ANOMALOUS TOPIC DISCOVERY BASED ON TOPIC MODELING FROM DOCUMENT CLUSTER

Brejit Lilly Abraham[1], Anjana.P.Nair[2]

[1]M.Tech Computer Science & Engineering, Computer Science Department, Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India.
[2]Second Assistant Professor of Computer Science & Engineering, Computer Science Department, Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India.

---***---

Abstract - *Not at all like most anomaly detection (AD) techniques, which distinguish individual anomalies, the proposed strategy recognizes groups (clusters) of anomalies; i.e., sets of focuses which altogether display anomalous patterns. An algorithm for detecting patterns displayed by anomalous clusters in high-dimensional discrete data is proposed in this paper. In numerous applications, this can prompt a better understanding of the nature of the atypical behavior and to identifying the sources of the anomalies. Likewise, consider the circumstance where the typical examples display on just a little (salient) subset of the high dimensional component space. Singular AD methodology and methodologies that recognize anomalies utilizing every one of the elements normally neglects to distinguish such inconsistencies (anomalies), in any case, our methodology can distinguish such cases however, in general, find the mutual strange examples showed by them, and distinguish the subsets of salient components. In this paper, we focus on recognizing particular themes in a cluster of content records, building up our calculation in light of topic models. Aftereffects of Eventual outcomes of our investigations show that our methodology can precisely identify strange topic and remarkable components (words) under each such topic in a synthetic data set and accomplishes better execution contrasted with both standard group AD and individual AD procedures.*

Key Words: **Topic Modelling, Anomaly Detection, Pattern Detection, Document Cluster, Topic Discovery.**

1. INTRODUCTION

This document is template. Data mining is the task of finding fascinating examples from a lot of information or data. Data mining is the computational procedure of finding examples in substantial informational collections including techniques at the crossing point of artificial intelligence, machine learning, statistics, and database system. It is an interdisciplinary subfield of software engineering. The general objective of the data mining procedure is to concentrate data from an informational index and change it into a reasonable structure for further utilize. Text mining has turned into a prevalent range in data mining. Collection of document gathering records is available in the text database. Sources of these records incorporate email messages, computerized libraries, news articles, papers, books, and research papers. As the measure of data accessible in the electronic frame has expanded step by step message databases likewise develop quickly. Text mining, likewise known text data mining, this is comparable to text examination, which refers to the way toward separating incomparable nature of data from text.

The AD techniques typically detect individual sample anomalies. In this work, however, we concentrate on identifying abnormal patterns exhibited by anomalous groups (clusters) of tests. An anomalous cluster is a set of data samples which manifest similar patterns of a typicality [1]. Each of the example in such a group may not be highly atypical by itself, but, when considered collectively, the cluster demonstrates a distinct pattern which is significantly different from expected (normal) behavior [2]. In this paper, we propose a structure to identify such groups of anomalies and the atypical patterns they exhibit.

Topic modeling is the strategy for communicating and clarifying the content of a document to a leaner or to some publishing journals. This way of presenting ones idea or knowledge will be beneficial to the viewers and he/she will think, see effectively under what criteria or space each report are. It basically clarifies the center purposes of the topic. Generating the topic and cluster consequently will reduce time and effort to the viewer and include the main points of the paper like mining the primary focuses from the paper or pointing out the main words of a particular or specific document.

Distinguish the clusters of anomalous samples in the test batch and identify the salient feature subset for each such group. For example [1], the test group could consist of 1000 samples, each characterized on an element space. There could be two anomalous clusters in the test batch, with one cluster consisting of 50 samples, all of which show abnormal conduct concerning the (low) four - dimensional all of which show abnormal conduct concerning the (low) four - dimensional feature subspace. Another anomalous cluster could consist of 50 samples, each displaying anomalous behavior with respect to the six-dimensional feature subspace. Take note of that these two clusters, each display ordinary behavior on the (very large) remaining subset of the full feature space.

The issue of identifying clusters of data points which exhibit similar anomalous patterns is sometimes referred to as group anomaly detection [11],[13],[14]. We will synonymously refer to detecting clusters of anomalies and group anomaly detection. In the remainder of this paper, we categorize and analyze as follows. In section 2, we display related work of anomaly detection system and techniques. Section 3, diagram of an existing framework. Section 4, we talk about the proposed work. Section 5, we describe about the implementation of SMARTDOZ (name of our actualized software). Section 6. Finally, At last, we talk about the critical elements of our work and present our decisions.

2. RELATED WORKS

2.1 A survey of outlier detection methodologies

In this section, we review some previous works on group anomaly detection. The creator [2] have attempted to give a wide samples of current strategies and has presented an overview of contemporary systems for Outlier discovery, Outlier detection has been utilized for quite a long time to identify and, where proper, remove anomalous perceptions from data. Anomalies emerge because of mechanical shortcomings, changes in framework conduct, fake conduct, human blunder, instrument mistake or just through characteristic deviations in populaces. Their identification can recognize framework issues and extortion before they rise with possibly disastrous results. It can recognize error and evacuate their tainting impact on the informational collection and all things considered to refine the information for handling. They have additional classifications and break down a wide scope of anomaly detection methodologies. What's more, has called attention to how each handles anomalies and make proposals for when every philosophy is suitable for clustering, grouping and additionally acknowledgment.

2.2 Anomaly Detection: A Survey

Sample paragraph, Abnormality detection review [3] is an essential issue that has been investigated inside assorted research ranges and application domain. Numerous abnormality detection methods have been particularly created for certain application areas, while others are more nonspecific. Abnormality detection finds broad use in a wide assortment of uses, for example, misrepresentation location for charge cards, protection, or social insurance, interruption discovery for digital security, blame identification in wellbeing basic frameworks, and military observation for exercises. The significance of Abnormality detection is because of the way that inconsistencies in information mean huge, and regularly basic, noteworthy data in a wide assortment of use areas. This review tries to give an organized and thorough diagram of the exploration on peculiarity discovery. This review is an endeavor to give an organized and expansive outline of broad research on irregularity identification methods crossing numerous examination ranges and application spaces. They have likewise included two more classifications of Abnormality detection strategies, data theoretic and spectral procedures.

While a portion of the current reviews says the diverse uses of anomaly detection, we give a point by point discourse of the application areas where irregularity recognition procedures have been utilized. For every area, we talk about the idea of an anomaly, the different aspects of the anomaly detection issue, and the difficulties confronted by the abnormality detection methods. What's more, has recorded the strategies that have been connected in every application space. The current studies talk about abnormality detection procedures that identify the most straightforward type of anomalies. The paper has recognized basic abnormalities from complex oddities. The examination of utilizations of oddity location uncovers that for most application spaces, the fascinating irregularities are mind boggling in nature, while the majority of the algorithmic research has concentrated on simple anomalies. Uses of Abnormality detection in a few uses of abnormality detection. For every application domain they have examined the accompanying four perspectives [3]:

- The notion of anomaly;
- Nature of the data;
- Challenges associated with detecting anomalies;
- Existing anomaly detection techniques.

This survey [3] tries to provide a structured and comprehensive overview of the research on anomaly detection. They have grouped existing techniques into different categories based on the underlying approach adopted by each technique. For each category they have identified key assumptions, which are used by the techniques to differentiate between normal and anomalous behavior. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain. For each category, and has provide a basic anomaly detection technique, and then show how the different existing techniques in that category are variants of the basic technique. It has also provided a discussion on the computational complexity of the techniques since it is an important issue in real application domains.

2.3 Hierarchical Probabilistic Models for Group Anomaly Detection

[4]Propose a Mixture of Gaussian Mixture Models (MGMM) for gathering anomaly detection. Accept every information guide has a place toward one gathering and that all focuses in a gathering are demonstrated by the gathering's Gaussian blend show. Blending extents of the blend show for each gathering, notwithstanding, are not uninhibitedly assessed, but instead, progressively, are chosen from a restricted arrangement of T conceivable blending extent "sorts" (types). These sorts speak to the typical practices. A

test gathering is called abnormal on the off chance that it has a low probability of the typical sorts.

A drawback of the model above is that it uses a Dirichlet distribution to generate topics distributions. This Dirichlet is uni-modal peaking at a single topic distribution 2, and thus unable to generate multiple normal topic distributions. In other words, there is essentially only one normal topic distribution for the whole data set. This is often too restrictive for real data sets. To address this problem, a second model in which the topic distributions come from a pool of multinomial distributions. This allows multiple types of normal groups that have different topic distributions. Efficient learning algorithms are derived for both models based on variation EM techniques. They have also demonstrate the performance of the proposed methods on synthetic data sets, and shows that they are able to identify anomalies that cannot be found by other generative model based detectors. Empirical results are also shown for the SDSS astronomical data.

2.4 Group Anomaly Detection using Flexible Genre Models

This thought is then stretched out to Flexible Genre Models (FGM) in [5] by regarding blending extends as arbitrary factors which are adapted on conceivable ordinary classifications. One huge inadequacy of these strategies is that they expect that the gathering participation for each information point is known from the earlier. Since this data is not accessible as a rule, one should by and by performing hard bunching of the information into gatherings preceding applying FGM or MGMM. Such grouping, working in the full (high-dimensional) include space, might be profoundly incorrect when the odd example lies on a low-dimensional element subspace. Another real issue with these strategies is that they don't give any noteworthiness test to group anomalies —they just announce a hopeful group atypical in the event that it is among the top K% of bunches with most noteworthy inconsistency scores or if its oddity score is higher than a pre-set edge esteem. Take note of that the best possible decision of such limits is issue of threshold —an inadequately chosen threshold may lead either to a high false identification rate or weak detection control.

2.5 Latent Dirichlet allocation

Before Topic models such as Latent Dirichlet Allocation (LDA) [6] are widely used to model data having this kind of group structure. The original LDA model was proposed for text processing. It represents the distribution of points (words) in a group (document) as a mixture of K global topics each of which is a distribution. Let M ( ) be the multinomial distribution parameterized by θ € and Dir ( ) be the Dirichlet distribution with parameter $\propto$ €. LDA generates the *m*th group by first drawing its topic distribution m from the prior distribution Dir ( ). Then for each point *Xmn* in the *m*th group it draws one of the K topics and then generates the point according to that topic. Although topic models are very useful in estimating the topics and topic distributions in groups, the existing methods are incapable of detecting group anomalies comprehensively. In order to detect anomalies, the model should be flexible enough to enable complex normal behaviors. LDA, however, only uses a single Dirichlet distribution to generate topic distributions, and cannot effectively define what normal and abnormal distributions should be. It also uses the same K topics for every group, which makes groups in differentiable when looking at their topics. In addition, these shared topics are not adapted to each group either. In our proposed paper we use LDA for generating the topics of each document and with its strength.

2.6 GLAD: Group Anomaly Detection in Social Media Analysis

Ref. [7] addresses the first issue by presenting a method, specifically for network analysis, for jointly detecting groups of similar nodes and computing anomaly scores for the discovered groups. Nevertheless, unlike our method, [7] does not have an algorithmic procedure for discovering "hard" anomalous clusters one by one—some post-processing effort is required to hard-assign each data point to the cluster with highest membership degree. Moreover, [7] does not provide any statistical significance testing and relies on choosing an appropriate threshold for detecting anomalous clusters. And they mainly focused on a generative approach by proposing a hierarchical Bayes model: Group Latent Anomaly Detection (GLAD) model. GLAD takes both pair-wise and point-wise data as input automatically infers the groups and detects group anomalies simultaneously. To account for the dynamic properties of the social media data, further generalize GLAD to its dynamic extension d-GLAD.

2.7 One-Class Support Measure Machines for Group Anomaly Detection

Ref. [8] follows a discriminative approach to group anomaly detection and generalizes the idea of one-class support vector machines to a space of probability measures, proposing one-class support measure machines. A simple and efficient discriminative way of detecting group anomaly is illustrated in this work [8], M groups of data points are represented by a set of M probability distributions assumed to be i.i.d. realization of some unknown distribution. To handle aggregate behaviors of data points, groups are represented as probability distributions which account for higher-order information arising from those behaviors. The set of distributions are represented as mean functions in the RKHS via the kernel mean embedding. And also extend the relationship between the OCSVM and the KDE to the OCSMM in the context of variable kernel density estimation, bridging the gap between large-margin approach and kernel density estimation.

Groups in this method are represented as probability distributions which are mapped into a reproducing kernel Hilbert space using kernel methods. Similar to MGMM, this method requires hard-clustering of the data prior to detecting any anomalous group.

Ref. [9] proposes a rule-based anomalous pattern discovery algorithm calculation for detecting illness episodes. Bizarre examples in this strategy are portrayed by first or second request "rules". Each control is basically an arrangement of conceivable qualities that a subset of all out elements goes up against. Noteworthiness of each govern is measured by looking at event recurrence of each control in the test set with respect to the preparation set by directing Fisher's correct test and a randomization test. This thought is then reached out in [10], which utilizes Bayesian systems to quantify relative noteworthiness of each run the show. [11] Uses a comparative strategy, yet first identifies individual peculiar focuses and afterward looks for conceivable examples among them. These strategies do give factual testing systems to gauge essentialness of each group. They can likewise (for low dimensional issues) distinguish notable components for each bunch. In any case, not at all like our technique, they don't give an enhancement calculation to mutually identifying groups and their related low-dimensional anomalous patterns. This, specifically, makes these techniques less reasonable for high dimensional areas, (for example, text document).

Ref. [12] proposes Fast Generalized Subset Scan (FGSS) to detect anomalous patterns in downright informational indexes. Not at all like numerous different strategies, FGSS gives a calculation to building strange bunches by mutually seeking over subsets of information occurrences and subsets of peculiar characteristics. FGSS has preferable scaling attributes over [10] and [11] and, thus, can detect anomalous patterns which lie on higher dimensional feature spaces. However, FGSS requires computing a p-value for each feature of every sample based on a Bayesian network learned on the training set. Learning Bayesian systems may not be for all intents and purposes practical for high-dimensional issues, for example, content reports where there might be a huge number of components. Besides, FGSS can just distinguish a subset of atypical components for each group—not at all like our strategy, does FGSS not give a model to the normal example of irregularities displayed by the cluster.

A fairly related issue to anomalous topic detection in text documents is the issue of Topic Detection and Tracking (TDT) in the data recovery writing. The primary concentration of TDT is following subjects and distinguishing new occasions in a transiently requested stream of articles [13], [14]. TDT strategies, by and large, fall into the classification of grouping advancing information streams and single shot bunching, and broadly depend on the fleeting area of each archive and other related meta-information. Truth be told, even in disconnected (bunch) TDT, time is a focal piece of the examination [15]. Our technique, then again, considers a batch of documents (bag-of-word questions) and finds strange subjects exclusively in view of the substance of the records.

3. EXISTING SYSTEM

We propose an algorithm for detecting patterns exhibited by anomalous clusters in high-dimensional discrete information. Not at all like anomaly detection (AD) techniques, which detect individual anomalies, our proposed strategy detects groups (clusters) of anomalies; i.e. sets of points which collectively exhibit abnormal patterns. Additionally, we consider the situation where the atypical patterns exhibit on only a small (salient) subset of the very high dimensional feature space. Individual AD techniques and techniques that detect anomalies using all the features typically neglect to identify such anomalies.

3.1 Disadvantages of Existing System

- In existing AD methods can detects only individual anomalies.
- Prior works require separate procedures for clustering the data and for measuring the degree of anomaly.

4. PROPOSED SYSTEM

In proposed system, we focus on detecting anomalous topics in a batch of text documents, developing our algorithm based on topic models. Results of our experiments show that our method can accurately detect anomalous topics and salient features (words) under each such topic in a synthetic data set and two real-world text corpora and achieves better performance compared to both standard group AD and individual AD techniques.

Which is also used for document clustering and its specificity and similar document p-value is viewed and anomaly % of each clustered document will be viewed in the Gantt chart as output.

4.1 Advantages of Proposed System

In proposed method detects groups (clusters) of anomalies. Proposed algorithm to jointly learn and detect anomalous clusters and the (low dimensional) anomalous patterns that they exhibit.

4.2 System Architecture

As shown in the fig 1, the procedures of processing the topic modeling based on anomaly detection from a document cluster is described below:

- Input can be taken from real world as well as synthetic datasets.

- Preprocessing is done on these datasets which is Stop-word removal process.
- On this preprocessed data topic processing is done which involves topic modeling and strength calculation.
- Topic modeling involves topic generation and strength calculation which is done using LDA Algorithm. LDA Algorithm generates probability of the topic with the help of document clustering.
- After topic modeling we go for document clustering and calculate its bootstrap testing and then p-value of likely hood candidate cluster and bootstrap testing ratio will be displayed.
- Finally its anomaly % of each document cluster is viewed on the Gantt chart.

The proposed system is applicable to all the users and is designed with the creation of anomaly detection from document clustering and its similarity and anomaly % is analyzed and displayed to the users of this website.

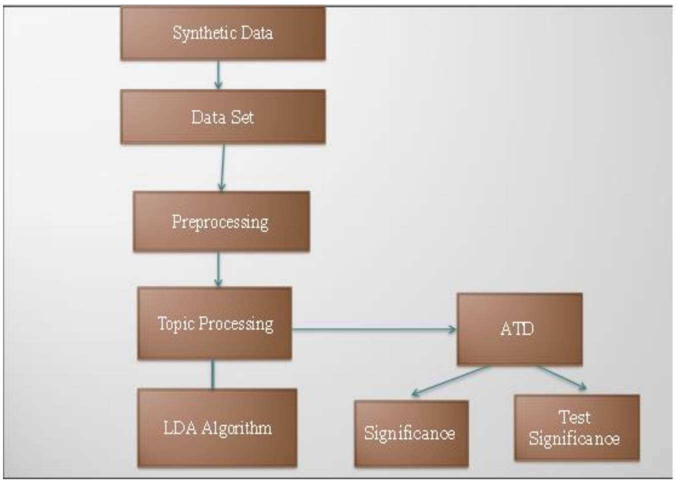

Fig -1: System Architecture of ATD

5. IMPLEMENTATION OF SMART DOCS

The Implementation is one of the most important tasks in a project. Implementation is the phase, in which one has to be cautious, because all the efforts undertaken during this project will be fruitful only if the software is properly implemented according to the plans made. Implementation is the stage in the project where the theoretical design is turned into a working system. The crucial stage is achieving successful new system and giving the users confidence in that the system will work effectively and efficiently.

It involves careful planning, investigation of the current system and its constraints on implementation and design of methods to achieve changeover. Apart from these, the major task of preparing for implementation is education and training of users and system testing.

The proposed system has mainly four phases namely:

- Data Preprocessing
- Topic Modeling
- Parsimonious topic model
- Anomalous topic discovery
- Determining the significance
- Significance test for a cluster.

5.1 Data Preprocessing

This module includes Topic discovery and its quality. Read syntactic data from the dataset, change over to data to the text documents. We first apply LDA model to the document. We utilize the same pre-and post-processing steps for learning topic, and afterward utilize the learned topic model and its behavioral quality. The data preparing task also perform following functions.

5.2 Stop Word Removal

The data may be processed further to remove stop words like auxiliary verbs, prepositions etc. The corpus data will be having only relevant terms mainly containing nouns and verbs.

5.3 Topic Modeling

The experiment automatically identifies the topics of every original document. This step is conducted for every time window, independently from each others. We first upload the needed document. We then remove from remaining document all stop words, slang words, 2 and non-English phrases. Next, we iteratively filter away words. After filtering each words of the document, these minimum thresholds are designed to ensure that for each word, we have enough observations to learn the latent topics accurately. A set of topics will be generated from each document, and each topics strength, i.e., number of times it has been used in the particular document. These data will be used to find the similarity and also for bootstrap testing.

5.4.1 LDA Algorithm

LDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion:

When writing each document:

- The Decide on the number of words N the document will have (say, according to a Poisson distribution).
- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). For example, assuming that we have the two

food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

- Generate each word in the document by:
- First picking a topic.
- Then using the topic to generate the word itself.

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the number of times it has been repeated in that particular document.

5.5 Anomalous Topic Discovery

We assume that we have a collection of normal documents which sufficiently characterizes all normal topics. We learn PTM as described before on this training corpus to discover the normal topics. Then, in the detection phase, our goal is to detect any and all patterns in the test corpus which are anomalous (unusual) with respect to the normal topics. In our proposed algorithm, we detect anomalous topics in the test set one by one. That is, at each step, we detect the cluster of test documents S (candidate anomalous cluster) that exhibits the pattern with maximum "deviance" from normal topics. Then, we conduct a statistical test to measure the significance of S and the topic exhibited by it, compared to the normal topics hypothesis. If the cluster candidate is determined to be significantly anomalous, we declare it as detected; we remove all documents in S from the test set, and then repeat this process until no statistically significant anomalous topic is found.

5.6 Determining the Significance

In this paper, we follow a more practical approach by proposing a bootstrap algorithm. First, we note that since the major difference between the null and alternative models is the new topic, our decision on whether to include the candidate document in the cluster or not can be reliably made based on the contribution of the new topic in modeling words in the candidate document. That is, if the new topic is not used in modeling a significant percentage of the words in the document, it is sufficient to rely on the null model to describe all contents of this document.

5.7 Significance of a cluster

These steps will be repeated until the entire document cluster will be processed and its specificity and bootstrap specification value will be displayed with its p-value of each document of a cluster. After growing of a cluster document, we need to determine whether the anomalous topic exhibited by the documents in that cluster is significant. Again, we note that due to small sample size, asymptotic distributions commonly known for the likelihood ratio test do not hold. Instead, we perform bootstrap testing to compare significance of a candidate cluster S to normal clusters.

For generating bootstrap document, we generate |S| bootstrap documents based on the null distribution from a collection of validation documents and compare the likelihood ratio score of this bootstrap cluster with that of the candidate cluster. Similar to the last section, for each document in the candidate cluster S, we generate a bootstrap document with similar topic proportions under the null model and with the same length. Then, we learn the alternative model and compute the log-likelihood ratio score, score (Sb). We repeat this process B2 times and compute the empirical p-value to measure significance of the candidate cluster.

After getting its similarity value, Anomaly Percentage = (Convert.ToDouble (rd ["similarity"].ToString ()) / total) * 100. We can calculate its anomaly of each cluster by dividing that by 100. So that anomaly % of each document cluster will be displayed and can view that in the Gantt chart.

6. CONCLUSIONS

In this Project work, an algorithm for detecting atypical topics exhibited by clusters of anomalous text documents. Not at all like individual-based AD systems, has our strategy identified clusters of anomalous documents which jointly manifest atypical topics on a small subset of (salient) features. Given a collection of ordinary documents, we first take in an (invalid) for the typical topics. At that point, in a different test set batch, we identify all cluster of abnormal documents and the topics showed by them, one by one. We utilize statistical tests to determine the significance of any detected cluster. Our trials demonstrate that our strategy can accurately detect anomalous topics and the subset of salient features under each such topic. Additionally, we demonstrate that since just a small subset of words are salient in any anomalous topic, some standard AD techniques, which evaluate atypicality on the full feature space. By contrast, our strategy accurately detects such anomalies by finding salient feature subsets and detecting clusters of anomalies.

REFERENCES

[1] Hossein Soleimani and David J. Miller, Senior Member, IEEE, "ATD: Anomalous Topic Discovery in High Dimensional Discrete Data" , IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 9, September 2016.

[2] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, 2004.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surveys, vol. 41, pp. 1–58, 2009.

[4] L. Xiong, S. P. Barnab_a, J. G. Schneider, A. Connolly, and V. Jake, "Hierarchical probabilistic models for group anomaly detection," in Proc. Int. Conf. Artif. Intell. Statist., 2011, pp. 789–797.

[5] L. Xiong, B. P_oczos, and J. Schneider, "Group anomaly detection using flexible genre models," in Proc. Adv. Neural Inform. Process. Syst., 2011, pp. 1071–1079.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. JMLR, 3:993–1022, 2003.

[7] R. Yu, X. He, and Y. Liu, "GLAD: Group anomaly detection in social media analysis," in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 372–381.

[8] K. Muandet and B. Scheolkopf, "One-class support measure machines for group anomaly detection," in Proc. 29th Conf. Uncertainty Artif. Intell., 2013, pp. 449–458.

[9] W. Wong, A. Moore, G. Cooper, and M. Wagner, "Rule-based anomaly pattern detection for detecting disease outbreaks," in Proc. 18th National Conf. Artif. Intell., 2002, pp. 217–223.

[10] W. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in Proc. Int. Conf. Mach. Learn., 2003, pp. 808–815.

[11] K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 169–176.

[12] E. McFowland, S. Speakman, and D. Neill, "Fast generalized subset scan for anomalous pattern detection," J. Mach. Learn. Res., vol. 14, no. 1, pp. 1533–1561, 2013.

[13] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 1998, pp. 37–45.

[14] X. Dai, Q. Chen, X. Wang, and J. Xu, "Online topic detection and tracking of financial news based on hierarchical clustering," in Proc. Int. Conf. Mach. Learn. Cybern., 2010, pp. 3341–3346 .

[15] Q. He, K. Chang, E.-P. Lim, and A. Banerjee, "Keep it simple with time: A reexamination of probabilistic topic detection models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 10, pp. 1795–1808, Oct. 2010.

[16] H. Soleimani and D. J. Miller, "Parsimonious topic models with salient word discovery," IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 824–837, Mar. 2015.

BIOGRAPHIES

Brejit Lilly Abraham received the Bachelor's Degree in Computer Science and Engineering from Karpagam University, Tamil nadu, India in 2017. She is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Kerala, India. His research area of interest includes the field of internet security, data mining and technologies in Department of Computer Science and Engineering.

Anjana.P.Nair received the bachelor's degree in LBS Institute of Technology for Women, Kerala, India. And master's degree in Computer Science and Engineering from Sree Buddha College of Engineering, Kerala, India in 2013. She is a lecturer in the Department of Computer Science and Engineering, Sree Buddha College of Engineering. Her main area of interest is Core Computers and has published more than 10 referred papers.