# Data Leak Identification in Social Networks using K Means Clustering & Tabu K Means Clustering

**Jayavarapu Karthik, V. Tamizhazhagan, S.Narayana**

***Abstract***: ***The prevention of leakage of data has been defined as a process or solution which identifies data that is confidential, tracks the data in a way in which it moves in and out of its enterprise to prevent any unauthorized data disclosure in an intentional or an unintentional manner. As data that is confidential is able to reside on various computing devices and move through several network access points or different types of social networks such as emails. Leakage of emails has been defined as if the email either deliberately or accidentally goes to an addressee to whom it should not be addressed. Data Leak Prevention (DLP) is the technique or product that tries mitigating threats to data leaks. In this work, the technique of clustering will be combined with the frequency of the term or the inverse document frequency in order to identify the right centroids for analysing the various emails that are communicated among members of an organization. Every member will fit in to various topic clusters and one such topic cluster can also comprise of several members in the organization who have not communicated with each other earlier. At the time when a new email is composed, every addressee will be categorized to be a potential leak recipient or one that is legal. Such classification was based on the emails sent among the sender and the receiver and also on their topic clusters. The work had investigated the technique of K-Means clustering and also proposed a Tabu - K-Means (TABU-KM) technique of clustering to identify points of optimal clustering. The proposed TABU-KM optimizes the K-Means clustering. Experimental results demonstrated that the proposed method achieves higher True Positive Rate (TPR) for known and unknown recipient and lower False Positive Rate (FPR) for known and unknown recipient.***

***Keywords: Data leakage prevention, email leakage, K-Means clustering technique and Tabu K-Means (Tabu-KM) clustering technique.***

## I. INTRODUCTION

Today, the organizations are being harmed by data that is exposed to parties that are unauthorised [1]. These data leaks may result in harm in different ways. When confidential data is handled in an improper manner, government regulations may get violated and this can result in fines. The companies are held liable for releasing the information on employees or customers such as their social security numbers or credit cards. Furthermore, it may result in loss of business and can pose a threat to the organization if the loss of proprietary information is to competitors.

A leak of data also involves the release of various sensitive information to any third party that is untrusted which may either be intentional or unintentional. There are different vendors offering products to prevent data leakage but academic research on the problem is surprisingly rare. According to the study made by survey reports [2], most threats to

information security have been triggered by data leakage. Such internal threat contain an estimated level of 29% private or sensitive data leakage that is accidental, another 16% approximately to the theft of intellectual property, another 15% to other thefts that include client information or monetary data or both. Also, there is a consensus of about 67% of organizations showing damage from internal threats to be more serious compared to the external threats. Since the activities of modern business are dependent on extensive usage of emails, an email leakage with “wrong recipients” has now become very widespread causing severe damage thus resulting in a problem that is very disturbing to the individuals and the organizations. There have been several solutions that were attempted for analysing exchange of emails to avert them from being sent to the wrong addressees but have failed to bring about a satisfactory solution. There are several mistakes of email addressing that remain undetected and in several cases, the right recipients are marked wrongly to be an addressing mistake.

Data Leak Prevention (DLP) is the technique or product that tries mitigating threats to data leaks. The products of DLP are now available from several vendors like McAfee, Trend Micro, CA Technologies and Symantec. Contrastingly, the prevention of data leakage has been getting only scant attention in the research. This does not mean the problem has been solved but that the products have been limited to the threats addressed.

Some of the most common methods employed for email leaks are: The techniques of Textual Content and classification. In the technique of textual content for the problem of leak prediction was based on the messages and their textural content. The primary aim was to model the pairs of “addressee-message” and also forecast the pair that is least probable. The email text content was compared to the cosine similarity or the techniques of clustering [6]. For the method based on classification, there is information on the social network that is used. There are certain features of social networks that are used as the number of messages received, the number of messages sent and the number of times both addressees had been copied in one message. For the purpose of combining the features of the social and textual network, the scheme based on classification was employed. The primary idea had been to perform the prediction of the leak in two different steps.

**Revised Manuscript Received on December 30, 2019.**
***** Correspondence Author

**Jayavarapu Karthik***, Research Scholar, Computer Science and Engineering, Annamalai University, Chidambaram, India.
**Dr.V. Tamizhazhagan**, Assistant Professor, Information TechnologyAnnamalai University, Chidambaram, India.
**Dr.S.Narayana**, Professor, Computer Science and Engineering, Gudlavalleru Engineering College, Gudlavalleru, India.

The first one is to compute the textual similarity that scores with the procedure of cross-validation. In the next step, the network features are mined and will learn a new function with textual scores. In this work, methods of clustering were used for the identification of email leaks. Clustering algorithms generally look for building clusters using interrelated criteria to choose objects that are in the same cluster and are quite similar while at the same time undertaking an assurance that objects in different clusters are dissimilar. K-Means algorithms will provide efficient means to solve this as it finds optimal placement of K centres which act as the centroid of clusters formed. The Tabu Search (TS) algorithm takes the initial solutions as inputs and further accomplishes a local search with the memory structures and neighbourhood structures. It also mitigates local minima by permitting solutions which don't improve their objective function. The TS is utilized to solve several other problems. With regard to the problems of clustering, the complexity of high computations is a challenging and selection of parameter needed by the TS makes it unviable when compared to K-Means algorithm. The TS further keeps the memory of the best solution at a point of search and will return the solution where the algorithm is terminated.

Based on a TS structure, the Tabu – KM hybrid algorithm [3] utilizes the properties of optimization of the TS with the local ability of search of the K-Means algorithm and thus enhancing the clustering. In this algorithm, a spherical Tabu space in the solutions obtained so far for every iteration. The object also has the radius which is the least and the other objects not found in the Tabu space can choose a different cluster. It also has the smallest radius of the best-so-far centre. This will cause an avoidance from the local optima to identify improved solutions.

In this work, the TABU-KM used for a social network in the DLP is proposed. The remainder of the investigation has been organized thus. The related work in literature is discussed in Section 2. The methods employed are explained in Section 3. Section 4 discusses the experimental results.

## II. RELATED WORKS

Alsayat and El-Sayed [4] had proposed a context used for the task of detection of communities using clustering messages taken from large social data streams. The proposed framework makes use of the K-Means clustering algorithm with the Genetic Algorithm and the Optimized Cluster Distance (OCD) for the clustering of data. The proposed framework has a twofold goal which is overcoming problems in general K-Means and selecting the primary centroids that are the best with the Genetic Algorithm and maximizing the distance observed between the clusters by using the OCD methods pairwise for getting accurate clusters. There were several other metrics of cluster validation used for the evaluation of the proposed algorithm and its performance. This analysis proved the proposed method to provide better results of clustering with a novel use-case of user community grouping. The approach was further optimized and was also scalable for the social media data and its real-time clustering. Another approach for the prevention of email "slip-ups" was proposed by Zilberman et al., [5]. This approach was built on mail exchange analysis within the organization and its associates and also the identification of the associates who exchange emails that have mutual subjects. Every associate's subjects were used at the time of the phase of enforcement to detect any potential leakage. At the time a new mail is created and is to be sent, the addressee of every email are analysed. An approval of the recipient in case the content of the email belongs to a minimum of one single topic which is common to both the recipient and the sender. There is, however, a critical issue of prevention of information leaks of emails which is at the time a message is addressed accidentally to the recipients that are non-desired. This has become a very common issue that can harm both corporations and individuals. Carvalho and Cohen [6] had presented an initial attempt to solve this issue. This began by means of redefining this to be a task of outlier detection in which all unintended recipients were the outliers. After this, all real email examples are combined (from Enron Corpus) along with some of the leak-recipients that are carefully simulated in order to learn patterns of a textual network that are associated to the email leaks. The method could distinguish mail leaks for about eighty percent of the test cases and also outperformed existing methods. Kalyan and Chandrasekaran [7] had made a new proposal of selecting input variables that were relevant to this domain that was a straightforward and simple scheme of learning which was able to detect leak of information by using an analysis of the mail pattern. It employed this method on the real-life emails obtained from economic related organizations. Selecting these variables in a judicious manner helped in learning the patterns of mails and detected the violations in an efficient manner. There were encouraging preliminary results that had an accuracy of about 92%. The technique is being implemented as a commercial tool.Shvartzshnaider et al., [8] had advocated another novel design methodology which was employed for the DLP systems that were centred on a notion with Contextual Integrity (CI). The work used a framework of the CI to the abstract real-world exchange of communication to the formally defined flow of information in which the privacy policies had described the sequences of all admissible flows. The current methods have discovered all misdirected emails from the gateway or the user agent but this was not appropriate for the different environments of application. Pu et al., [9] had made a new proposal of a method of misdirected email detection that was based on the multi-attributes that are deployed on the side of the server. There are three different attributes that include a fingerprinting of email content, meta information, and social relationship. On the basis of the classification algorithm of the Support Vector Machine (VM), the experiments proved that they can detect all misdirected emails with an accuracy of about 91.6%.

## III. METHODOLOGY

For the purpose of this section, the TABU-KM clustering algorithm, the K-Means Clustering and the Hierarchical Agglomerative Clustering (HAC) were discussed.

### A. Dataset

There was an Enron email dataset that had been made public at the time of this investigation, courtesy of the Federal Energy Regulatory Commission. There were several veracity problems for this. These were gathered and then organized for a project known as the Cognitive Assistant that Learns and Organizes (CALO). A lot of the issues of integrity for this dataset were ideally resolved. It also contained different types of emails both official and personal [10]. Some emails were deleted owing to the appeals from the employees who were affected. This particular data set version contained about 517,431 mails retrieved from hundred and fifty one users that were spread across three thousand five hundred folders. The messages did not include any attachments.This dataset had the information on folders for all 151 employees. Every message in the folder contained the email address of both the receiver and the sender, the time, date, body, text, subject and certain other technical details.

### B. Hierarchical Agglomerative Clustering (HAC)

Analysis of Hierarchical Clustering is a method used widely that is separated into two different types: the agglomerative techniques that make a set of amalgamates of x instances into some universal sets with disruptive techniques that separates x instances into finer sets. However, practically speaking, the HAC is utilized by specifying the clusters in a manual way. For the HAC, common criteria will comprise of a sum of the within-group sums of squares along with the shortest possible distance among sets that triggers in a technique of single-link [11].The HAC is iterative that builds a tree, T, made over a dataset based on the linkage function. The linkage function l: 2X × 2X → R will score the actual merger of two different nodes that correspond to clusters with data points that are stored at the descendant leaves. This is initialized by the creation of a node for every data point. It proceeds in a new series of rounds. For every HAC round, two nodes minimizing linkage function will be merged thus making them siblings to one another and further creating new nodes as parents. This algorithm will terminate after its final merge and this creates the actual root for the tree [12]. The advantages of this HAC approach are shown below:

- Easy and simplicity of computation and implementation.
- Less number of restrictions and more flexibility levels: the HAC uses simple information on the connectivity of quantitative data using the Received Signal Strength (RSS) or the GPS. Additionally, there are other factors that are incorporated into this algorithm. For example, there are different weights assigned to connections or nodes for various scenarios.
- Lesser resource requirements for the establishment of clusters: With the HAC approach, the nodes will complete the election, announcement, establishment, and scheduling of the clusters simultaneously. This reduced the dissipation of resources to a great extent.
- Works without any need to periodically re-cluster or update network: This HAC approach can generate a CH backup that was logical in the process of cluster generation making them easily adaptive to changes. They do not need any additional periodic updates.

### C. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a statistical measure that reflects the importance of a word to documents in the corpus or collection. TF – denotes the actual amount of times a word appears in a document measured with raw frequency divided by maximum raw frequency for a word in a document [13]. The cluster centroids are computed by using the K-Mean Clustering on the vectors of TF-IDF(c) found in a training dataset.

### D. K-Means Clustering Algorithm

The K-Means is a very simple learning algorithm that is unsupervised to solve problems on clustering. It tags the given dataset by defining the amount of K clusters that are fixed a priori. For every cluster, there is a centroid associated and this is the centre of gravity placed inside the problem space. The primary notion here is to split the samples of the dataset into the K groups (or clusters) where every object has some common traits to the other objects on a similar cluster maximizing inter-cluster distances and simultaneously minimizing intra-cluster distances.This algorithm begins using a random set of the K initial centre points of a cluster which is Ci (i = 1 …K), and these denote the current centroids. Firstly, the distance is computed (also called the measure of dissimilarity) for every object to every cluster and this is associated with its closest centroid. Once this is done, K-new centroids are recalculated. As a result of the previous step. Both steps will be repeated iteratively until a convergence takes place and this may be the assignment of centroids that do not change until a finite set of iterations is reached. With regard to computing the distance between instances and their clusters, it aims at optimizing objective function (f) in (4):

$$f = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ i \in G_i}}^{N} \| x_j - C_i \|^2 \tag{4}$$

Wherein, K denotes the amount of clusters, N the quantity of instances, $x_j$ the coordinate of instance, $C_i$ the coordinate of cluster i and $G_i$ the group of all instances of cluster i.

This algorithm further pushes around cluster centres within the space to bring down the squared distances within the cluster. For every cluster, there is a new centroid that is recalculated by means of averaging the object and locations. Computing the centroid is done as in equation (5):

$$C_i = \frac{1}{|G_i|}\sum_{\substack{j=1 \\ i\in G_i}}^{N} x_j \qquad (5)$$

Wherein, | $G_i$ | denotes the actual quantity of objects that are in cluster i.The K-Means will produce a new separation of objects into some groups that have the metrics that are minimized which is computed. The K-Means is a very popular algorithm as it is flexible, easy to implement, straight-forward and simple. Even though it is useful in an extensive manner, it does suffer from certain limitations. The number of the K clusters should be known well in advance. The K-means objective function will not be convex and may also contain some local optima. Thus, it tends to get caught in the local minima (or the local maxima and sometimes the saddle point). Its efficiency is dependent on the initial centroids and it is also quite sensitive to outlier and noises. The clustering of data is not very suitable to the clusters in density. It cannot be applied to the calculation of average and data collection and is limited only to numerical data.

### E. Tabu Search (TS) Algorithm

The TS denotes a strategy of metaheuristic search that was introduced by Glover which was applied to a varied range of problems in optimization. It is a method of a single-solution neighbourhood search that employs flexible memory for avoiding getting trapped in the local optima. The principle of the TS was to pursue a search at the time of encountering local optimum. Moving back to the solutions visited earlier has been banned by making use of a memory known as the TABU list. This list will record all recent moves and will direct the search correctly. For overcoming all unwanted limitations of the TABU list, the TS is equipped with a very efficient device called the aspiration criteria which permits revoking unwanted TABU. Another simple aspiration criteria were defined as the move that had an objective value which was better compared to the current solutions. This will be in case they attempt at minimizing objective function f (θ) on the proper domain. There have been various methods for stopping the process of search and once a set of pre-defined iterations are complete, the objective value may be identified to be a value that is lower than any small threshold.

### F. Proposed Tabu-KM Algorithm

The proposed scheme of classification is based on the exchange of email traffic. It has been assumed that any user can be a part of various groups that work on several distinct topics. In the next phase, every new email to be sent will be analysed as follows: for every recipient, the email will be checked to see if both recipient and sender are part of a common topic group. In case there is no such group, they come to the conclusion that there is no common topic for both users. Then the recipient mentioned above is not correct. If not, the email and its content will be compared to the email content that has been exchanged.

A new model of classification has two phases which are the training and the classification phase. Training will be used on a new set of mails called "leak-free" and are classified which will be utilized on the new emails composed that are characterized to be the queries. Every email will have content that is represented by means of a TF-IDF vector. Taking the TS structure into consideration, a hybrid of the TABU-KM has been intended and further applied the optimization quality of the TS with the local ability of search of the K-Means. In the proposed algorithm, there is a spherical Tabu space that is around the solution that is the current best for every repetition. Furthermore, the other instances unavailable inside the Tabu space have been allowed to select the new cluster as its centre. The object also has the least radius of the centre that is the best-so-far. This may result in it escaping from the local optima and identify some solutions that are better [3].The configuration will be a solution of numeric to the variables. In the initial stage of this algorithm, the K-Means will generate an answer that is feasible. After this, the centroids for the clusters will be computed. These clusters will be chosen sequentially for the generation of a new solution by means of the logic of this algorithm. The starting point for the algorithm will be the centre of the chosen cluster. Identified are two sorts of Tabu found in the Tabu-KM system which are the Tabu space and the Tabu list. The reason behind this was the use of a Tabu space to contain all forbidden centres of the cluster centre. There are two strategies used to implement the Tabu spaces that have been investigated and they are the Static and the Dynamic. Once the Tabu space is made or extended, its object will be located out of the space to be a new centre for the cluster. Once the algorithm takes place, an investigation of the cluster centre is made. If the new cluster centre is inside the TABU space, this will only mean they are found inside the local optimal condition and will have to change their direction by means of restricting the object's reselection. If the new cluster centre is not found within the TABU space, the subsequent movement's direction is determined in accordance with the value of the objective function and the determination of either an improvement of a solution or its non-improvement. A move denotes a new process that generates a viable answer to the clustering issue which is connected to the present answer. During the iteration, a new centre of the cluster found in the neighbourhood is chosen from its current cluster centre that is unavailable in the Tabu space. The three different strategies of the candidate changes in the cluster is explored.

(a) Move towards the instance that is closest to the centre of the K-Means answer: the space that is spherical in shape which is in the centre of the K-means will be investigated. For this strategy, the spherical space's radius will be increased based on the remoteness of an instance to the centre of its initial K-means.

(b) Move towards the instance that is the closest to the centre of its current solution: space which is spherical and is currently around the centre of the cluster is now investigated. This new centre is chosen from various instances found in the neighbourhood of its current centre and this is unavailable in the Tabu space. It has to be understood that the present answer doesn't have to be improved and so a feasible solution can be generated in a different direction also.

(c) Move towards the instance that is the closest to the centre of the solution which is best-so-far: this will be the spherical space that is around the cluster centre for the solution that is the best-so-far which is examined. The spherical search space and its radius will be increased in order to allow the other instances that are unavailable in the TABU space to be the centre of its new cluster. Once a better solution is identified, this spherical space will be taken into consideration and this will be around the best-so-far solution's centre and this can be in a different direction as well. Once the K-Means in the TABU-KM algorithm is performed, in case the new centre is within the TABU space, the solution and its value will be computed in accordance with its new objective function. In case the solution has a much more improved value of its best solution until now, this solution will be accepted to be the solution that is best-so-far. Even though it comes back to the TABU space, it will have to restrict a reselection of this object since the cluster centre owing to the satisfying of the criterion of aspiration will have the solution that is best-so-fat to be updated by its current solution. This process will come to an end at the time all the clusters have been investigated in a sequential manner without any major enhancement being found in its best-so-far solution. For the study of every such cluster, the centre of the other clusters can be altered and thus it is important to be able to restudy all the clusters in order to arrive at a answer that is better. The diagram for the TABU K-Means algorithm is depicted in Figure 1.

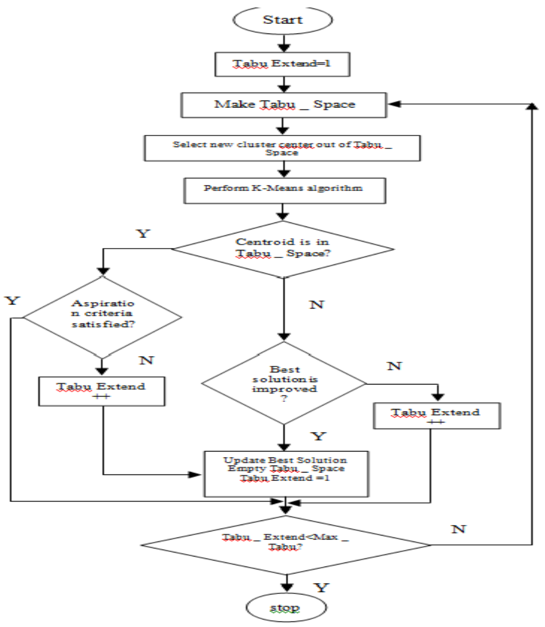
Start
Tabu Extend=1
Make Tabu _ Space
Select new cluster center out of Tabu _ Space
Perform K-Means algorithm
Centroid is in Tabu _ Space?
Y
N
Aspiration criteria satisfied?
Y
N
Tabu Extend ++
Best solution is improved ?
N
Y
Tabu Extend ++
Update Best Solution
Empty Tabu _ Space
Tabu Extend =1
Tabu _ Extend<Max _ Tabu?
N
Y
stop

**Figure 1 Flow Diagram of Tabu K-Means Algorithm**

## IV. RESULTS AND DISCUSSION

In this section, the HAC, K-Means and Tabu K-Means methods are used. Table 1 shows the summary of results. Table 2 shows the True Positive (TP) obtained. The True Positive Rate (TPR) for known and unknown recipient and False Positive Rate (FPR) for known and unknown recipient and TP obtained as shown in figures 2 to 4.

**Table 1 Summary of Results**

| | HAC | K-Means | Tabu K-Means |
|---|---|---|---|
| TPR for Known recipient | 0.8231 | 0.8377 | 0.8468 |
| TPR for Unknown recipient | 0.8407 | 0.8466 | 0.8642 |
| FPR for Known recipient | 0.1593 | 0.1534 | 0.1358 |
| FPR for Unknown recipient | 0.1769 | 0.1623 | 0.1532 |

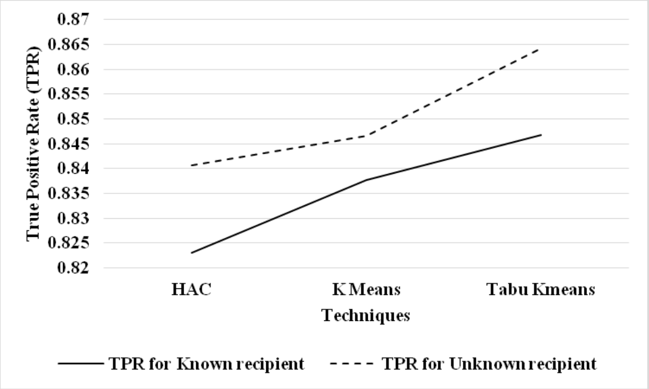

**Figure 2 True Positive Rate (TPR)**

From the figure 2, it can be observed that the Tabu K-Means has higher TPR for known recipient by 2.83% for HAC and by 1.08% for K-Means. The Tabu K-Means has higher TPR for unknown recipient by 2.75% for HAC and by 2.05% for K-Means.

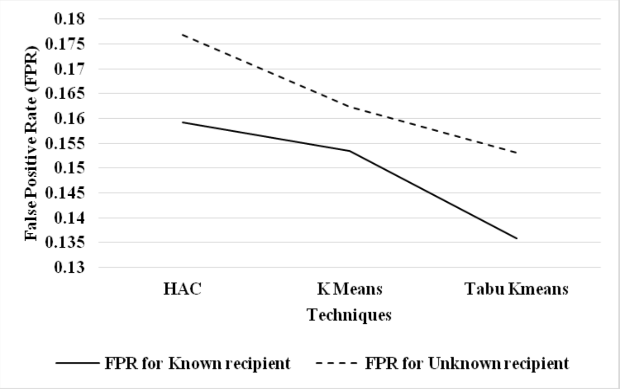

**Figure 3 False Positive Rate (FPR)**

From the figure 3, it can be observed that the Tabu K-Means has higher FPR for known recipient by 15.92% for HAC and by 12.17% for K-Means. The Tabu K-Means has higher FPR for unknown recipient by 14.35% for HAC and by 5.77% for K-Means.

**Table 2 True Positive (TP) Obtained**

| FP | HAC | K Means | Tabu K-means |
|---|---|---|---|
| 0 | 0.66 | 0.68 | 0.69 |
| 0.1 | 0.66 | 0.68 | 0.7 |
| 0.1 | 0.65 | 0.68 | 0.69 |
| 0.1 | 0.64 | 0.67 | 0.68 |
| 0.1 | 0.65 | 0.66 | 0.69 |
| 0.1 | 0.66 | 0.67 | 0.69 |
| 0.1 | 0.64 | 0.67 | 0.7 |
| 0.1 | 0.67 | 0.69 | 0.71 |
| 0.1 | 0.66 | 0.68 | 0.7 |
| 0.1 | 0.65 | 0.68 | 0.7 |
| 0.1 | 0.65 | 0.67 | 0.7 |
| 0.1 | 0.65 | 0.68 | 0.7 |
| 0.1 | 0.65 | 0.67 | 0.69 |
| 0.1 | 0.66 | 0.68 | 0.7 |
| 0.1 | 0.64 | 0.67 | 0.68 |
| 0.1 | 0.64 | 0.67 | 0.67 |
| 0.1 | 0.66 | 0.67 | 0.69 |
| 0.1 | 0.66 | 0.68 | 0.69 |
| 0.1 | 0.69 | 0.7 | 0.72 |
| 0.1 | 0.69 | 0.71 | 0.74 |
| 0.1 | 0.68 | 0.7 | 0.73 |
| 0.1 | 0.7 | 0.72 | 0.74 |
| 0.1 | 0.72 | 0.74 | 0.77 |
| 0.1 | 0.7 | 0.73 | 0.75 |
| 0.1 | 0.72 | 0.75 | 0.77 |
| 0.12 | 0.71 | 0.74 | 0.76 |
| 0.12 | 0.71 | 0.74 | 0.77 |
| 0.12 | 0.72 | 0.75 | 0.77 |
| 0.14 | 0.72 | 0.74 | 0.76 |
| 0.16 | 0.72 | 0.75 | 0.77 |
| 0.18 | 0.72 | 0.74 | 0.77 |
| 0.18 | 0.74 | 0.77 | 0.79 |
| 0.2 | 0.72 | 0.75 | 0.78 |
| 0.24 | 0.75 | 0.77 | 0.79 |
| 0.24 | 0.76 | 0.78 | 0.8 |
| 0.24 | 0.76 | 0.79 | 0.81 |
| 0.26 | 0.76 | 0.78 | 0.8 |
| 0.32 | 0.76 | 0.78 | 0.8 |
| 0.32 | 0.77 | 0.79 | 0.82 |
| 0.32 | 0.75 | 0.78 | 0.81 |
| 0.34 | 0.78 | 0.8 | 0.82 |
| 0.36 | 0.78 | 0.8 | 0.82 |
| 0.36 | 0.76 | 0.79 | 0.81 |
| 0.36 | 0.76 | 0.79 | 0.81 |
| 0.38 | 0.78 | 0.8 | 0.82 |
| 0.4 | 0.78 | 0.8 | 0.82 |
| 0.42 | 0.81 | 0.83 | 0.85 |
| 0.42 | 0.81 | 0.84 | 0.86 |
| 0.44 | 0.79 | 0.82 | 0.84 |
| 0.44 | 0.8 | 0.83 | 0.86 |
| 0.48 | 0.8 | 0.82 | 0.85 |
| 0.5 | 0.8 | 0.83 | 0.86 |
| 0.52 | 0.79 | 0.81 | 0.84 |
| 0.54 | 0.8 | 0.83 | 0.86 |
| 0.56 | 0.82 | 0.84 | 0.87 |
| 0.56 | 0.82 | 0.84 | 0.88 |
| 0.58 | 0.81 | 0.83 | 0.86 |
| 0.6 | 0.82 | 0.84 | 0.86 |
| 0.6 | 0.82 | 0.84 | 0.86 |
| 0.62 | 0.81 | 0.84 | 0.87 |
| 0.64 | 0.83 | 0.85 | 0.87 |
| 0.66 | 0.83 | 0.86 | 0.88 |
| 0.66 | 0.84 | 0.86 | 0.89 |
| 0.68 | 0.83 | 0.86 | 0.89 |
| 0.7 | 0.83 | 0.86 | 0.89 |
| 0.72 | 0.83 | 0.86 | 0.89 |
| 0.72 | 0.83 | 0.86 | 0.88 |
| 0.74 | 0.82 | 0.85 | 0.88 |
| 0.76 | 0.84 | 0.87 | 0.9 |
| 0.78 | 0.83 | 0.86 | 0.89 |
| 0.8 | 0.86 | 0.9 | 0.92 |
| 0.8 | 0.89 | 0.92 | 0.94 |
| 0.82 | 0.87 | 0.9 | 1 |
| 0.84 | 0.89 | 0.91 | 1 |
| 0.86 | 0.87 | 1 | 1 |
| 0.88 | 0.92 | 1 | 1 |
| 0.88 | 0.94 | 1 | 1 |
| 0.9 | 1 | 1 | 1 |
| 0.92 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 |

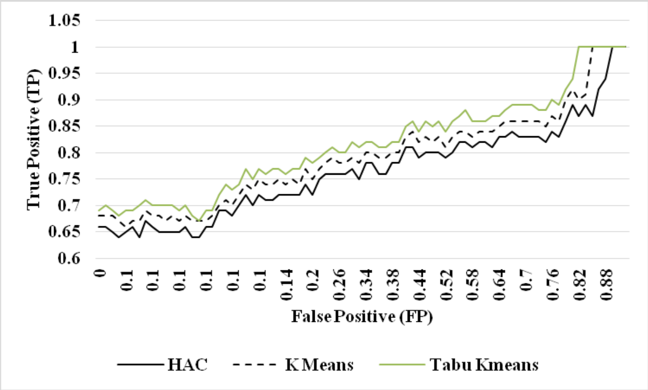

**Figure 4 True Positive (TP) Obtained**

From the figure 4, it can be observed that the Tabu K-Means has higher average TP for 6.2% for HAC and by 2.86% for K-Means respectively.

## IV. CONCLUSION

The increase in the leakage incidents and its resulting cost, the threat to the leakage of data has become a critical security issue to the organizations. DLP techniques are utilized to prevent leakage. One of the popular method of DLP is based on clustering which groups instances into several classes in order to ensure the objects of each class (or cluster) are very similar to the rules of the criteria. The K-Means algorithm will be the primary method or partition clustering that is popular owing to the simplicity of its computation. But, it can be a highly sensitive aspect to selecting of the various initial cluster centroids which is trapped inside the local minima. In this work, an effective algorithm that is based on the approach of Tabu Search known as the TABU- KM was developed by the integration of the Tabu space and change the generator to restrict all instances to choose it as the cluster centre. The Tabu-KM algorithm has been used for mitigating the local optima to find solutions for better clustering. The results have proved that the TABU K-Means will have a higher TPR which is for the recipient by about 2.83% for the HAC and by about 1.08% for the K-Means. The TABU K-Means will have a higher TPR is for an unknown recipient by about 2.75% for the HAC and further by about 2.05% for the K-Means. This TABU K-Means will have a higher FPR for the known recipient by about 15.92% for the HAC and further by about 12.17% for the K-Means. TABU K-Means also has a higher FPR for the unknown recipient by about 14.35% for the HAC and further by about 5.77% for the K-Means.

## REFERENCE

1. Shvartzshnaider, Y., Pavlinovic, Z., Balashankar, A., Wies, T., Subramanian, L., Nissenbaum, H., & Mittal, P. (2019, May). VACCINE: Using Contextual Integrity ForData Leakage Detection. In The World Wide Web Conference (pp. 1702-1712). ACM.
2. Pu, Y., Shi, J., Chen, X., Guo, L., & Liu, T. (2015, July). Towards misdirected email detection based on multi-attributes. In 2015 IEEE Symposium on Computers and Communication (ISCC) (pp. 796-802). IEEE.
3. Shetty, J., & Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. Information sciences institute technical report, University of Southern California, 4(1), 120-128.
4. Tu, Q., Lu, J. F., Yuan, B., Tang, J. B., & Yang, J. Y. (2012). Density-based hierarchical clustering for streaming data. Pattern Recognition Letters, 33(5), 641-645.
5. Yadav, N., Kobren, A., Monath, N., & McCallum, A. (2019). Supervised Hierarchical Clustering with Exponential Linkage. arXiv preprint arXiv:1906.07859.
6. Dang, N. C., De la Prieta, F., Corchado, J. M., & Moreno, M. N. (2016, June). Framework for retrieving relevant contents related to fashion from online social network data. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 335-347). Springer, Cham.
7. Raman, P., Kayacık, H. G., & Somayaji, A. (2011, June). Understanding data leak prevention. In 6th Annual Symposium on Information Assurance (ASIA'11) (p. 27).
8. Yu, X., Tian, Z., Qiu, J., & Jiang, F. (2018). A data leakage prevention method based on the reduction of confidential and context terms for smart mobile devices. Wireless Communications and Mobile Computing, 2018.
9. Yaghini, M., & Ghazanfari, N. (2010). Tabu-KM: a hybrid clustering algorithm based on tabu search approach.
10. Alsayat, A., & El-Sayed, H. (2016, June). Social media analysis using optimized K-Means clustering. In 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 61-66). IEEE.
11. Zilberman, P., Dolev, S., Katz, G., Elovici, Y., & Shabtai, A. (2011, July). Analyzing group communication for preventing data leakage via email. In Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics (pp. 37-41). IEEE.
12. Carvalho, V. R., & Cohen, W. W. (2007, April). Preventing information leaks in email. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 68-77). Society for Industrial and Applied Mathematics.
13. Kalyan, C., & Chandrasekaran, K. (2007, August). Information leak detection in financial e-mails using mail pattern analysis under partial information. In AIC'07: Proceedings of the 7th Conference on 7th WSEAS International Conference on Applied Informatics and Communications (pp. 104-109).

## AUTHORS PROFILE

**Mr.J.Karthik** completed his B.Tech(CSE) from J.N.T.University,Kakinada in 2010 and M.Tech(CSE) fromJ.N.T.University,Kakinada in 2012. He is currently pursuing Ph.D. in Annamalai University and working as Assistant Professor in Department of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru.. He has published five research papers in reputed international and 1 in International conference and it's also available online. His main research work focuses on Data Mining , Information Security, Cloud Computing. He has 7 years 6 months of teaching experience.

**Dr.V.Tamizhazhagan** working as Assistant Professor in Department of Information Technology, Annamalai University. He has published five research papers in reputed International Journals and Three National and two International conferences and it's also available online. He attended & organized various workshops. His main research work focuses on Data Mining, Network Security, Mobile Computing and Computer Networks. He has 15 years of teaching experience. He guided 7 Post graduation and many under graduation projects.

**Dr.Narayana Satyala** completed his B.Tech, M.TechandPh.D.fromJ.N.T.University,hyderbad.He is working as Professor in Department of Computer Science Engineering, Gudlavalleru EngineeringCollege,Gudlavalleru..Hehaspublished 11 research papers in reputed international and 2 in International conference and it's also available online. His main research work focuses on Data Mining,Machine Learning. He has 20 years of teaching experience. He guided 12 Post graduation and 20 Under graduation projects.