International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Event Identification in Social News Streams Using Keyword Analysis

Tabiya Manzoor Beigh[1], Shuchita Upadhyaya[2], Girdhar Gopal[3]

[1]M.tech Student, Department Of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana 136119 India

[2] Professor, Department Of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana 136119 India

[3] Assistant Professor, Department Of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana 136119 India

Abstract - *Millions of users share their experiences, thoughts, and interests online, through social media sites e.g., Twitter, Flickr, YouTube. As a result, these sites host a substantial number of user-contributed documents e.g., textual messages, photographs, videos for different types of events e.g., concerts, political demonstrations, earthquakes. In this paper, a method is proposed for leveraging the wealth of available social media documents in the form of continuous news streams to identify and characterize events. A news stream consists of news stories related to a particular topic. A bursty event is represented as a group of highly correlated bursty features. Bursty features cluster around bursty events in a single or multiple documents. Bursty features are identified with different sliding window sizes. The paper also discusses the evolution of burst events along the timeline. In this paper, a formal insight is given to the above problem and a solution is presented. Proposed method is extensively evaluated on the Reuters Corpus Volume 1. Experimental results show that our methods can detect bursty events in a timely way and effectively discover their evolution.*

Keywords: - Big data, Bursty Feature, Bursty Event, Event Detection, News Streams, Social media, Sliding window.

1. INTRODUCTION

Huge amount of electronic data is available in the current age. This data includes mainly big data. Big data represents the large and rapidly growing volume of information, more or less structured made available from different and diverse sources. The information is dispersed among various agencies including social media, retail sector, health sector, banking sector and customer feedbacks in call centers in the form of chats. People find themselves well suited if they express their feelings, emotions, thoughts and interests electronically. The best outlets for individuals to share their interests are Facebook, Twitter, WhatsApp and YouTube. The people appears to be moving away from the traditional approach using specific media environments such as newspapers, magazines, or television shows and instead tap themselves with technologies that reach targeted people at optimal times in optimal locations. So, individuals are revolving around electronic data. To efficiently utilize this electronic data, it has been a necessity for organizations to mine this huge amount of data. This would reduce their competitive edge and help businesses to define and redefine their policies and strategies. This huge and enormous amount of available information threatens human attention, giving new edge and direction for information retrieval technology. The data available on social media sites usually revolves around the trending issues. The notion of a topic has been modified and sharpened to be an "event". An event is some unique thing that happens at some point in time. An event can be defined as "something that occurs in a certain place during a particular interval of time". The term may also refer to a significant occurrence or happening, or a social gathering or activity [HYPERLINK \l "RJa081" 1]. WordNet defines an event as "something that happens at a given place and time2]**}**. The notion of an event differs from a wider category of events both in spatial and temporal localization. For example, Droughts that hit Latoor, Maharashtra in 2016 is an event whereas Droughts in general is a class of events. An event can be physical as well as virtual. It is not necessary that all the geographical and temporal information will be available with every event. Events can be expected as well as unexpected. Earthquakes, landslides, murder etc. fall under the category of unexpected events whereas political elections,

meetings and concerts fall under the category of expected events.

Event identification is the problem of identifying stories in several continuous news streams that amount to a new or previously unidentified event [HYPERLINK \l "JAl982" 3]. Identifying an event from accumulated collection may either result in retrospective identification or online identification. Retrospective identification refers to the discovery of already available unidentified events. Online identification refers to the identification of new events from live news feeds in an online manner from multiple document stories. In last few decades, the rapidly increasing number of electronically available news reports threatens to be overwhelming. All the events which are discussed on social media sites are stored as news streams. News stream contains information regarding different or similar events. The available amount of news streams forced the researchers to analyze the news streams and the best information out of it. Consider an example where a person was out of his country for a specific time period. Now he wants to know what has happened in the country in his absence. To read the whole news of a month or a year can be a daunting task. He could not even generate any queries because he is not having any knowledge of recent events. So, an individual requires some system which will produce the significant events and their evolution along the timeline. The Topic Detection and Tracking Community has been studying for decades to explore techniques for detecting the appearance of new topics and for tracking the reappearance and evolution of them.

The contributions of this paper can be enlisted below:-

- Find events from a text stream of documents which show the highest reporting on the social media in a single document or more. These news streams can further help to gain insights into the subject matter whenever it further needs discussion.
- News agencies can correlate the incoming event with the existing one and can go on the fly.
- News agencies can categorize the documents relating to the specific events. Thus headlines of the new stories can be retrieved in a simple and elegant manner.
- It greatly enhances user navigation.

The rest of the paper is organized as follows: - Section 2 describes related work on event detection in social media. Section 3 describes in detail all the facets of the proposed methodology. Section 4 presents the technology used to conduct the experiments. Section 5 discusses the results. Section 6 summarizes the work and concludes the paper. Section 7 describes the future scope of the proposed work.

2. RELATED WORK

A news stream can be defined as a continuous transmission or flow of information related to one or more topics with specific news updates as well. The news transmissions are done through web} sites or through a syndicated news service provider. Users subscribe for the news and then receive the news feed or web feeds. A news event is something that happens at a certain time in a certain place, which may be reported consecutively by many news reports over a period of time [4]. Event detection is an unsupervised learning task. The TDT community focused on retrospective news event detection (RED) [5] and new event detection (NED) [6]. RED detects previously unidentified events in a news corpus. [5] NED detects news stories about previously unseen events in a stream of news stories. The most popular approach of NED is based on the initial work of Yang et al. [5] and Allan et. al [6]. They compared every new incoming report with already received and existing ones. If the incoming report does not match any of the existing ones, it became the first story of a new event. Many researchers have used named entities (such as person, organization, location, date, time, etc.) in their NED methods to improve accuracy [7]. A multi-resolution indexing scheme has been proposed to online discover meaningful behavior and monitor this over variable window sizes [8]. Based on ratio aggregation pyramid (RAP) and slope pyramid (SP) data structure, algorithm can also detect bursts in multiple window sizes. Effective ways have been presented for identifying periodicities and bursts in the query logs of the MSN search engine [9]. The evolutionary patterns of themes have also been summarized in a text stream [10]

3. PROPOSED WORK

A news stream consists of news stories related to a particular topic. A topic is a seminal event or activity, along with all directly related events and activities. A story is a topically cohesive segment of news that includes two or more declarative independent clauses about a single event. A news stream consists of a set of new stories. Each story consists of a set of documents. A news story could be represented as S = $\{d_1, d_2, \ldots, d_i, \ldots\}$, where d_i is a document at some point of time t_i. Each document consists of a set of features in asset of vocabulary F = $\{f_1, f_2, \ldots, f_i, \ldots\}$. News

stream consists of thousands of features. Temporal representation of a feature can be defined as trial of a feature. Trial of a feature can be written as a discrete time series f_i[1,2,3,.....t], where each element f_i [t]denotes the feature f_i at some specific point of time t. The specific point may be one hour, half a day or one day. In this proposal, time period is assumed to be of one day. The main objective is to detect bursty features which will lead to the detection of bursty event. Bursty event is the minimal set of bursty features that occur together in a certain time window in a stream of documents [11]. Bursty Feature can be defined as the feature whose aggregate value of feature trial is greater than other features in documents within a certain time window. Aggregate used in this proposal is the sum of its feature trials. After detecting bursty features in the current window, bursty events will be identified. Bursty events could be represented in the current window as E = {e1,e2,e3,.....,ei,...}. After finding bursty events in the current window, closely related events could also be found. Closely related events could be well described as the events having concurrent evolution. This is known as bursty event's evolutionary trial. The steps involved in the proposed work are as under:-

- Data Preprocessing:- Data is available in the SGML format. Removal of noise is done. Those tags which do not produce any fruitful information are discarded.
- Data conversion:- Data is converted into the textual format.
- Keyword selection:- Keywords are generated by recording the occurrences of each word in the news stream.
- Removal of stop words:- Stop words are discarded as their frequency of occurrence is quite high.
- Bursty Feature Selection:- bursty features are selected as those keywords whose frequencies are above certain thresholds.
- Bursty Event Selection:- Bursty event clusters bursty features in it and are thus identified.
- Bursty Event's Evolution:- Two closely related events can be found by tracing their evolution and their relative strength to specific documents.

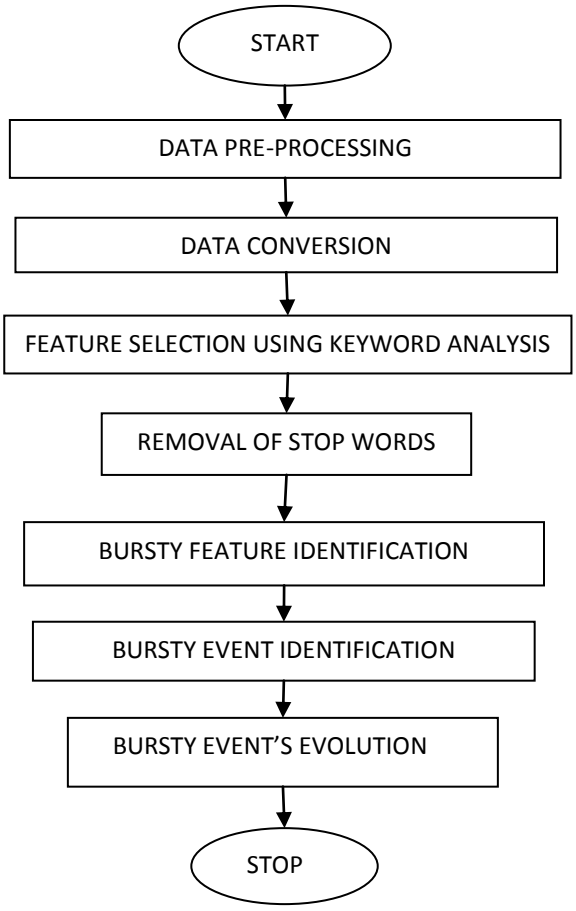

Figure 1: Steps for the online identification of news streams

4. Research Methodology

Data used to test our approach has been taken from the Reuters Corpus Volume1. RCV1 was used to evaluate proposed methods [12]. RCV1 is an archive of 806791 manually categorizes newswire stories made available by Reuters Ltd. for research purposes. These news stories are from '96/8/20' to '97/8/19' (totally 365 days) using a daily resolution. Data set was made available in SGML format. Only the text in <title> and <text> fields were processed. Remaining tags were discarded. Then the data in SGML format is converted into textual format. It becomes easy to identify events from textual stream. Keyword Analysis is done to extract the features. Words having higher frequencies are extracted. Then fluffy words and stop words are removed. Then after, bursty features are extracted in certain time window with different thresholds.

Bursty features help us to cluster bursty events. Closely related events can be found in the same window size and their evolution can be traced. Different size of sliding window reveals different number of features. All the experiments were carried out in Matlab and Java.

5. RESULTS AND DISCUSSIONS

In this approach, 21,578 files are taken from RCV1 to efficiently and effectively categorize the news streams based on the retrieved bursty features. Varying size of sliding window produced different number of bursty features. It is observed that as we compress the window size, the features extracted are numerous. In contrast as the window size is expanded, the generation of bursty features starts reducing. In this approach, sliding window size with values of W= 2, 4, 6, 8 and 10 are taken. The number of detected bursty features is approximately inversely proportional to the size of the sliding window. For example, the number of the features in the case W=2 is nearly double that in the case W=4. The number of features generated with varying sliding window sizes is shown in the following figures.

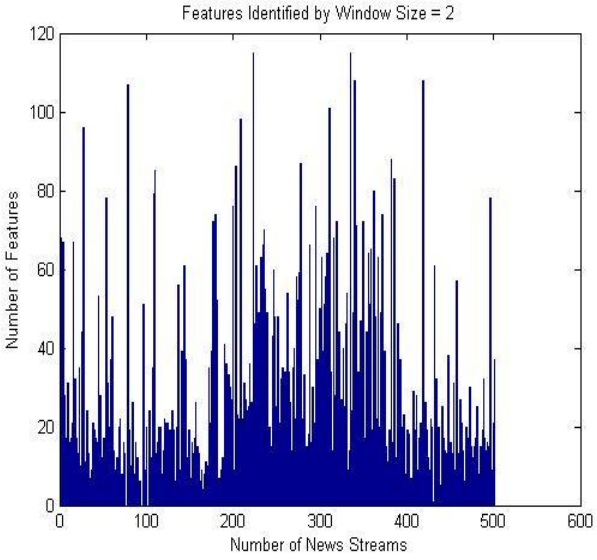

Figure 2: Bursty Features identified with sliding window W=2

To effectively and efficiently evaluate the bursty features, 21,578 files were taken. Figure 1 shows that when the sliding window size is least i.e, W=2, the features generated are maximum. The features clustered are approximately 120 within sliding window size W=2.

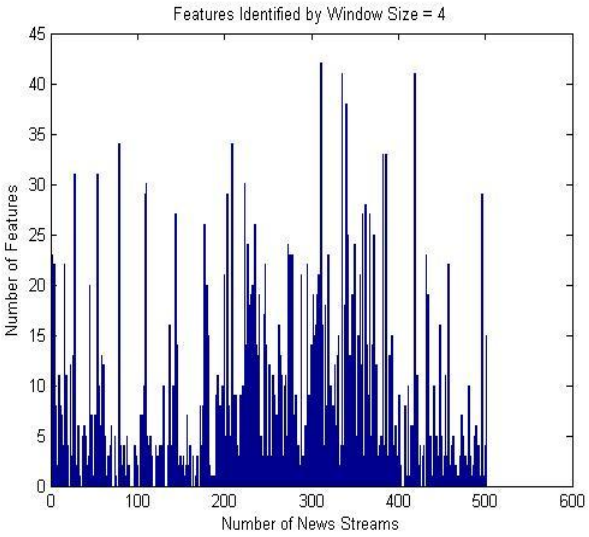

Figure 3: Bursty Event Identification with sliding window size W=4

To check the effect of varying window size, similar files were mined to generate the bursty features. In Figure 3, the window size was expanded to 4. With this window size, the features were approximately 45. This number is very low when compared with W=2.

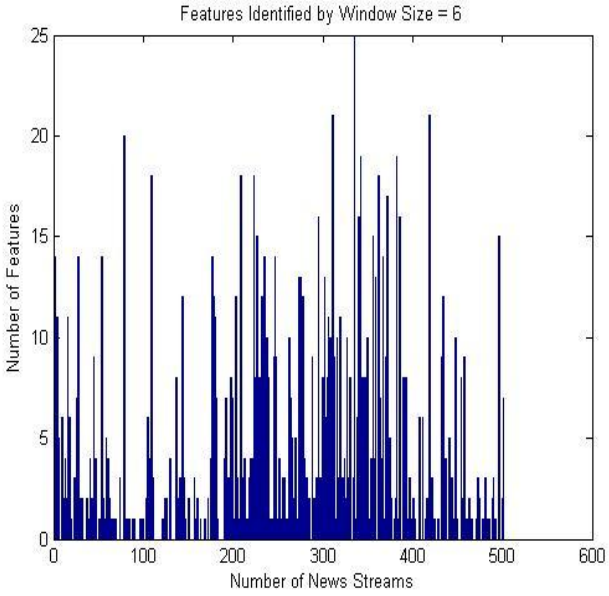

Figure 4: Bursty Event Identification with sliding window size W=6

21,578 files were mined to generate the bursty features. In Figure 4, the window size was set to 8. With this window size, the features were approximately 16. The number of features generated are very less in number.

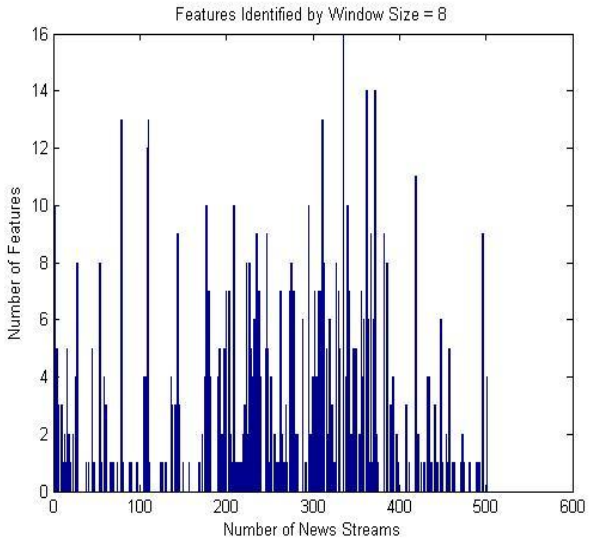

Figure 5: Bursty Event Identification with sliding window W=8

In Figure 5, 21,578 files were evaluated extensively with sliding window size W=10. This window size showed least number of features as compared to that of window size W=2, 4 and 6.

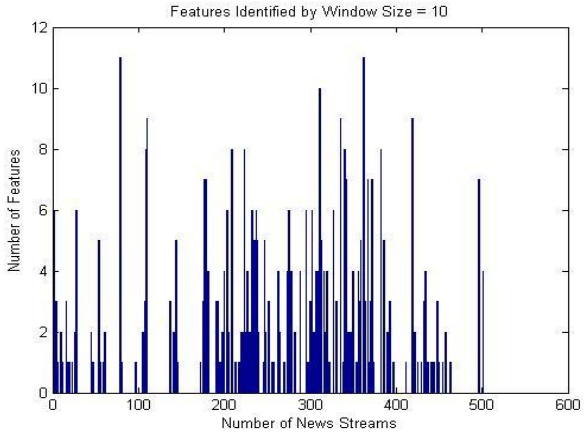

Figure 6: Bursty Event Identification with sliding window size W=10

In Figure 6, 21,578 files were evaluated with sliding window size W=10. This window size showed least number of features as compared to that of window size W=2, 4,6 and 8. The features identified are approximately 12.

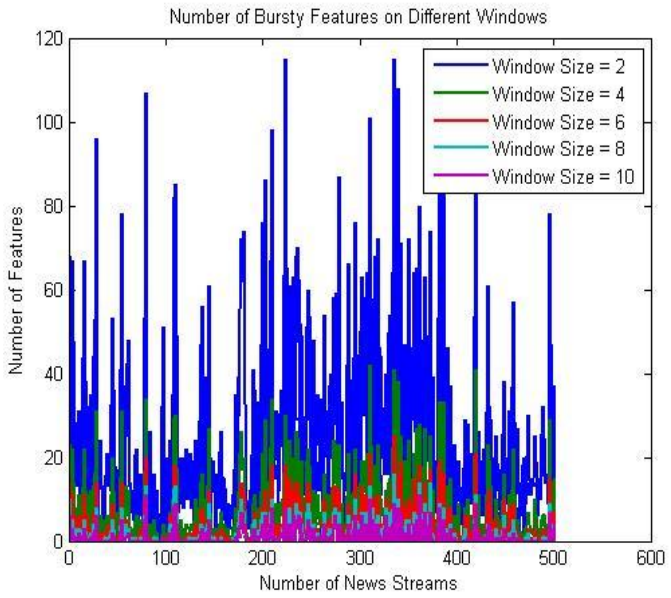

Figure 7: Bursty Event Identification with varying sliding window sizes.

Figure 7 shows the comparative study of different and varying window sizes. It shows the effect of window size on the number of bursty patterns.

6. CONCLUSION

In this paper, online detection of bursty events is studied. Bursty event is defined as a group of bursty features that have high correlation among them. Correlation among bursty features is calculated by considering the similarity between the existing bursty patterns and the incoming stream of data. The relation between closely related events has been found by tracing their evolution. If both the events correlate to similar document set, they are said to be closely related.

Proposed method is evaluated using Reuters Corpus Volume 1, which consists of 806791 news reports over one year. Experimental results show that the proposed method can mine meaningful features hidden in the news stream. It can also detect the evolution of news stories. It greatly facilitates user navigation.

7. FUTURE SCOPE

Many interesting aspects of the proposed work can be searched further. A news recommendation or retrieval system can be built to help users navigate in the news information space. Better methods should be proposed to make the event detection a supervised learning method and make the work more practical. It can assign labels automatically for detected events then after. There exists some interesting knowledge hidden in news streams that can be addressed in future work, such as events' association and events' periodicity.

REFERENCES

[1] R. Jain, "Eventweb: developing a huma-ncenterd computing system," *computer*, vol. 48, pp. 42-50, 2008.

[2] R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller G.A. Miller, "Introduction to wordnet:an on-line lexical database," *lexicogr*, vol. 3, pp. 235-244, 1999.

[3] J. Carbonell, G. Doddington, J. Yamron, Y. Yang, J. Allan, "Topic detection and tracking pilot study final report," in *Proceedings of the DARPA broadcast news transcription and understanding workshop,DARPA*, 1998, pp. 194-218.

[4] (2007, August) Topic Detection and Tracking Evaluation (TDT) Project. [Online]. http://www.itl.nist.gov/iad/mig//tests/tdt/

[5] Y.M., Pierce, T., Carbonell, J.G., Yang, "A Study on Retrospective and On-line Event Detection.," in *Proc. SIGIR Conf. on Research and Development in Information Retrieval*, 1998, pp. 28-36.

[6] J., Papka, R., Lavrenko, V., Allan, "Online event Detection and Tracking," in *Proc. SIGIR Conf. on Research and Development in Information Retrieval*, 1998, pp. 37-45.

[7] W., Meng, H., Wong, K., Yen, J., Lam, "Using contextual analysis for news event detection.," *Int. J. Intell. Syst.*, vol. 16(4), pp. 525-546, 2001.

[8] A., Singh, A.K., Bulut, "A Unified Framework for Monitoring Data Streams in Real Time.," in *Proc. 21st Int. Conf. on Data Engineering*, 2005, pp. 44-55.

[9] M., Meek, C., Vagena, Z., Gunopulos, D., Vlachos, "Identifying Similarities, Periodicities and Bursts for Search Queries," in *Int. Conf. on Management of Data*, 2004, pp. 131-142.

[10] Q.Z., Zhai, C.X., Mei, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining.," in *Proc. 11th ACM SIGKDD Int. Conf. on KnowledgeDiscovery in Data Mining, p.198-207.*, 2005, pp. 198-207.

[11] G.P.C., Yu, J.X., Yu, P.S., Lu, H.J., Fung, "Parameter Free Bursty Events Detection in Text Streams.," in *Int. Conf. on Very Large Data Bases*, 2005, pp. 181-192.

[12] D.D., Yang, Y.M., Rose, T.G., Li, F., 2004. Lewis, "RCV1: A new benchmark collection for text categorization research.," *J. Mach. Learn. Res.,* vol. 5, pp. 361-397.