Using News Articles for Real-time Cross-Lingual Event Detection and Filtering.," in NewsIR@ ECIR, 2016, pp. 33–38.
9. Ahmed Al-Rawi, "News values on social media: News organizations' Facebook use," Journalism, pp. 1–19, 2016.
10. M. Allahyari et al., "A brief survey of text mining: Classification, clustering and extraction techniques," arXiv Prepr. arXiv1707.02919, 2017.
11. B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," IEEE Trans. Evol. Comput., vol. 20, no. 4, pp. 606–626, 2016.
12. T. R. Chandran, A. V Reddy, and B. Janet, "Text Clustering Quality Improvement using a hybrid Social spider optimization," Int. J. Appl. Eng. Res., vol. 12, no. 6, pp. 995–1008, 2017.
13. N. Panagiotou, I. Katakis, and D. Gunopulos, "Detecting events in online social networks: Definitions, trends and challenges," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9580, Springer, 2016, pp. 42–84.
14. R. Hassanian-esfahani and M. Kargar, "A survey on web news retrieval and mining," in Web Research (ICWR), 2016 Second International Conference on, 2016, pp. 90–101.
15. Q. H. Ramadan and M. Mohd, "A review of retrospective news event detection," in Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on, 2011, pp. 209–214.
16. X. Dai, Y. He, and Y. Sun, "A Two-layer text clustering approach for retrospective news event detection," in Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on, 2010, vol. 1, pp. 364–368.
17. J. Deng, F. Qiao, H. Li, X. Zhang, and H. Wang, "An Overview of Event Extraction from Twitter," in Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on, 2015, pp. 251–256.
18. F. Zarrinkalam and E. Bagheri, "Event identification in social networks," Encycl. with Semant. Comput. Robot. Intell., vol. 1, no. 01, p. 1630002, 2017.
19. C. S. Cracs and P. Porto, "A Three-Step Data-Mining Analysis of Top-Ranked Higher Education Institutions ' Communication on Facebook," 2018.
20. A. Alsaeedi, M. A. Fattah, and K. Aloufi, "A hybrid feature selection model for text clustering," in System Engineering and Technology (ICSET), 2017 7th IEEE International Conference on, 2017, pp. 7–11.
21. S. Arora and P. Anand, "Binary butterfly optimization approaches for feature selection," Expert Syst. Appl., vol. 116, pp. 147–160, 2019.
22. B. Manaskasemsak, B. Chinthanet, and A. Rungsawang, "Graph Clustering-Based Emerging Event Detection from Twitter Data Stream," in Proceedings of the Fifth International Conference on Network, Communication and Computing, 2016, pp. 37–41.
23. S. . Abdulsahib, A.K., Kamaruddin, "Graph based text representation for document clustering," J. Theor. Appl. Inf. Technol., vol. 76, no. 1, pp. 1–13, 2015.
24. A. Edouard, "Event detection and analysis on short text messages," 2018.
25. S. Katragadda, R. Benton, and V. Raghavan, "Framework for real-time event detection using multiple social media sources," 2017.
26. B. S. Khan and M. A. Niazi, "Network Community Detection : A Review and Visual Survey," 2016.
27. Q. Chen, X. Guo, and H. Bai, "Semantic-based topic detection using Markov decision processes," Neurocomputing, vol. 242, pp. 40–50, 2017.
28. Y. Zhang, C. Szabo, and Q. Z. Sheng, "Sense and focus: towards effective location inference and event detection on Twitter," in International Conference on Web Information Systems Engineering, 2015, pp. 463–477.

AUTHORS PROFILE

Wafa Zubair Al-Dyani received the bachelor's degree in Computer Science (with Honors) from Hadramout University, Yemen, and M.S. degree in Information Technology from University Utara Malaysia (UUM), in 2008 and 2016, respectively. She is currently pursuing the Ph.D. degree in Computer Science with UUM, where she is currently a Research Student with the Data Science Research Laboratory, School of Computing, UUM. She is also a lecturer with the Department of Computer Science, Hadramout University, Hadramout Province, Yemen. Her research interests include Text Mining, Natural Language Processing, Machine Learning, Optimization Algorithms (Bat Algorithm).

Dr Farzana Kabir Ahmad is a senior lecturer at School of Computing, Universiti Utara Malaysia. She holds a bachelor's degree of Computer Science (with Honors) from Universiti Sains Malaysia in 2003 and a Master degree in Computer Science from the same university later in 2005. She pursued her Ph.D. in Computer Science (Bioinformatics) from Universiti Teknologi Malaysia in 2012 and her doctoral work involves the development of synergy network for breast cancer progression using gene expression profiles. Her research interests include areas such as Computational Intelligence, Machine Learning, Bioinformatics, Neuroscience, Neuroinformatic, Text Mining, and Social Media Analytic.
Email: farzana58@uum.edu.my.

Associate Prof. Dr. Siti Sakira Kamaruddin received her Diploma in Computer Science from Universiti Putra Malaysia (UPM), Serdang, Selangor in 1990. She completed her Bachelor's degree and Masters in Computer Science from Universiti Teknologi Malaysia (UTM), Skudai, Johor in 1995 and 1998 respectively. She obtained her PhD in Science and System Management from Universiti Kebangsaan Malaysia, Bangi, Selangor in 2011 and currently is an Associate Professor in the School of Computing, Universiti Utara Malaysia (UUM), Sintok, Kedah. She has published various articles in scientific journals and international conferences in the area of computational intelligence, text mining, natural language processing, social media analytics and data mining