

EVENT MONITORING AND SENTIMENT ANALYSIS SYSTEM OVER TEXT STREAMS

Authors

N.Bharathiraja^{1*}, N.V. Ravindhar², S.Sandhya³, R.Abhishek⁴, M.Shanthi⁵

^{1,2}Assistant Professors / CSE, Saveetha Engineering College, Anna University Chennai.

^{3,4,5}UG students / CSE, Saveetha Engineering College, Anna University Chennai.

*rajamesoft@gmail.com

ABSTRACT

Micro blog stages have been strikingly famous in the gigantic data time assignment as a result of its predictable scrambling of information. It is central to consider the whimsical events that are inclining on the accommodating connection and can screen their advance and find related whimsies. In this paper we propose an enduring rising irregularity surveying structure humbler scale blog content streams. It joins the undertakings on both rising irregularity watching system and structure inquire about. This gives a graphical sensible approach to manage regulate perceive making assortments from the standard at a starting time. It is among the first to discover rising irregularities relationship spoutingly. This can organized to screen the advance of erraticisms at different time scales from minutes to months. It revives time-fluctuated catchphrase request execution of a full-content web record to improve the capacity of watching mannerism development. It in like way improves the dynamic chart getting ready execution of Spark and recognizes design stream appear on it. It can process giant data to the entire Twitter content stream and to find positive negative scores by mindful examination over little scale blog stages to ensure versatile deficiency.

Keywords: *Event detection, Indexing, Monitoring, sentiment analysis, trending tweets.*

1. INTRODUCTION

Micro blog stages have been an extremely main stream of huge information because of its viral dispersion of data. An immense number of odd occasions rise up out of such stages extending from political or day by day issues to catastrophic events or some other open hazard. These stages are generally used to be the first to report any sort of occasions like earth tremors or mishaps or even any presidential crusades. Be that as it may, it might have numerous collapsed sees, might be certain or negative. It is important to distinguish and react to crisis occasions in a convenient way and track its development and abridge the slanting occasions. The current framework doesn't give an early location of such abnormal occasions. This framework performs superior to anything the current techniques in the occasion checking undertakings.

[7]Twitter Monitor gives online recognition to developing atypical occasions yet couldn't uncover its other perspective or give relationship investigation into distinguished occasion. CLEAr gives constant bizarre occasion recognition and following however couldn't give relationship investigation of them. Neither of them could distinguish potential bizarre occasions previously they are prevalent and spreading at scale. [5]Signitrend could recognize potential strange occasions at little scale yet is confined to just location, and couldn't give following and connection investigation. The strategy would likewise have a tendency to produce peculiar occasions that exclusive contain a solitary watchword as depiction, which is difficult to understand for clients. Nothing unless there are other options strategies furnish the even versatility with appropriated executions of their calculations, nor do they explore framework advancements for their applications. In spite of the difficulties, there are numerous

coveted highlights when constructing a developing abnormality checking framework over short content stream.

Our major contributions in this paper include:

- We adopt anomaly detection method to monitor each keyword for early detection of trends. The proposed graph stream model is fully distributed with an efficient context statistics maintenance strategy and linear scalability.
- We provide a scalable anomaly monitoring approach meeting all the listed requirements. Especially, we are among the first to provide detailed correlation analysis of anomalies under the real-time emerging anomaly monitoring scenario.

In Figure 1, we present a real-time emerging anomaly monitoring system over micro blog text streams. Emerging anomaly monitoring has attracted much attention from the research domain. Here we aim to monitor emerging anomalous events on micro blog platforms. Our emerging anomaly monitoring methods are based on graph mining techniques, which provide unique opportunities to integrate our emerging anomaly monitoring research and system optimizations. In the system, emerging anomaly monitoring includes early detection, correlation analysis and temporal evolution tracking of anomalous events. Early detection would capture emerging events before they go viral. Correlation analysis would automatically reveal multiple aspects of the anomalous event, or the causality of anomalous events, or categorical structure of related anomalies. For example, the history and current development of a political event reveal multiple aspects of the event. The capture of a criminal and the crime he had committed for relationship. Different genres of news from an agency detected at the same time would reveal categorical information. According to its popularity, anomalous events would emerge from different time granularities, e.g., a publicly concerned long trial or an overnight pop concert. The temporal evolution tracking of events could recover the evolution process of an anomalous event, to trace its origin and get the big picture. Such monitoring happens in real-time and provides valuable intelligence for government agencies, news groups and marketing agencies, etc.

Algorithms are implemented to have linear horizontal scalability to handle big data, i.e., full stream of Weibo or Twitter data. The full-text indexing engine is optimized for efficient time range queries, which benefits our evolution tracking algorithms and queries over events and tweets.

2. RELATED WORK

Micro blog services have developed as a vital method to keep up the interchanges among people and associations. These services advance convenient and dynamic discourses and remarks towards items, markets and open events, and have pulled in a considerable measure of considerations from associations. Specifically, developing themes are of quick worries to associations since they flag current worries of, and input by their clients. Two difficulties must be handled for compelling rising subject discovery. One is the issue of continuous important information gathering and the other is the capacity to display the rising attributes of identified points and recognize them before they end up hotly debated issues. To handle these difficulties, we first outline a novel plan to slither the applicable messages identified with the assigned association by checking multi-parts of miniaturized scale blog content, including clients, the developing watchwords and their transient grouping. We at that point build up an incremental bunching system to distinguish new subjects, and utilize a scope of substance and worldly highlights to help in immediately identifying hot developing themes. Broad assessments on an agent certifiable dataset in light of Twitter information show that our plan can describe developing points well and distinguish them before they wind up interesting issues.

Twitter has become one of the largest platforms for users around the world to share anything happening around them with friends and beyond. [3]A bursty topic in Twitter is one that triggers a surge of relevant tweets within a short time, which often reflects important events of mass interest. How to leverage Twitter for early detection of bursty topics has therefore become an important research problem with immense practical value. Despite the wealth of research work on topic modeling and analysis in Twitter, it remains a huge challenge to detect bursty topics in real-time. As existing methods can hardly scale to handle the task with the tweet stream in real-time, we propose in this paper Topic Sketch, a novel sketch-based topic model together with a set of techniques to achieve real-time detection. We evaluate our solution on a tweet stream with over 30 million tweets. Our experiment results show both efficiency and effectiveness of our approach. Especially it is also demonstrated that [12]Topic Sketch can potentially handle hundreds of millions tweets per day which is close to the total number of daily remote tweets in Twitter and present bursty event in finer-granularity.

Social media such as Twitter are some popular source for live textual data. Much of this popularity is due to the fast rate at which this data arrives and there are a number of global events – such as the Arab Spring – where Twitter is reported to have had a major influence. However, this existing methods for emerging topic detection are often been able to detect events around the globe such as natural disasters like earth quakes or celebrity deaths, and can monitor user-selected keywords or operate on the hash tags only. [2]Earth quakes shakes can be detected with such twitter analysis. Interesting emerging topics may be of much smaller magnitude as it may involve the combination of different words which are not unusually hot in the globe. Our contributions to the monitoring of emerging trends are threefold: first of all, we propose a essential measure that could be used to detect emerging topics early, long before they become “hot tags”, by drawing upon experience from outlier detection. Secondly, by using hash tables in a heavy-hitters type algorithm for establishing a noise baseline, we can show how to track even all keyword pairs using only a fixed amount of memory. Finally, we aggregate the detected co-trends into larger topics using clustering approaches, as often as a single event will cause multiple word combinations to trend at the same time.

We show a novel exploratory visual systematic framework called TIARA (Text Insight through Automated Responsive Analytics), which consolidates content examination and intuitive perception to enable clients to investigate and dissect substantial accumulations of content. Given a gathering of records, TIARA utilizes a few intelligent content representation procedures to clarify the rundown comes about and consistently connection such outcomes to the first content. We have connected TIARA to a few certifiable applications, including email outline and patient record examination. Our trial results and introductory client input propose that TIARA is powerful in helping clients in their exploratory content logical undertakings.

3. SYSTEM ARCHITECTURE

3.1 Data Collection: A scattered crawler is tried to assemble data from Twitter, the best cut back scale blog deal with on the planet. The crawler reliably totals the latest little scale web

diaries scattered by customers preferably with a liberal number of disciples, i.e., feeling pioneers. The crawler has a master/slave format. The ace concentration uses key/regard store to perform undertaking arranging. Slave center concentrations get the assignments from the ace and crawl data. An endeavor would screen the reposts and comments of an extraordinary tweet and recoup the repost and comment list, from which we can build up the forward chart of each tweet. The errands are sorted out by posts' need, which is weighed by the measure of reposts and comments. Every errand has a proximity cycle with the objective that the checking for each post can be feasibly kept up or wrapped up

3.2 Data Processing: For cut back scale blog stages, steadfast search for after is visit. It's for the reason that most customers are amped up for microblogs posted in late time. More on an amazingly fundamental level, history look is basic to event change following application. The application issues wide measure of watchword demand to recoup related tweets in the substance offer to for equivalence estimation. Each request is issued for a specific time between time as portrayed in the event approval count. Both consistent request and history interest can be directed as the eagerness with a specific time go, with various time relate for question.

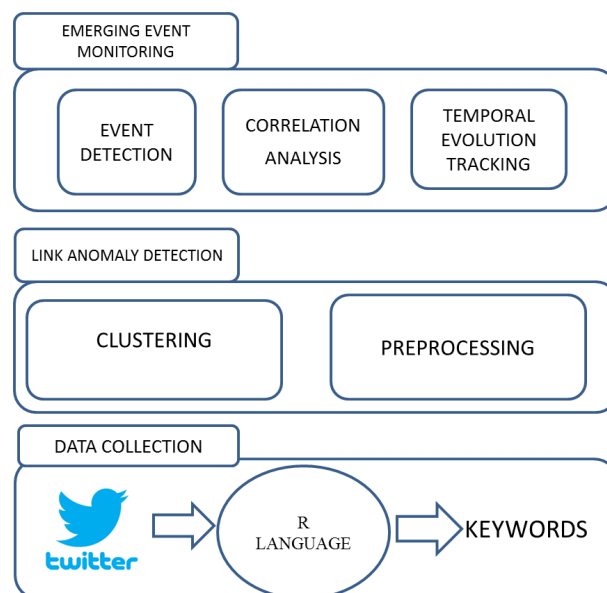


Figure 1: Architectiure of the Anomaly Monitoring System over Text Streams

3.3 Event Detection: The depictions of these occasions are ousted regularly either especially from tweets or from assign tweets that contained predefined fragments indicating appropriate titles.

3.4 Correlation Analysis: We can see making events of strong watchword association with the strategy. We watch that unmistakable parts of an event would make in a lone event. Since the events are ousted from pack closeness relationship, there would exist maximal inside circles in a get-together that are not clearly close to each other. This showcases potential masterminded parts of an event. These edges may consolidate unmitigated information, causality affiliations or particular ideas on a comparative event. Every so often these perspectives could just co-occur by shot. To decrease co-occasion by probability, truncate event edges into sub-events and make an other leveled point of view of the event.

4. IMPLEMENTATION AND RESULT

The screenshot displays the 'Scalable Uncertainty-Aware' web application. At the top, there's a navigation bar with 'E4-Engines - Shiny' and a 'Shiny' logo. Below this, a breadcrumb trail shows 'shiny.rstudio.com > Open in Browser'. The main heading is 'Scalable Uncertainty-Aware'. The interface is divided into several sections: a search bar labeled 'Enter data to be searched with "V"', a date range selector for 'Number of recent tweets to use for analysis:' (set to 100), and a list of 'Top Trending Tweets Today'. The trending tweets list includes topics like #WinterChills, #VermontSubzero2018, #Tatort, #SDLive, DO NOT CONGRATULATE, FELICES ON THE, Goodell, Marissa Morris, So a Chicka, Cato, Day Jams, and Jayson Tatum. The interface is designed to allow users to explore uncertainty-aware data analysis results for various locations and time periods.

[illegible]

Word cloud is the most used words associated with the hash tags. It consists of huge number of words that are extracted from the tweets in the twitter. Word cloud is the visual representation of text data typically used to depict keyword metadata (tags).on websites or visualize free form text. This Figure 3 is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence.

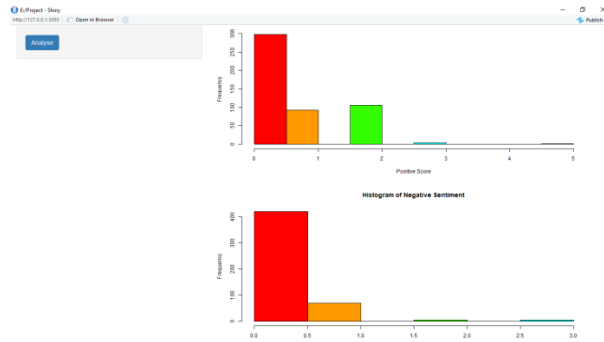


Figure 4: Histogram

Figure 4 shows the graphical representation of the negative and positive sentiment. [20]Sentiment Analysis is the technique used here. This graph is plotted between frequency and positive or negative score. It ranges from 1 to 6 scales.

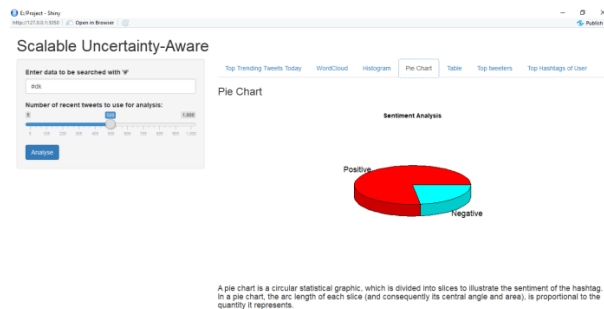


Figure 5: Pie Chart

The Figure 5 is a pie chart that shows a graphical view in a statistical manner which is divided into slices to illustrate the sentiment of the hash tag. In a pie chart, the arc length of each slice is proportional to its quantity it represents.

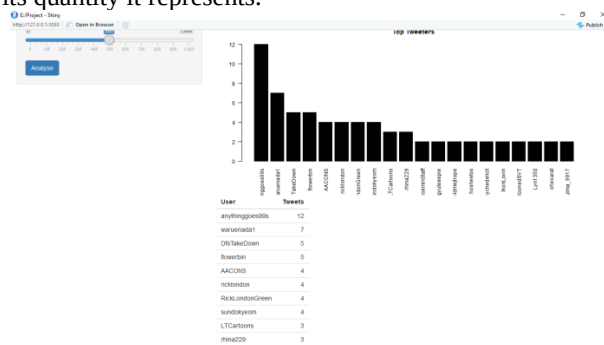


Figure 6: Top Tweeters

The Figure 6 represents the top tweeters of the given hash tag and their number of tweets is also displayed.

5. CONCLUSION AND FUTURE WORK

We have shown a relentless rising occasion assessing framework little scale blog organizes that combines our endeavors on both making occasion watching examination and structure explore. It can screen rising occasion as to recognize rising occasions, assemble occasion affiliations and take after occasion upgrades. Further, this foundation is outfitted with changed update on its full-content web searcher and scattered outline preparing motor to perform occasion checking all the more efficiently. The framework unmistakably exhibits its slants over existing structures and techniques from both specific and framework points of view forward making occasion watching errand. It besides gives extra bits of information about the negative and positive suppositions in a graphical approach close-by the best tweeters inconspicuous parts.

Later on, information examination through goliath information in electronic frameworks organization content streams has upgraded understands the little scale blog stages. We would organize investigate on redid portraying and envolving learning base into our outcomes to give better data.

6. REFERENCES

- [1]. Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., & Lin, J. (2012, April). Earlybird: Real-time search at twitter. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on* (pp. 1360-1369). IEEE.
- [2]. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). ACM.
- [3]. Guzman, J., & Poblete, B. (2013, August). On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model. In *Proceedings of the acm sigkdd workshop on outlier detection and description* (pp. 31-39). ACM.
- [4]. McCreadie, R., Macdonald, C., Ounis, I., Osborne, M., & Petrovic, S. (2013, October). Scalable distributed event detection for twitter. In *Big Data, 2013 IEEE International Conference on* (pp. 543-549). IEEE.
- [5]. Schubert, E., Weiler, M., & Kriegel, H. P. (2014, August). Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 871-880). ACM.
- [6]. Chen, C., Li, F., Ooi, B. C., & Wu, S. (2011, June). Ti: an efficient indexing mechanism for real-time search on tweets. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (pp. 649-660). ACM.
- [7]. Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1155-1158). ACM.
- [8]. Cai, H., Huang, Z., Srivastava, D., & Zhang, Q. (2015). Indexing evolving events from tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 27(11), 3001-3015.
- [9]. Yu, W., Aggarwal, C. C., Ma, S., & Wang, H. (2013, December). On anomalous hotspot discovery in graph streams. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on* (pp. 1271-1276). IEEE.
- [10]. Huang, C., Wang, D., & Chawla, N. (2017). Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems. *IEEE Transactions on Big Data*.
- [11]. Long, R., Wang, H., Chen, Y., Jin, O., & Yu, Y. (2011, September). Towards effective event detection, tracking and summarization on microblog data. In *International Conference on Web-Age Information Management* (pp. 652-663).

Springer, Berlin, Heidelberg.

- [12]. Xie, W., Zhu, F., Jiang, J., Lim, E. P., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8), 2216-2229.
- [13]. Saha, A., & Sindhwani, V. (2012, February). Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 693-702). ACM.
- [14]. Allan, J., Papka, R., & Lavrenko, V. (1998, August). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 37-45). ACM.
- [15]. Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1), 132-164.
- [16]. Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the Association for Information Science and Technology*, 60(11), 2169-2188.
- [17]. Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675-684). ACM.
- [18]. Stieglitz, S., & Dang-Xuan, L. (2012, January). Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (pp. 3500-3509). IEEE.
- [19]. Younis, E. M. (2015). Sentiment analysis and text mining for social media microblogs using open source tools: an empirical study. *International Journal of Computer Applications*, 112(5).
- [20]. Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejó-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, 20(1), 1-28.

