JCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-7, July 2019 E-ISSN: 2347-2693

Anomaly Detection In Practice Using Python

Shirishkumar Bari[1*], Abhijit Patankar[2]

[1]Department of Computer Science, Alard College Engineering and Management, Pune, India
[2]Department of Information Technology, D.Y. Patil College of Engineering Aakurdi, Pune, India

Corresponding Author: shirishkumargbari@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i7.241246 | Available online at: www.ijcseonline.org

Accepted: 13/Jul/2019, Published: 31/Jul/2019

Abstract - On 8th August 2018, Kerala had a very heavy rainfall, resulting filling of dams caused flood situation in Kerala. Many people started posting twits about this and people living in that area were alerted. Administration department started their rescue operations. Here social media played key role in locating people and providing help to them. A lot of campaigns were started to collect financial aid to the affected people. Here we again felt power of social media that can positively impact the society.

Twitter, a popular micro blogging service, has received much attention recently. An important characteristic of Twitter is its real-time nature. It is also extremely popular because the information gets spread more widely and rapidly.

It's important to detect anomalous events which are trending on the social media and be able to monitor their evolution and find related events. This paper talks about how to detect the anomalies in tweets.

Keywords— Anomaly, Types of Anomalies, Machine Learning, Text Stream, Twitter Data, Social media analysis

I. INTRODUCTION

Social media is plays great role in people's life. Social media can act as effective medium of communication across the globe. Social media has large set of application in public sector, NGO's, political parties and governments etc. Social Media can act as great platform for marketing and advertising. Social platforms are synergetic web applications or mobile apps like Facebook and Twitter etc.

Due to the social platform and application like above there is might be positive and negative impacts on us. The popular component and feature of Twitter are re-tweeting we can say in other word is reply on some points via messages. Twitter allows people to keep up with important occasions and always stay connected with their peers and can contribute their opinion and thoughts in various ways throughout social platform. Re-tweeting is favourable or advantageous strategy for us because it reports or notifies individuals on Twitter about popular trends, posts and occasions. Based on these popular trends some abnormal things are happen in this media or in other word we can say that the anomaly. To avoid this things/anomaly we use the new technique which is shown in our paper. In this paper we are introducing anomaly monitoring system over text stream data. This system will be monitoring the online incoming text data and find out the appeared topics with abnormal things. Early detection of anomalous event in tweets is very beneficial [1]

To understand how big social media platforms have expanded its reach we can refer to dreamgrow[15] data, it has published the list of top 15 platforms according to active number of users refer ***Image1.*** Facebook has 2,230,000,000 active users and Twitter has 336,000,000 active users per month(at the time data was collected).Which clearly demonstrates huge impact of that can be caused.

There is a need of system which has capability to alert to prescribed users in case anomaly is detected on twitter. For e.g. system can alert Police if anti-nation, crime related tweets/post are published. It can alert government if natural calamity related posts are published like flood / earthquake.

- **Methods for Anomaly Detection**
  Several methods have been designed to detect anomalies like Model based, Clustering Based, Graph Based Approaches Each approach has different advantages and challenges.[2]

➢ **Different Types of Anomalies** [3]

1. Collective anomalies: When you analyze set of data to determine the anomalies is called collective anomalies. Example - If you are receiving thousands of emails in an hour it could be potentially spam attack.

2. Point anomalies: When a single data item is very different from the rest of the dataset. Example - A transaction done

from a totally new location it could be a fraudulent transaction.

3. Contextual anomalies: When the data can be categorized as anomalous under certain situations only. Example - A person spending 5000rs on shopping during weekend could be a normal but if someone spends similar amount in a shopping mall during office hours could be anomalous transaction.

- **Basics of Anomaly Detection**

Definitions: [4]

"*Anomaly detection is a process to the identify items or events that do not conform to an expected pattern or to other items present in a dataset. Typically, these anomalous items have the potential of getting translated into some kind of problems such as structural defects, errors or frauds.* "

"*Anomaly detection (also outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions*

➢ **Techniques for anomaly detections**[5]

These techniques can be broadly categorized into Supervised and Unsupervised techniques.

1. Supervised Learning: It requires a training dataset with identified labels for normal and anomalous data point. It creates a classifier model from these labels. New data set would be evaluated based on the model, it will predict the class to which the data would belong.

- Support Vector Machine:

This technique identifies the soft boundaries from training dataset which forms clusters of normal data. The data points outside these boundaries would be considered as anomalous data.

- K -Nearest Neighbor (k-NN):

It classifies a data point with respect to its neighbor's classification. The unlabeled point will be assigned to the class of its K-nearest neighbors based on the approximate distance's calculation between various points. Please keep in mind that selection of value of *k* is very crucial part here. We should try with different values and then select the best one.

2. Unsupervised Machine Learning: This technique doesn't require labeled training data. It has an assumption that the normal instances are more frequent than anomalous instance so it can label output groups as normal. It often targeted to search groups of instances having same characteristics and similarities based on similarity measure. These techniques build a model which is used for decision making.

- Clustering Techniques

These techniques work on principle of grouping data into groups(clusters). We can group data based on Distance or Similarity. There exist many different variations of these techniques like Unsupervised Neural Network, K-Means, Fuzzy C Means (FCM), Unsupervised Niche Clustering (UNC) and

- One - Class Support Vector Machine (OCSVM)

This technique works on principle that we should try group all training data into one class and anything which does not fit into this class would be outlier. This technique has gain popularity now a days.

II. LITERATURE SURVEY

Social Monitoring System for Dynamically Evolving Anomalies over Text Streams,[6] this paper explains a blog having a small group of members with their discussion. These blogs are smaller than a traditional blog and contains very short entries is called as micro blog. Using this real-time diffusion of information is going on. So, in that so many abnormal discussions on micro blog, hence we can analyze and monitor these things which are currently very popular on social Medias and can trace the related anomalies. The new users register on the social Network like twitter and then login. The user here has to post their comment on the social network and be able to monitor their evolution and find related anomalies and the system then retrieves the anomaly thing of that particular post from database which the user has posted and these posts are displayed in the list and detect all anomaly events in trends which store into database. So, for every post there is a different list and there is also facility for user to view all the post. Here the user can see the post in their system.

Second paper studied is Earthquake shakes twitter users [7]. In this Paper, As discussed earlier, we can use twitter to get a real time feel is to what is happening around the world per seconds. We can track various events from any one of this social media platform in the data streams e.g. identify the various accidents and automatically report the road accidents for emergency services. It is to be useful; events need to be identified within the stream with a very low potential. Due to the high volume of posts within these social streams this is challenging.

The explosion of much of big databases and the World Wide Web must be created extraordinary opportunities for mainly monitoring, analyzing and predicting global economic, political, geographical, medical, demographic and most of other processes in the world. Although, despite the huge amount of data being available, specific events of interests are still rare. This type of rare events, mostly called anomalies, which are known as events that happen much in frequent way (their frequency ranges are between 5% to 0.01% based on which application we are using). Detection of anomalies (rare events) must be gained recently a lot of attention in much domains, ranging from detecting fraudulent transactions and intrusion detection to engineering health management (prognostics and diagnostics) and direct marketing. Consider an example, in the network intrusion detection domain, the many of the cyber-attacks on the network is particularly a very minor fraction of the total network traffic. Many of the data records corresponding to the failures which may be occur in the specific engines or components that are correspond only to small portion of all the data records collected during the process of monitoring in the prognostics and diagnostics applications. In most of the cases of anomaly detection for e.g. Network malware detection or health of aircraft, we need to continuously receive the data that too real time. This paper is having a unique anomaly detection method which was proposed for the detection of anomaly over notion drifting statistics streams. In this providing almost equivalent performance to the static LOF algorithm while with the many of lower cost This is showing that detection of anomaly over the stream data which only requires limited frequently of concept-drifting and so the model update time complexity which is not based on the total number of the instances in the data streams.

The real time anomaly detection in most extensive area checking of smart grids is very difficult to improve the reliability of power systems. Although, capturing the characteristics of anomalous interruption and then finding them at real time is become hard for wide-scale smart grids, due to the measurement data volume and complication rises severely with the exponential growth of data from the huge intelligent monitoring gadgets to be pulled out and the need for fast information renewal from those mass data. Many of existing anomaly detection methods was getting failed to handle it well This work proposes a spatial-temporal correlation based anomalous behaviour model to capture the characteristics of anomaly such as transmission line outages in smart grid. We are proposing such an algorithm which will be beneficial for small to large power grid systems where data size will be high. But we are taking care about the time complexity even if we process the large amount of data, the time consumption will be very less compared to the existing systems.

III. PROPOSED METHODOLOGY

Existing system has its own limitations. That is why to minimize the shortcoming of existing methods, we recommend a distributed online index system for temporal micro blog data. Our emerging anomaly monitoring methods introduces the graph stream model that combines our emerging anomaly monitoring research and system research.
In the existing system represented in this paper, early detection of anomalies, their relations and analysis and tracking of its inception are done through the graph plotting.
The idea is to catch or capture the events at a very early stage before they became viral. In this we can able to get multiple features of the anomalous events which can be found by Correlation analysis. Not only that we can also trace the connectedness of anomalous events, as well as we can or unconditional building of related anomalies**.**

I had faced multiple challenges where I was processing the runtime graphs. The basic challenge is that whenever we tend to modify the graph structure, the time complexity is high. The nodes in the graphs will continuously getting increased as the time has increased or passed. This tend to make the system very slow in processing and thus resulting the analysis of graphs very much time consuming. My prime objective is that whenever working on real time data for anomaly detection the user should get analysis in efficient time.

The second challenge which I had faced was that real time processing of the data is obviously is not on easy level as the graphs which we are plotting goes parallel on a global mode instead of time incremental mode.

Third challenge was even if we register for developer account on twitter it provides only few hundred's live tweets.

To handle these challenges, I have downloaded 10K tweets from Kaggle website and processed it offline.

IV. EXPERIMENT DETAILS

To prove the concepts, I took sample twitter data set from Kaggle.com. I ran offline processing to detect the anomaly using python programming.

Details of software/hardware system:
Software Details:
Programming Language: Python 3 / Java
Application Server: Django server
Operating System: Windows 8
Editor: Notepad ++ / Notepad
IDE: PyCharm IDE

Hardware Configuration:
Intel Core I3(6th Generation), 8GB RAM
Dataset collection site: kaggle.com

As mentioned in below figure 1, I have developed python application which uses below flow to detect anomalies on twitter datasets.

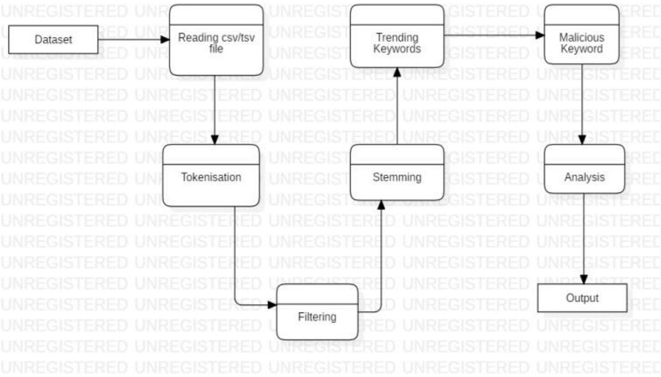


Figure 1. Anomaly Detection Processing Flow

Steps:

1. Dataset Specifications – I have collected 10 thousand tweets dataset from kaggle.com-
2. Datasets - Split the dataset into training and test in which we are using 70-30 common pattern to divide the dataset. 70% of overall data acts as training and remaining are test data set.
3. We are using two machine learning algorithms Naïve Bayes and K –nearest neighbour.
4. For first dataset it takes 8.66896 seconds time to prepare data and 5.04520 seconds to classify data into Naïve Bayes and KNN.
5. For Second dataset It takes 8.34595 seconds to prepare data and 4.84974 seconds to classify into naïve Bayes and KNN. (It can plus minus based on processor u are using.)
6. Overall time required to processing of data and classification of first dataset is 13.71416.
7. Overall time required to processing of data and classification of data is for second dataset is 13.19569 seconds.
8. Accuracy comparison of First dataset gives the 62.73 for both Naïve Bayes and KNN. Also, second dataset gives the 67.14 for both algorithms

- Use Case Diagram

Figure 2 would explain the typical activities that would happen with system. User should have valid credentials to login to system. Once authorized he can use other modules.

User has to download the latest dataset from website. He can then update with the existing dataset. He can then perform actions like tokenizing, removal of stop words stemming, cleaning and finally do the analysis. User should select the option's in the sequence given in the diagram.

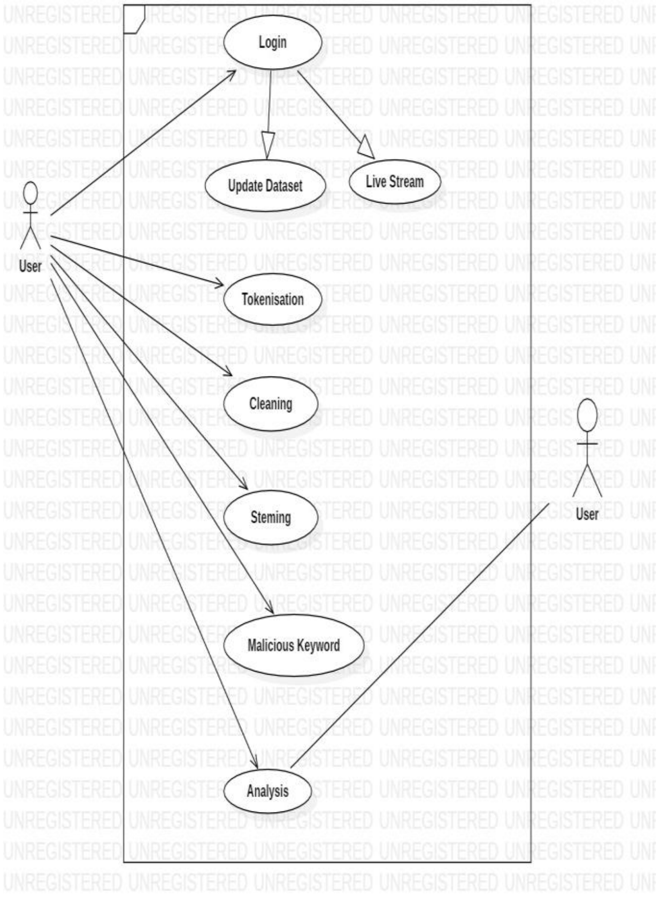


Figure 2. User Case Diagram

- Big Data Handling

As we know amount of data generated is huge and to be able to sustain this data and yet provide the anomaly detection, we should make use of Big Data techniques like parallel processing, Map-Reduce. Below architecture is proposed to handle Big-Data volume. We propose to use Kafka and Flume to process the dataset. This would solve the scalability and performance issues.

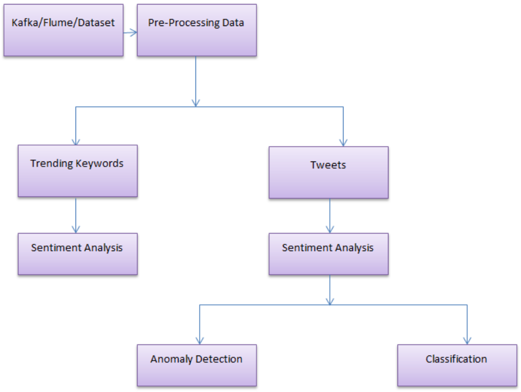


Figure 3. Big Data volume processing with Kafka

Mathematical model for Trending Keywords

We used a normalized version of the metric to assign its value range to [-1, +1]. This avoids negative infinity values and neutralizes its preference for low frequency events. The definition is as follows:

$$NPMI(x,y) = -1 \cdot \frac{1}{log[p(x,y)]} \cdot log\frac{p(x,y)}{p(x)p(y)}$$

where x is a random variable and p(x) is defined as the frequency of x in the document collection. The final NPMI score of an event E is defined as the mean of all NPMI (wi , wj) values of its K keywords:

$$NPMI(E) = \frac{2}{K(K-1)} \cdot \sum_{1<i<j<K} NPMI(w_i, w_j)$$

V. RESULTS

Below are result (Figure 4) for accuracy of different machine learning algorithms for detecting the anomalies on test data. Figure 5. shows result of sentiment analysis.

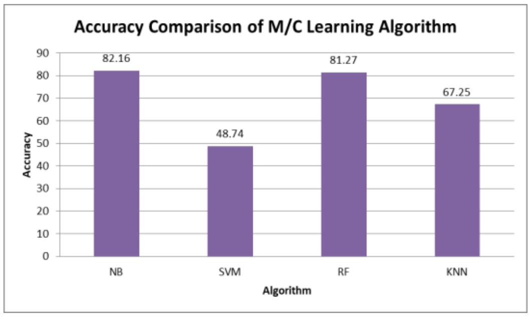
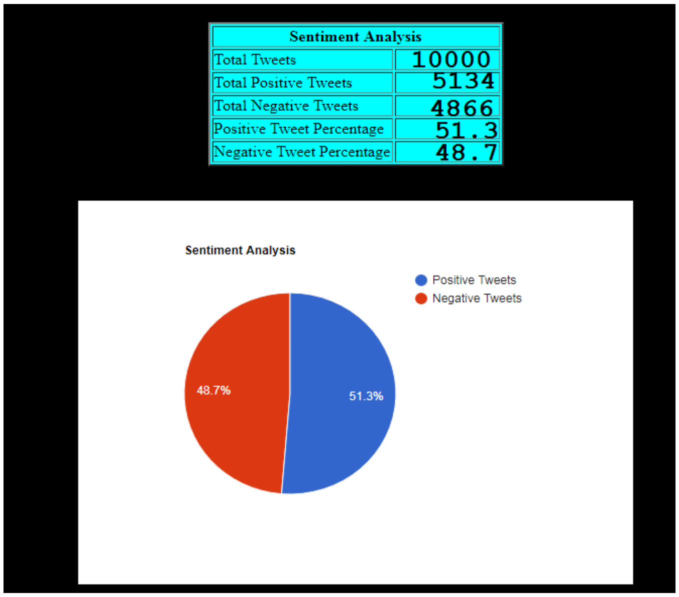


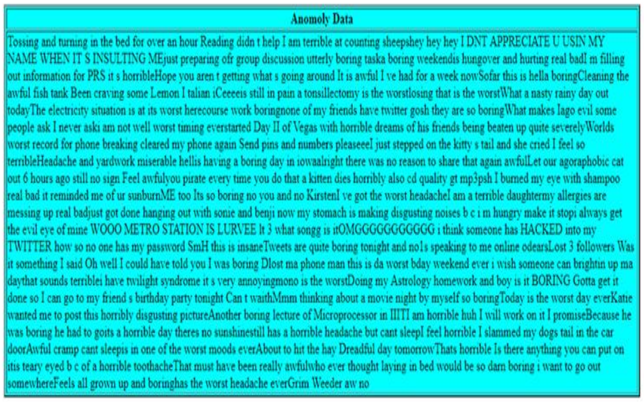
Figure 4. Accuracy Results



Sentiment Analysis	
Total Tweets	10000
Total Positive Tweets	5134
Total Negative Tweets	4866
Positive Tweet Percentage	51.3
Negative Tweet Percentage	48.7

Figure 5. Sentiment Analysis Results

Based on Sentiment analysis , please find anomalous tweets detected by system in figure 6.



Anomoly Data

Tossing and turning in the bed for over an hour Reading didn t help I am terrible at counting sheepshey hey hey I DNT APPRECIATE U USIN MY NAME WHEN IT S INSULTING MEjust preparing ofr group discussion utterly boring taska boring weekendis hungover and hurting real badI m filling out information for PRS it s horribleHope you aren t getting what s going around It is awful I ve had for a week nowSofar this is hella boringCleaning the awful fish tank Been craving some Lemon I talian iCeeeeis still in pain a tonsillectomy is the worstlosing that is the worstWhat a nasty rainy day out todayThe electricity situation is at its worst herecourse work boringnone of my friends have twitter gosh they are so boringWhat makes Iago evil some people ask I never aski am not well worst timing everstarted Day II of Vegas with horrible dreams of his friends being beaten up quite severelyWorlds worst record for phone breaking cleared my phone again Send pins and numbers pleaseeeI just stepped on the kitty s tail and she cried I feel so terribleHeadache and yardwork miserable hellis having a boring day in iowaalright there was no reason to share that again awfulLet our agoraphobic cat out 6 hours ago still no sign Feel awfulyou pirate every time you do that a kitten dies horribly also cd quality gt mp3psh I burned my eye with shampoo real bad it reminded me of ur sunburnME too Its so boring no you and no KirstenI ve got the worst headacheI am a terrible daughtermy allergies are messing up real badjust got done hanging out with sonie and benji now my stomach is making disgusting noises b c i m hungry make it stopi always get the evil eye of mine WOOO METRO STATION IS LURVEE lt 3 what songg is itOMGGGGGGGGGGG i think someone has HACKED into my TWITTER how so no one has my password SmH this is insaneTweets are quite boring tonight and no1s speaking to me online odearsLost 3 followers Was it something I said Oh well I could have told you I was boring Dlost ma phone man this is da worst bday weekend ever i wish someone can brightin up ma daythat sounds terriblei have twilight syndrome it s very annoyingmono is the worstDoing my Astrology homework and boy is it BORING Gotta get it done so I can go to my friend s birthday party tonight Can t waitMmm thinking about a movie night by myself so boringToday is the worst day everKatie wanted me to post this horribly disgusting pictureAnother boring lecture of Microprocessor in IIITI am horrible huh I will work on it I promiseBecause he was boring he had to goits a horrible day theres no sunshinestill has a horrible headache but cant sleepI feel horrible I slammed my dogs tail in the car doorAwful cramp cant sleepis in one of the worst moods everAbout to hit the hay Dreadful day tomorrowThats horrible Is there anything you can put on itis teary eyed b c of a horrible toothacheThat must have been really awfulwho ever thought laying in bed would be so darn boring i want to go out somewhereFeels all grown up and boringhas the worst headache everGrim Weeder aw no

Figure 6. Anomalous tweets

VI. CONCLUSION AND FUTURE SCOPE

Python provides a lot of libraries to help us for classification. With using this data identifying the sentiment we can detect anomalies.

The anomaly detection has a lot of use cases from both technical and non-technical point of view. I have studied an anomaly detection, which can also be used for a real-time emerging event for social media information. It can detect emerging anomalies that can be used for system research. Anomaly detection can be used for academic or scientific purposes.

- Future Scope

As I have mentioned this system can detect anomalies for offline datasets only. Even for few thousand (10K) it takes almost 8 seconds to process. This processing speed would not be enough for processing real time tweets feed which would be in lacs of tweets per second. I proposed to use Big Data architecture(Hadoop) to process real time data. Hadoop is a reliable, scalable and distributed computing framework. Hadoop uses Map- Reduce, Hive, and Pig techniques to extract and transform data. Then input data is imported into HDFS here Hadoop cluster can be used to transform large datasets in parallel.

Hadoop is proven to be very fast in processing very large volume of data.

Acknowledgment

I would like to thank Prof. Abhijit Patankar for giving his valuable guidance to me for preparing this paper. He acted as co-author and provided lots of review comments. It would have been impossible to complete this work without his guidance.

Image Reference

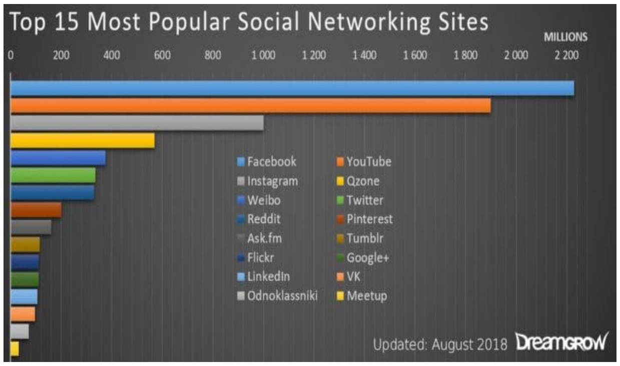


Image 1

References

[1] Weiren Yu, *Member, IEEE,* Jianxin LiMd Zakirul Alam Bhuiyan Richong Zhang and Jinpeng Huai ,"*Ring: Real-Time Emerging Anomaly Monitoring System over Text Streams*" , IEEE Big Data 2017. 2017.2672672

[2] Chao wang, Zhen Liu , Hui Gao and Yan Fu , "*Applying Anomaly Pattern Score for Outlier Detection*" , IEEE 2019. 2895094

[3] Varun Chandola, Arindam Banrjee and Vipin Kumar -- "*Anomaly Detection : A Survey*", in ACM Computing Surveys, September 2009

[4] Jagruti D. Parmar and Prof. Jalpa T. Patel ,"*Anomaly Detection in Data Mining: A Review* ", International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X, April 2017

[5] Salima Omar , Asri Ngadi and Hamid H. Jebur, "*Machine Learning Techniques for Anomaly Detection: An Overview* ",International Journal of Computer Applications (0975 – 8887) October 2013

[6] Pranali Ratnaparkhi, Rohini Jadhaw,Prakash Kshirsagar, "*Social Monitoring System for Dynamically Evolving Anomolies Over Text Stream*", vol.8, issue-4, April 2018 (*references*)

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "*Earthquake shakes twitter users: real-time event detection by social sensors*", in WWW, 2010.

[8] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic, "*Scalable distributed event detection for twitter*", in IEEE BigData, 2013.

[9] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang, "*Topicsketch: Realtime bursty topic detection from twitter*", in ICDM, 2013.

[10] E. Schubert, M. Weiler, and H.-P. Kriegel, "*Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds*", in KDD, 2014.

[11] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "*Tiara: a visual exploratory text analytic system*", in KDD, 2010.

[12] P. Lee, L. V. Lakshmanan, and E. E. Milios, "*Incremental cluster evolution tracking from highly dynamic network data*" , in IEEE International Conference on Data Engineering (ICDE), 2014, pp. 3–14.

[13] C. Li, A. Sun, and A. Datta, "*Twevent: segment-based event detection from tweets*" , in CIKM, 2012.

[14] D. Metzler, C. Cai, and E. Hovy, "*Structured event retrieval over microblog archives*", in HLT-NAACL, 2012.

[15] C. Budak, T. Georgiou, and D. A. A. El Abbadi, "*Geoscope: Online detection of geo-correlated information trends in social networks*", PVLDB, vol. 7, no. 4, 2013.

[16] J. Allan, R. Papka, and V. Lavrenko, "*On-line new event detection and tracking*", in SIGIR, 1998.

[17] T. Hofmann, "*Probabilistic latent semantic analysis*", in UAI, 1999.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "*Latent dirichlet allocation*", the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

Authors Profile

Mr. Shirishkumar Bari pursing Master of Engineering from Savitribai Phule Pune University, in 2019.

Mr. Abhijit Patankar
Asst Professor Department of Information Technology DYPCOE Akurdi Pune 411044