iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4437

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Feature Clustering Anomaly Detection from Document Pool Modeled with Topic Specifity and Utility Factors

Brejit Lilly Abraham[1], Anjana.P.Nair[2]
[1, 2]*M.Tech Computer Scince and Enginering, Assitant Professor Computer Science and Engineering.*

Abstract: ***The data of a record can be utilized to extract summaries documents or to compute summaries of the documents based on the words incorporated into them. Anomaly Detection is a method which is utilized to for recognizing patterns exhibited by anomalous groups (cluster). These patterns are extracted from document using LDA algorithm and used to discover specificity and anomaly. It is also used for users to analyze documents and help in similarities with them. The system can combine the different paper from different domain to same concepts. Later the document can be further classified using SVM algorithm for auto prediction of each document. Using CHUD algorithm the importance of each document will be analyzed by using each count of downloads of users using sequential mining. The result analysis of this project shows the accuracy of the result and also its time complexity comparison with two different algorithms.***
Keywords: ***Anomaly Detection, Cluster, Specificity, Anomaly, Utility Mining.***

I. INTRODUCTION

Data mining is the computational procedure of finding designs in huge informational collections including techniques at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is an interdisciplinary subfield of computer science. The general objective of the data mining process is to remove data from a data set and transform it into an understandable structure for additionally utilize. Aside from the raw analysis step, it includes database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

The Text Mining is the way toward extricating unstructured data or concentrates significant numeric records from the content and makes the information contained in the content open to the different data mining. This information can be utilized to separate outlines of the documents or to compute summaries of the reports in view of the words incorporated into them. Thus, users can analyze words, clusters of words in reports or could analyze documents and help in similarities between them or how it is identified to other variables in the document. Text mining will ”turn text into numbers” (meaningful indices), which would then be able to be utilized as a part of different, such as predictive data mining studies, the application of unsupervised learning methods (clustering), etc. Neither the data accumulation and data readiness, nor result translation and announcing is a bit of the data mining step, be that as it may, do have a place with the general KDD prepare as extra strides.

Text mining is a variation on a field called data mining that tries to discover interesting patterns from expansive databases. Text mining is otherwise called Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT). It refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is normally determined through the formulating of examples and patterns through means such as statistical pattern learning.

Anomaly detection (AD) methods ordinarily identify individual sample anomalies. In this work, however, it focuses on detecting abnormal patterns exhibited by anomalous groups (clusters) of tests and its related work. An anomalous cluster is an arrangement of data samples which show comparative examples of a regularity. Each of the samples in such a cluster 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. In this paper, it focuses on the framework to detect such groups of anomalies and the atypical patterns exhibit and its previous work.

A. Anomaly Detection

The Anomaly detection (AD) is the problem of identifying items or patterns which do not conform to normal or expected behavior. Anomaly detection techniques have been widely used e.g., to detect credit card fraud, insurance fraud, and network intrusions. AD techniques typically detect individual sample anomalies. In this work, however, it focuses on detecting abnormal patterns exhibited by anomalous groups (clusters) of samples. An anomalous cluster is a set of data samples which manifest similar patterns of a typicality. Each of the samples in such a cluster 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. A framework is proposed to detect such groups of anomalies and the atypical patterns it exhibit. Moreover, consider the case where the anomalous pattern may manifest on only a small subset of the features, not on the entire feature space; i.e., samples in the anomalous cluster may be far apart from each other measured on the full feature space, but on a subset of the feature space (the salient features), it exhibit a similar pattern of abnormality. In addition to detecting atypical clusters, the proposed method identifies each cluster's salient feature subset.

In some cases, no prior knowledge about normal behavior is available, and the goal is to detect anomalies (outliers) in a single data set consisting of normal and possibly abnormal instances, without any annotation of which samples are normal. More typically, there is a collection of normal data which sufficiently characterizes normal behavior. In the training phase, and use this data to build a (null) model. Then, in the detection phase, this model is used as a reference to help detect (possible) clusters of anomalous patterns in a different (test batch) data set.

Topic modeling is the method of expressing and explaining 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, understand easily under what criteria or domain each document are. It mainly explains the core points of the topic. Generating the topic and cluster automatically will reduce time and effort to the viewer and include the main points of the paper like mining the main points from the paper.

B. Problem statement

It formally defines our problem as follows:

Given: A training set consist of ordinary data to be used in learning a null model.

Furthermore, the abnormal sample subset may consist of clusters of samples, with each group particularly described by the way that its samples exhibit anomalous behavior (in respect to the null model) on the same low-dimensional subset of the full (high-dimensional) feature space.

C. Objective

Detect the clusters of anomalous samples in the test batch and identify the salient feature subset for each such cluster. The focus here is on group anomaly detection in very high-dimensional data domains, where the samples in a group are expected to manifest its anomalies on a (the same) low-dimensional (a priori unknown) subset of the high-dimensional feature space. This approach requires jointly detecting these clusters of samples and it's (in general, low-dimensional) salient feature subsets.

In this paper, it focuses 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. It can also find its specificity using the topic modeling autonomous strength. 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.

To make this work for efficient for user usage, the implementation of utility mining for the feedback mechanism and frequency count, i.e., how many times each document is viewed by different users, with the help of CHUD (Closed high utility item set discovery) algorithm. Then with the help of SVM (Support Vector Machine Classification) algorithm helps us to predict the document with its labels. Later with this, result evaluation analysis can be done and accuracy also will be calculated.

D. Overview of the project

Anomalous cluster detection approach consists of several fundamental steps repeatedly applied to the test batch:

1) Determining the best current candidate anomalous cluster;
2) Determining whether this candidate cluster is anomalous.
3) Check its anomalous percentage.

4) Calculate anomaly and specificity.

In this paper, the statistical tests to accomplish these steps; i.e., to determine which samples significantly belong to the best current cluster candidate and to test whether the candidate exhibits a statistically significant degree of relative to the null model. The proposed framework can be applied generally, to both continuous and discrete valued data. However, in this paper, it focus on detecting atypical patterns (topics) in text documents, a domain with a very high-dimensional (bag of-words) feature space. Anomalous topic discovery (ATD) for document databases represents a challenging domain due to the high feature dimensionality, with many candidate low-dimensional subspaces that may exhibit anomalous patterns.

The proposed framework is based on focusing topic models. Topic models have been used in modeling different types of data such as images and text documents. The formulation of topic models primarily developed for modeling text documents, based on a multinomial distribution model for each topic.

Topic models are a class of statistical models often used for discovering latent patterns (topics) in a collection of text documents. Each topic specifies a pattern of words; i.e., words that appear more or less frequently than others under that topic. A simple and yet widely popular topic model is Latent Dirichlet Allocation (LDA), which posits document- specific mixing proportions over the topics, with each topic a multinomial distribution over the given vocabulary. LDA in its basic structure is a parameter-rich model, which, when applied to high-dimensional problems such as text documents. Second, in the detection phase, under the alternative hypothesis, it posit that a cluster of documents in the test set may contain an additional topic

Accordingly, build an alternative model M_1 with M + 1 topic by simply adding one topic to the null model. Then, in the spirit of a generalized likelihood ratio test, next seek the best candidate anomalous cluster by, alternately, learning the parameters of the new topic and choosing the documents from the test set for which the new topic has significant presence, until a convergence criterion is met. Then it can also check the specificity with the help length of patterns by 0.5.Finally, it measures the statistical significance of the candidate cluster. If the cluster is significant, it detect it, remove all its documents from the test set, and repeat this detection process until no further significant topics are discovered. That is, detect anomalous clusters in the test set one by one. Note that the new topic represents the anomalous pattern in each cluster and the set of topic-specific words under that topic are the salient features of that pattern. Then apply non-parametric bootstrap testing both to determine;

a) Whether a document belongs to a candidate anomalous cluster and

b) If a candidate cluster is significantly anomalous.

For the first task, it compares the empirical topic proportion of the new topic in the candidate document with that of a set of normal bootstrap documents. Similarly, for measuring significance of a cluster, it computes the ratio of the candidate cluster's likelihood under the alternative and null models, comparing it with normal bootstrap clusters, and computes an empirical p-value. It calls a candidate cluster anomalous only if the empirical p-value is lower than a pre-set significance level.

After analyzing its consistency by anomaly and specificity it can then evaluate it with the help of SVM Algorithm for the predicting them with them with three different test stages, at first it is used for test i.e., shows its features. Second it is used for training with some SVM parameters. Later it is used for train model and they create its value, finally it predicts the document with the help of labels as 1 and 0.

With the help of CHUD Algorithm the importance of each document can be analyzed. Later the result accuracy will be also calculated and consistency of that document will be analyzed.

II. SYSTEM ANALYSIS

The development of this project is done in Microsoft Visual Studio.Net, which is an integrated development environment (IDE) from Microsoft. C# programs run on the .NET Framework, an integral component of Windows that includes a virtual execution system called the common language runtime (CLR) and a unified set of class libraries. ASP.NET is a web development platform, which provides a programming model, a comprehensive software infrastructure and various services required to build up robust web applications for PC as well as mobile devices. The project is proposed to be developed in ASP.Net as front end and SQL Server 2008 R2 as back end which develop to help powerful software. With the help of these tools and techniques, this project can be implemented.

A. Existing system

Used to detect individual anomalies for the given documents. Document clustering and its topic were analyzed by PTM. It also has some limitations for giving the significance of anomaly, thus it use LDA for topics autonomous strength and it gives us a better result.

1) Disadvantages of an existing system

a) In existing AD methods can detects only individual anomalies,

b) Prior works require separate procedures for clustering the data and for measuring the degree of anomaly.

B. Proposed system

In this paper, it is focused on detecting anomalous topics in a batch of text documents, developing our algorithm based on topic models. Results for this 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. By calculating its anomaly and specificity. 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.

Each document download count is also calculated for utility mining by sequential patterns and using SVM the prediction of each document is also displayed. Later result analysis and evaluation is analyzed and by variation of time complexity taken for both two different algorithms is also displayed. This project is also designed as for user-friendly for all the readers. The architecture steps done for proposed system are:

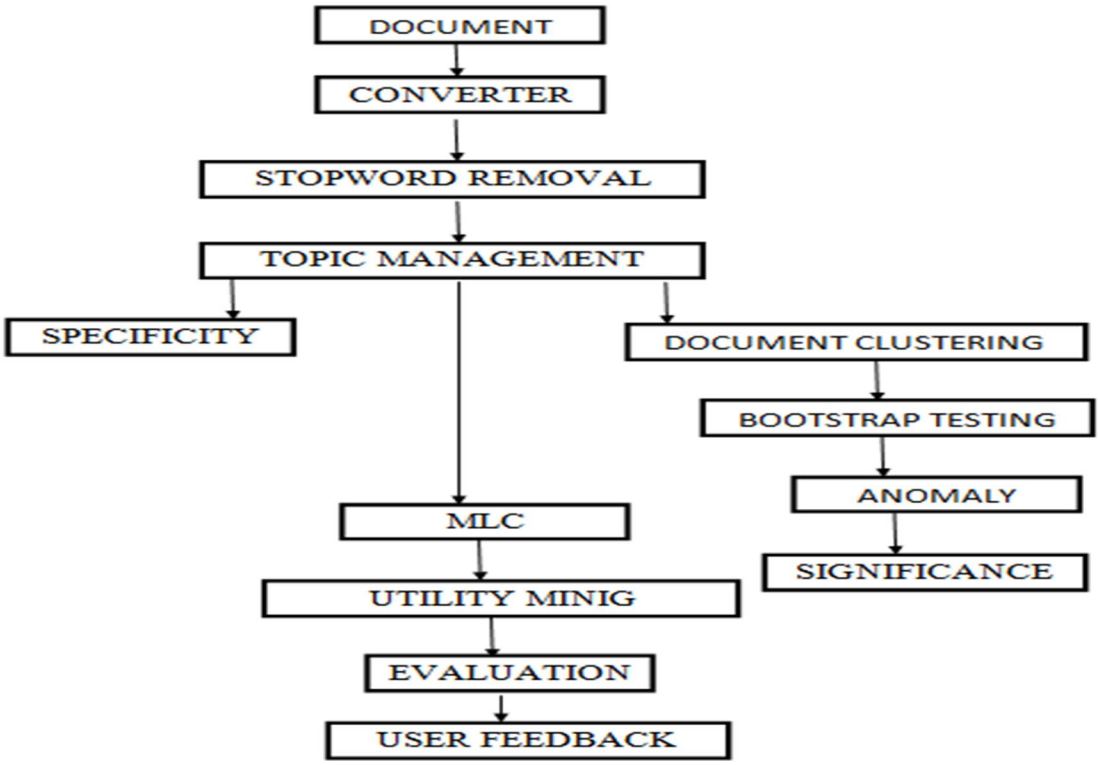


Fig 1. Proposed System Architecture

1) Input can be taken from real world (pdf, word, notepad) as well as synthetic datasets

2) The documents uploaded will be converted using the converter for the further processing steps into a text document.

3) Preprocessing is done on these datasets which is Stop-word removal process.

4) On this preprocessed data topic processing is done which involves topic modeling and strength calculation. Later with this topic specificity can be calculated.

5) 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.

6) After topic modeling it 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.

7) Finally its anomaly % of each document cluster is viewed on the Gantt chart.

8) Thus anomaly and specificity will be calculated.

9) To make this project more efficiently for the users, the user can search its needed domain topic and the relevant topic as it have will be displayed below and it can view and download as per its need.
10) Utility mining is enhanced for the feedback mechanisms and for the frequency strength is each person as it has viewed, the count can be analyzed with the help of CHUD algorithm.
11) The frequency count of how much times that particular topic is been downloaded.
12) A reminder can be also given for user as the feedback mechanism, for its rating.
13) The advantage of this utility factor, it decides the importance of topic in a cluster.
14) SVM (Support Vector Machine classification) is used to predict each document.
15) Finally result Evaluation is analyzed and accuracy is verified.
16) Comparison of anomaly and specificity is also displayed.
17) Comparison of two different algorithms and its time complexity variation is also show in bar chart.

a) Advantages of proposed system

i. In proposed method detects groups (clusters) of anomalies
ii. Proposed algorithm to jointly learn and detect anomalous clusters and the (low dimensional) anomalous patterns that it exhibit.
iii. Can be applicable to all the users.
iv. The advantage of this utility factor, it decides the importance of topic in a cluster.

C. Modules

The proposed system mainly consists of three modules:

1) author Module
2) Admin Module
3) User Module

The main functionalities of the Author module are:

Function	Description
Registration	A new author can register to the SMARTDOZ site with his personal details. Registered user can log in with his ID and password.
Upload document	Author can upload his document and also with its details. And these entire uploaded documents will be stored in dataset and other details will be stored in database.
Check anomaly	As the author has uploaded the number of files, he/she can select the needed file and view its anomaly % and similarity will be displayed in the Gantt chart.
Enhancement ATD	i. At first it predicts each document after analyzing the anomaly and specificity. ii. Utility mining is used in this project to know the importance of each document. iii. Result analysis and experiment evaluation is done and its result is displayed in bar charts.
Document Search	If the author wants to know any details about a particular file he/she can search that particular topic by giving its title name.

TABLE 1: Author module function and description

The main functionalities of the admin module are:

Function	Description
Admin Homepage	Functionalities added and performed in that will be displayed.
Data Set Management	Admin can view the entire document which is uploaded and stored in the dataset.
Document Management	Admin can view the author details and delete the unnecessary files.
Stopword Management	If the topic modeling file document has word which has been used for many times and if that word is need not to be added in topic modeling process, then it have to save that word in stopword management.
Topic Modeling	i. Performs LDA algorithm and shows Topic, strength evaluation. ii. After the autonomous strength calculation it is used for calculating specificity, by patterns length.
Document Clustering	Specify the number of document need to be clustered. And perform prepare documents and cluster.
Bootstrap testing	Load the content and shows its probability, perform ATD and its anomaly, and then check its anomaly. The admin has to perform all these functions for each document stored in the dataset, so that author can view the anomaly of each document.
Check anomaly percentage	The cluster documents and its anomaly percentage can be calculated and will be represented with the help of a Gantt chart.
Enhancement ATD	i. At first it predicts each document after analyzing the anomaly and specificity. ii. Utility mining is used in this project to know the importance of each document. iii. Result analysis and experiment evaluation is done and its result is displayed in bar charts.

TABLE 2: Admin module function and description

The main functionalities of the User module are

Function	Description
Registration	A new user can register to the SMARTDOZ site with his/her personal details. Registered user can log in with that ID and Password.
Search	Search for a file with necessary topic, and document with above pattern will be displayed, with that other document cluster and its similarity will be viewed.
Utility Mining	i. Utility mining is used in this project to know the importance of each document. ii. This is calculated using the count of each document downloaded by users.

TABLE 3: User module function and description

III. SYSTEM DESIGN

A. Implementation

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.

B. Data Preprocessing

The module includes Topic discovery and its strength. At first the document should be converted to into a .txt for that it use a converter. Read syntactic data from dataset, convert to data to the text files. It first apply LDA model to the document. The same pre- and post- processing steps are used for learning topics, and then use the learnt topic model and its behavioral strength. The data processing task also perform following functions.

C. Stop Word Removal Algorithm

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. A dictionary based approach is been utilized to remove stopwords from document. A generic stopword list containing 75 stopwords created using hybrid approach is used.

D. Topic Modeling

The experiment automatically identifies the topics of every original document. This step is conducted for every time window, independently from each others. It first uploads the needed document. And then remove from remaining document all stop words, slang words, 2 and non-English phrases. Next, it iteratively filters away words. After filtering each words of the document, these minimum thresholds are designed to ensure that for each word, it 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.

E. 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;

1) Decide on the number of words N the document will have (say, according to a Poisson distribution).
2) Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). For example, assuming that it 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.
3) Generate each word in the document by:
4) First picking a topic.
5) Then using the topic to generate the word itself.
6) 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.

F. Specificity Calculation

The Specificity refers to the influence of a set of topic in every document in a cluster. Specificity also decides the importance of each document by calculating its patterns length.

$$\text{Specificity} = \frac{length\ of\ the\ patterns}{0.5}$$

Patterns length will be calculated by taking limited number of topic will the order of its strength calculated by topic modeling.

G. Document Clustering

To calculate Document cluster at first it have to read all doc &display that in topics and after that document vector library file is used checks similarity between two documents and creates an array and compare that together. Later all documents will be calculated similarly using similarity metrics and vector will be created with the help of vector space model library files.

In this proposed algorithm, it detects anomalous topics in the test set one by one. That is, at each step, it detects the cluster of test documents S (candidate anomalous cluster) that exhibits the pattern with maximum “deviance” from normal topics. Then, it conducts 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, and declare it as detected; it removes all documents in S from the test set, and then repeats this process until no statistically significant anomalous topic is found

H. Determining The Significance

In this paper, it follows a more practical approach by proposing a bootstrap algorithm. It 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.

I. Significance Test For A Cluster

The After the calculation of bootstrap testing and its empirical value of likelihood it is then used for anomaly. 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, it needs to determine whether the anomalous topic exhibited by the documents in that cluster is significant. Again, note that due to small sample size, asymptotic distributions commonly known for the likelihood ratio test do not hold. Instead, it performs bootstrap testing to compare significance of a candidate cluster S to normal clusters.

For generating bootstrap document, it 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, it generates a bootstrap document with similar topic proportions under the null model and with the same length. Then, it learns the alternative model and compute the log-likelihood ratio score, score (Sb). It repeats this process B2 times and compute the empirical p-value to measure significance of the candidate cluster. After getting its similarity value, then anomaly percentage can be calculated and divide by100

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.

Thus the consistency of a document can be determined by this calculation of specificity and anomaly, which is to know whether it is relevant or not-relevant.

J. Support Vector Model Classification

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

This SVM is defined by three levels in this project.

1) Test model
a) Show the features(anomaly and specificity)
b) Do scaling for all values.
2) Train model
a) It is done with SVM parameters.
3) Perform train using SVM
a) They create and write.
4) Predict model
a) Predict label using SVM as 1 or 0.
b) 1- relevant

0- non-relevant

The process of this SVM model classification steps is as follows:

a) Open a text file
b) Read all document Id and its anomaly value, specificity.

c) Write it into a text files
d) Define a class label, as example 1 or 0.
e) If 0.5 specificity> threshold and
0.5 anomaly > threshold
f) Label it as = 1
Else
Label =0
g) Write label to train set.
h) Close file.

K. Utility Mining

Mining high utility itemsets (HUIs) from databases is an important data mining task, which refers to the discovery of itemsets with high utilities. To achieve high efficiency for the mining task and to provide a concise mining result to users it goes for CHUD (Closed High Utility Itemset Discovery). The utility of an itemset represents its importance, which can be measured in terms of weight, profit, cost, quantity or other information depending on the user preference. An itemset is called a high utility itemset.The input for utility mining is taken as fig. 2 as shown below.

Freq	doctitle	docdomain	docId
3	barley stock	Trading	103
1	black hole	Networking	104
4	anomaly	Data Mining	105
2	ethernet	Networking	204

Fig 2. Input counts of the users

Here it uses sequential mining for calculating the count of downloads by each user. Used for analyze the utility by the number of downloads and know about utility member with their document Id as per their registration.
Utility = frequency * consumption
With this it will be able to find the importance of each document needed for users. The analyzes of our calculation result will be shown in a grid as per the document number. Thus an absolute utility of an itemset is calculated.

docid	utility
204	2
103	3
105	4

Fig 3. Utility of document as per document Id

IV. RESULT AND DISCUSSION

The purpose of system testing is to identify and correct errors in the candidate system. Testing is an important element of the software quality assurance and represents the ultimate review of specification, design and coding. The increasing visibility of the software as a system element and the costs associated with a software failure are motivated forces for well planned testing.
In this section, it compares performance of our algorithm against methods on a synthetic data set and real-world text document. It uses class labels of the data sets to define anomalous classes. In each data set, it chooses some classes as anomalous and takes all

documents from those classes out of the training and validation sets. It then randomly selects some documents from normal classes and some documents from anomalous classes to create the test set. Our goal is to detect clusters of documents from the anomalous classes in the test set. It uses the class labels for creating the training and test sets and for evaluating test set performance.

A. Specificity Comparison

The below grid result calculation fig 4 shows the specificity variation of value occurred for each document obtained from specificity patterns according to their topic strength. And the bar chart displayed will help us analyze the variation difference of each document

docid	specificity
07587426.pdf.txt	13.92839
aba.pdf.txt	13.92839
adhoc.pdf.txt	13.82027
ethrnet.pdf.txt	13.78405
IJREAMV02I082007.pdf.txt	13.96424

Fig 4. Specificity calculations for each document

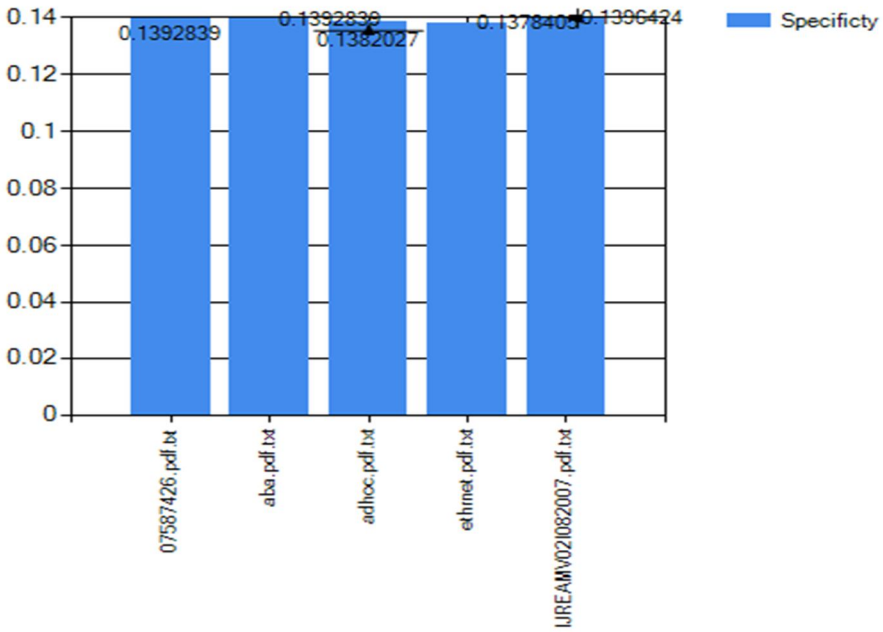


Fig 5. Bar chart diagram of specificity in each document

clusterid	docid	anomaly	specificity
2	aba.pdf.txt	0.333	13.92839
2	adhoc.pdf.txt	1.000	13.82027
2	ethrnet.pdf.txt	0.333	13.78405
2	IJREAMV02I082007.pdf.txt	0.333	13.96424

Fig 6. Show the difference value calculated in each document

The above fig 6 shows our result calculated from our new algorithm i.e., specificity calculation and the next row is anomaly from our calculated p-value.

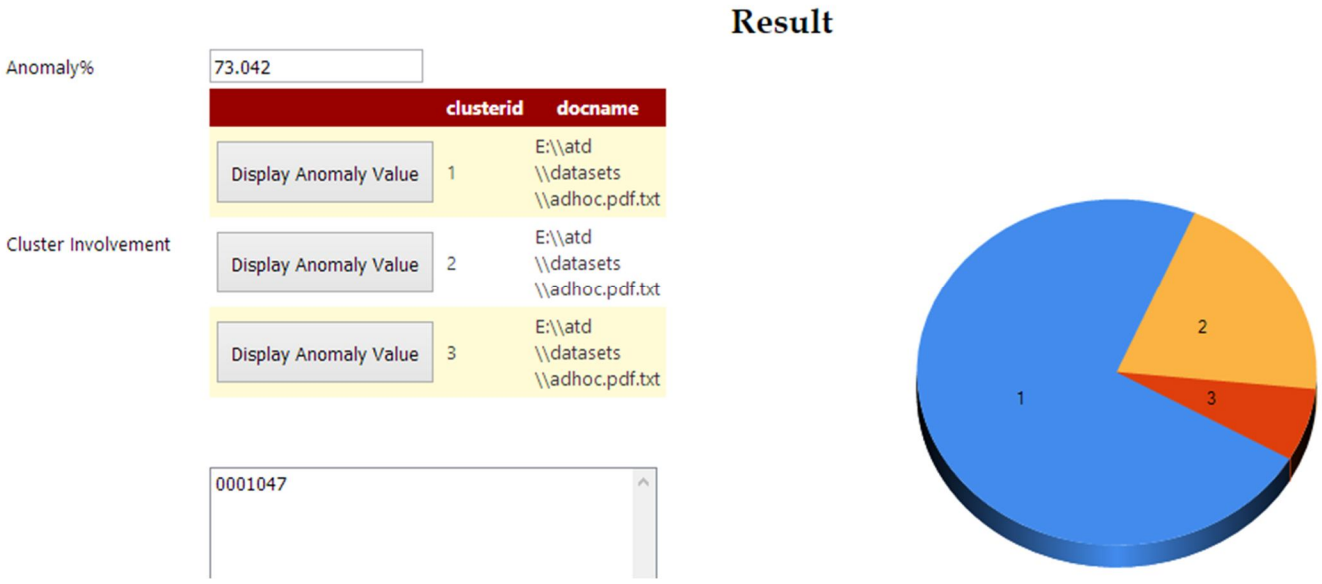


Fig 7. Show the significance of anomaly percentage

This diagram fig 7 is used to represent the significance of anomaly percentage as per calculated in anomaly, and it also shows the document cluster in each document as per their document Id and the brief description is displayed by the help of a Gantt chart.

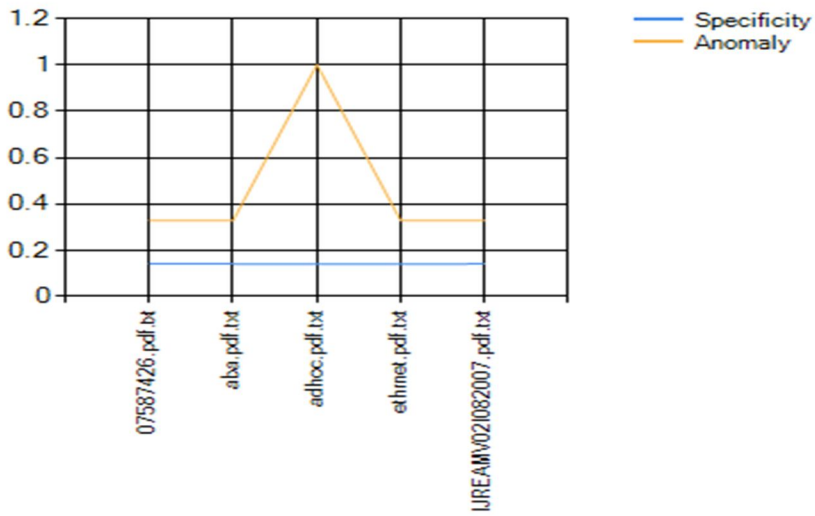


Figure 4.5 shows the bar chart diagram of comparison between specificity and anomaly

B. Accuracy View of Machine Learning

For calculating accuracy it takes values of accuracy and recall. The recall can be defined as the fraction of relevant instances that having retrieved over the total amount of instances. The measure of relevance is measured by precision and recall. The accuracy calculation view of machine learning can be done as follows, and fp- flase positive, fn- false negative, tp-true positive, tn- true negative:

Precision = fp / (tp + fp)

Recall = fp / (tp + fn)

Sense = fp / (tp + fn)

Specifcity = fn / (fp + tn)

Accuracy = (sense * (tp + fn) / (tp + fn + fp + tn)) + (specif * (fp + tn) / (tp + fn + fp + tn))
Error = (fp + fn) / (tp + fn + fp + tn)

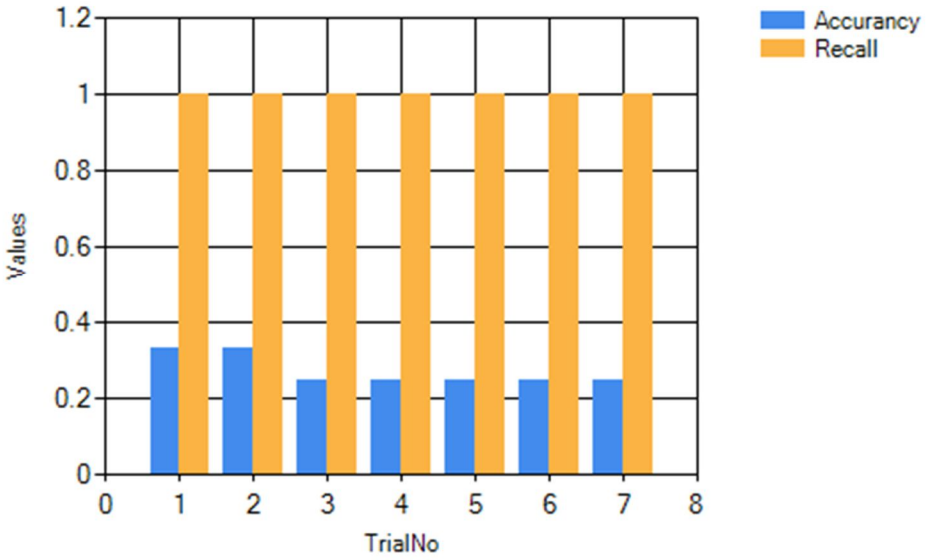


Fig 8. Diagram represents the accuracy view of machine learning

C. Time Complexity

The testing is performed in the evaluation phase, in this result analysis time Complexity is used to define the efficiency of our project. It calculates the time efficiency of the project. It is done by calculating the time in millisecond for different our two different algorithms.i.e., specificity and anomaly. The above figure shows us the variation of time taken for both two algorithms in the bar chart. Thus then we can say that the project is efficient and scalable.

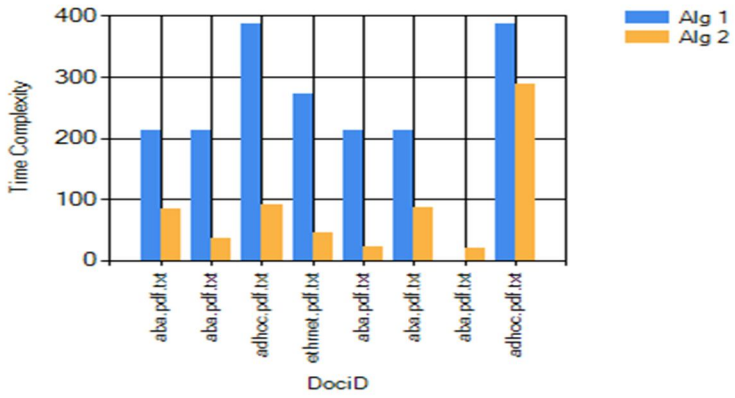


Fig 9. Show the variation of time taken for two different algorithms

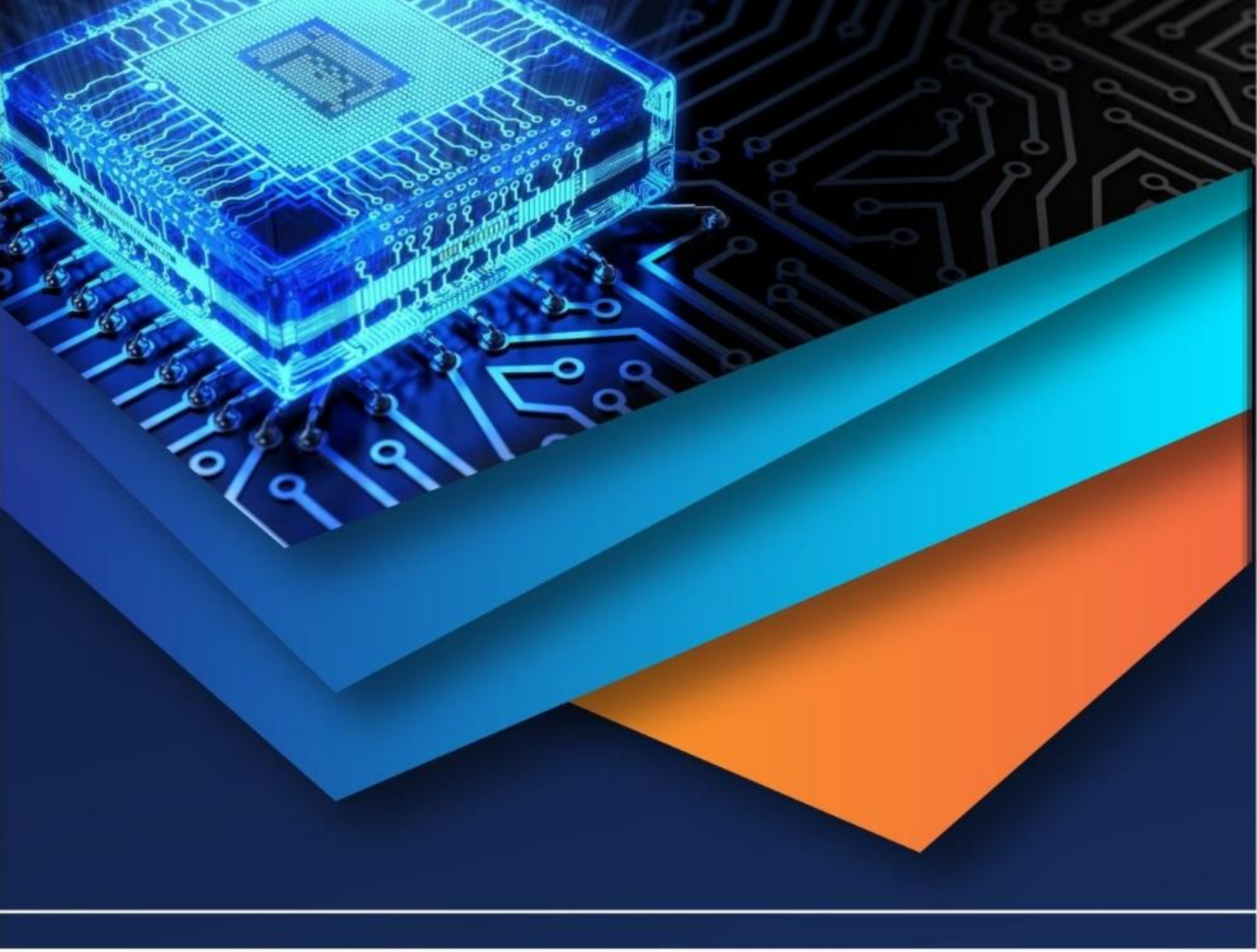
V. CONCLUSION

The proposed algorithm is used for detecting atypical topics exhibited by clusters of anomalous text documents. Given a collection of normal documents, it first learn a (null) model for the typical topics using topic modeling analysis and using that strength it can calculate the specificity of each document by the specificity of patterns. Then, in a separate test set batch, it detects all clusters of abnormal documents and the topics exhibited by them, one by one. It helps us to define each document to their particular cluster. Thus with help of bootstrap testing anomaly p-value is obtained and later anomaly significance % is displayed. Again to obtain more efficiency it then go for SVM for prediction. And utility is also used for knowing the importance of each document from users. This experiments show that our method can accurately detect anomalous topics and their sub-topics efficiently. The result analysis is

compared in different ways to evaluate the project efficiency. By contrast, the method accurately detects such anomalies by discovering their consistency of relevant and non-relevant. For future enhancement Utility mining, for related data document can be added in the future enhancement it can go for Real data (such as pdf, doc, and notepad) can be used for output with bigger size and the speed efficiency of calculating consistency can be increased.

REFERENCES

[1] Hossein Soleimani and David J. Miller, (September 2016) Senior Member, ATD: Anomalous Topic Discovery in High Dimensional Discrete Data, IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 9.
[2] V. J. Hodge and J. Austin, (2004) A survey of outlier detection methodologies, Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126.
[3] V. Chandola, A. Banerjee, and V. Kumar,(2009) Anomaly detection: A survey, ACM Comput. Surveys, vol. 41, pp. 1–58.
[4] A. Srivastava and A. Kundu, (Jan. /Mar. 2008) Credit card fraud detection using hidden Markov model, IEEE Trans.DSC,vol. 5.
[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, (2003) Latent dirichlet allocation, J. Mach. Learning Res., vol. 3, pp. 993–1022.
[6] D. Blei, L. Carin, and D. Dunson, (Nov. 2012) Probabilistic topic models, ACM Commun., vol. 55, pp. 77–84, Nov. 2012.
[7] H. Soleimani and D. J. Miller, (Mar. 2015) Parsimonious topic models with salient word discovery, IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 824–83.
[8] B. Efron, (1979) Bootstrap methods: Another look at the jackknife," Ann. Statist., vol. 7, pp. 1–26.
[9] K. Wang and S. Stolfo, (2004) Anomalous payload-based network intrusion detection, in Proc. 7th Int. Symp. Recent Adv. Intrusion Detection, pp. 203–222.
[10] L. M. Manevitz and M. Yousef, (2001) One-Class SVMs for document classification, J. Mach. Learning Res., vol. 2, pp. 139–154.
[11] L. Xiong, S. P. Barnab_a, J. G. Schneider, A. Connolly, and V. Jake, (2011) Hierarchical probabilistic models for group anomaly detection, in Proc. Int. Conf. Artif. Intell. Statist., pp. 789–797.
[12] L. Xiong, B. P_oczos, and J. Schneider, (2011) Group anomaly detection using flexible genre models, in Proc. Adv. Neural Inform. Process.Syst., pp. 1071–1079.
[13] R. Yu, X. He, and Y. Liu, (2014) GLAD: Group anomaly detection in social media analysis, in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 372–381.
[14] K. Muandet and B. Scholkopf, (2013) One-class support measure machines for group anomaly detection, in Proc. 29th Conf. Uncertainty Artif. Intell., pp. 449–458.
[15] J. Major and D. Riedinger, (2002) EFD: A hybrid knowledge/ statistical based system for the detection of fraud, J. Risk Insurance, vol. 69, no. 3, pp. 309–324..



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY