DECEIVING COMMUNICATION LINKS ON AN ORGANIZATION EMAIL CORPUS

R. Divya , Ch. Aswani Kumar*, S. Saijanani and M. Priyadharshini

School of Information Technology and Engineering, VIT University,
Vellore – 632014, India.
*Email: cherukuri@acm.org

ABSTRACT

Word use in email is correlated to the role of the individuals within the organization and relationships among them are based on their patterns of word use. In this paper, using the matrix-decomposition techniques we determine the communication links with individuals who conversed frequently based on their word usage pattern. Further we analyze the emails of all the members of the communication link and isolate the individuals who have surprising results. Later we follow the deception theory, which suggests that deceptive writing is characterized by reduced frequency of first person pronouns, exclusive words, elevated frequency of negative emotion words and action verbs.

Keywords: *Deceptive emails, Non-Negative Matrix Factorization, Semi Discrete Decomposition, Singular Value Decomposition*.

1. Introduction

Email use is now ubiquitous in many large and small organizations in industry, government and in academic circles. Few issues associated with email are spamming, phishing, computer worms, etc [6, 18]. At the organization level two main types of crimes, namely White -collar crime and Blue-collar crimes exist. The white-collar crime is a term that is applied to nonviolent crimes committed in business situations by individuals, groups, or corporations for the purpose of financial gain. Many white-collar criminals rely on cellular phones, electronic banking, emails and computer databases to commit their crimes. The increase in frequency and magnitude of white-collar crimes is approximately during the same period that many organizations adopted email as their primary form of communication [3, 20]. Therefore as an electronic means of communication, email should be analyzed in a manner unlike its predecessors to detect such crimes. The core objective of our work is to understand the characteristics and structure of both normal and abnormal emails that provide information about how such email data might be better analyzed in an intelligence setting. Further to examine the structure of the email data, we look for the most frequently used words in the email dataset, and thereby understand what kind of information about the individuals within the organization could be extracted from their word usage (members with the same word usage pattern form a communication link) [15]. On categorizing emails of the individuals based on a few cues (i.e. official and unusual), suspicious members are spotted. The appearance of every individual is then detected to make the decision about the communication link. Matrix decomposition techniques are used to accomplish these tasks.

Section 2 provides the background on deception by presenting the related works and the list of document collections. Section 3 reviews the matrix decomposition techniques, which are used to decompose the large matrix to produce the minimized plot values. We propose a novel technique, which discovers both email communication links and the status of those links within the organization in Section 4. In the penultimate section we present the experimental analysis on the constructed matrices and an analysis of the results, followed by the conclusion, acknowledgement and references.

2. Related Work

As individuals increase the usage of electronic communication, there has been an increase in the research in detecting deception in these new forms of communication. Models of deception assume that deception leaves a footprint. The following are the few other works carried on with the email dataset regarding the deception detection. Skillicorn [19] shows that matrix decompositions, in particular Singular Value Decomposition (SVD) and Semi Discrete Decomposition (SDD), have several useful properties to identify a terrorist group with few false positives. Berry and Browne [2] applied a Non-negative Matrix Factorization (NMF) approach for the extraction and detection of concepts or topics from electronic mail messages. Rajaram and Balamurugan [16] developed a data mining approach to detect deceptive communication in email text. They applied deception model to the set of 300 emails and then applied ID3 algorithm to generate the decision tree which is used to test the email as deceptive or not. Keila and Skillicon [5] developed a method based on the Singular Value Decomposition (SVD) to detect unusual and deceptive communication in emails. However, their work did not concentrate on tracing the email communication links in the organization or categorizing the individuals, thus the status of the communication link could not be determined. We have extended upon this analysis by finding the communication links with an organizational email corpus, where employees with frequent and/or similar email exchanges formed a group. Later on the behavior of the individuals on the communication link was scrutinized to check if their profiles are official or unofficial. Further, the deception model is applied to justify the type of the linkage. In order to analyze the emails we consider Enron dataset, word list (stop words, organizational words, and business words) and cues for categorizing emails (official and unusual).

Fewer first-person pronouns (I, me, mine, etc.) may indicate authors' attempts to "dissociate" themselves from their words. Fewer exclusive words (but, except, without, although, etc.) indicate a less cognitively complex 'story' that is easier to create and to remember consistently [21]. An increased frequency of action verbs (lead, going, taking, action, etc.) may be an artifact of reducing the number of exclusive words, or the result of attempts to distract from the lack of subtlety by including plenty of action. The increased frequencies of negative emotion (despise, dislike, abandon, afraid, agony, etc.) words indicate some degree of self-respect dissonance about the fact of the deception [5]. The Enron email dataset is the first large-scale collection of real world email released into the public domain [4]. Though the vast majority of the communication is completely innocent, the emails of a number of top executives who are currently being prosecuted are in the dataset. Hence, it is reasonable to believe that evidence of deception exists within the dataset. We considered the inbox collection (inbox and sent_items) of a few employees in Enron, which consists of 46 employees and their 60549 mails. We collected a set of business words, a set of words pertaining to Enron and retained those words that pertain to the organization. From the obtained list, stop words were filtered. The cues for categorizing emails into unusual and official mails were collected (terms relating to sports, personal, legal contracts... etc) [12]. The attributes for deception (first person pronoun, exclusive words, negative emotion words and action words) were identified [14, 22]. Finally, we got the list of words that are most frequently used within the organization (Enron speak words).Later cues for categorizing emails are framed.

3. Matrix Decomposition Techniques

A dataset contains information about n objects with m attributes about each of them. Such a dataset is naturally viewed as a matrix **A** with n rows and m columns. Matrix decomposition expresses the matrix **A** as a product of other matrices in a way that reveals matrix structure. Several decomposition techniques are found in the literature [7, 8, 9, 10]. In our work we consider three matrix decomposition techniques termed, SVD, SDD and NMF.

3.1 Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) of a matrix **A** reduces it into three matrices as, $\mathbf{A}=\mathbf{USV}^T$. If **A** is of dimension $n \times m$ and has rank r, then **U** is $n \times r$, **S** is an $r \times r$ diagonal matrix with non-increasing entries $\sigma_1, \sigma_2 \ldots \sigma_r$ (the singular values), and **V** is $m \times r$. Both **U** and **V** are orthogonal, so that $\mathbf{UU}^T= \mathbf{I}$ and $\mathbf{VV}^T = \mathbf{I}$ [17, 19]. **U** is a matrix whose columns are the eigenvectors of the $\mathbf{AA}^T$ matrix and are termed as the left eigenvectors. **S** is a matrix whose diagonal elements are the singular values of **A**. **V** is a matrix whose columns are the eigenvectors of the $\mathbf{A}^T\mathbf{A}$ matrix and termed as the right eigenvectors. This decomposition not only provides a direct method for computing the rank of a matrix, but exposes other equally interesting properties and features of matrices [1, 24]. SVD has a number of subtle, yet useful properties: 1. The correlation between two objects is proportional to the dot product between their positions regarded as vectors from the origin. Two highly correlated objects will have a large and positive dot product. Two negatively correlated objects will have a large and negative dot product. Uncorrelated objects will have a dot product close to zero. 2. Multiplying row(s) or column(s) of the dataset by a scalar effectively changes their influence on the entire decomposition. If the scalar is greater than one, the effect is to move the points corresponding to these rows or columns further from the origin, and hence makes these points seem more interesting. However, this also has the useful side-effect of 'pulling' points that are correlated with the weighed points further from the origin as well.

3.2 Semi Discrete Decomposition (SDD)

The Semi Discrete Decomposition (SDD) of a matrix decomposes it into three matrices as, $\mathbf{A}=\mathbf{XDY}$, where the entries of **X** and **Y** come from the set $\{-1, 0, +1\}$, **D** is a diagonal matrix. If **A** is of dimension $n \times m$ and has rank r, then **X** is $n \times$

r, **D** is $r \times r$ and **Y** is $r \times m$ [6, 26]. The natural interpretation of SDD is a layered one. Each $\mathbf{A}_i$ corresponds to a column of **X** and a row of **Y**, weighed by an entry from **D**. The product of $\mathbf{X}_i$ and $\mathbf{Y}_i$ is a stencil representing a 'bump' (where the product has a +1) and corresponding 'ditch' (where the product has a −1) [13]. SDD generates a ternary, unsupervised hierarchical classification of the samples, based on the values in each successive column of the X matrix. Consider the first column of X. Those samples for which this column has the value +1 can be grouped; those samples for which this column has the value −1 forms another group; and those samples for which this column has the value 0 are unclassified at this level. This can be repeated for columns 2, 3, and so on, to produce a classification tree. An effective approach to analyze complex datasets is to use SVD and SDD together; i.e. use SVD to position items and SDD to classify them.

3.3 Non Negative Matrix Factorization (NMF)

The Non Negative Matrix Factorization of a Matrix **X** with dimension m x n is, **X**=**WH** where each column is an m-dimensional non-negative matrix of n vectors. NMF finds two new reduced dimensional matrices **W** and **H**, in order to approximate the original matrix **X** by the product **WH** in terms of a metric. The dimensions of matrices **W** and **H** are $m \times r$ and $r \times n$, respectively [9]. NMF is a vector space method used to obtain a representation of data using non-negativity constraints. These constraints can lead to a parts-based representation because they allow only additive, not subtractive, combinations of the original data. Advantages of NMF are, Sparsity and Nonnegativity; Reduction in storage; Interpretability [2, 10].

4. Proposed Technique

We propose a novel technique, which discovers both email communication links and the status of those links within the organization through the steps framed as follows, (1) emails are parsed into the suspicious link sensor, through which both the normal and abnormal link are revealed. A human reader is involved to distinguish them. (2) Link members profiles are checked by the categorizer. (3) The occurrence of link members within the deceptive region are identified.

4.1 Suspicious Link Sensor

For our work on word usage of individuals, we create a matrix whose objects are individuals and whose columns are word frequency, aggregated over all of their emails in the dataset. This captures a characteristic word use pattern for an individual. Individuals having analogous responsibilities in the organization and same designation might use similar words, because of likelihood in the area under discussion and their distinct way of writing. Therefore, on applying decomposition techniques to the matrix, the correlation in word use patterns, places the individuals with similar word patterns close together. This can be done by one of the SVD properties, in which any of the employee's profile value (the particular employee's row value in the matrix) is weighed by a factor. Now the other individuals with similar word-use will move along the weighed employee. Using this approach with other employees, authorities can ascertain a set of communication links.

By emphasizing certain words that key individuals use (in our case, Lay and Skilling), other individuals who used the same words change their position in the plot. Besides several reasons for sharing the same words, this does indicate some form of communication link. Now, authorities can explore links between individuals who are not obviously associated with one another by their role or position in the company. We create a ranking of the words based on their global frequencies to correctly isolate individuals who use less "Enron speak words" initially [18]. Later on the individuals who have similar ranking of their profile words are made visible. So this remains as another method to detect a communication link. Thus, word-frequency profiles clearly show the role played by an individual in the organization and the relationships among individuals based on their patterns of word use in email as word use is associated with function within the organization. The word use among those involved in suspected criminal activity may be slightly distinctive leading us to a forceful analysis of those individuals who are likely to be deceptive. As a result, our approach creates a visual representation of the data that allows us quickly, and easily, to pick out the key players.

4.2 Categorizer

Our next approach is to inspect those individuals who had a suspicious communication link from the results of the previous approach more deeply. In this strategy we will determine the nature of the individual's communication i.e. if their mails are linked to unusual subjects (sports, personal. etc.) or related to official matter (legal contracts, document collection.). For this we create a matrix, wherein the rows represent each email within the dataset and its columns are cues to categorize the emails (unusual and official).Individuals falling under the unusual category may be officially responsible for it like (sports event coordination ... etc.) or must be misusing the resources provided to them (wasting time on personal matters .. etc). Thus these individuals are not deceptive in behavior. Individuals coming under the official category whose mails are more of organizational issues are expected to be deceptive who might be involved in fraudulent activities. Thus for our ensuing work on deception detection we focus more of our attention on these individuals.

4.3 Deception Detector

Deception theory proposes that deceptive writing generally includes reduced use of first - person pronouns and exclusive words, and more of negative emotion words and action verbs. The use of first-person pronouns are a slight declaration of a person's possession over his statement. Hence in order to detach themselves from their words, deceivers use fewer first-person pronouns. More usage of exclusive words might raise a doubt if the message is fallacious. Hence liars employ reduced exclusive words in their mail. This clearly shows the negative association between the use of exclusive words and action-verbs. Hence, for a deceptive communication to be less knowledgably complex, it should have a greater frequency of action-verbs. Further, due to inner feelings of guilt, deceptive communication is characterized by greater use of words denoting negative emotions [14]. The use of more deceptive attributes in the cues may weaken the actual detection of detection. Therefore, we further refine the attributes for deception and retain only those attributes that contribute more to detecting deception. The result will give a more sensitive method of perceiving deception. We then label emails by their senders (based on their designation) to identify the distinction in the writing style between Enron employees and others. The decomposition techniques reflect the importance of attributes with increased values, but our work requires reduced number of first person pronouns and exclusive words. Hence, we modify the raw frequencies of these two cues by choosing a frequency pyramid. For example, the characteristic signature of reduced first person pronouns in other settings seems likely to be present in emails that contain 5-15 occurrences (some emails contain 50 or more occurrences). Hence we alter first person pronoun data to select a frequency pyramid centered around frequency 10, and set frequencies outside the range 5-15 to zero (used in section 5.2.3).

The above mentioned steps for deception detection are applied to a matrix in which the rows correspond to emails and columns represent deception cues. Each deception cue has a list of deceptive attributes. The entries for the matrix count the frequency of the deceptive cues in each mail. We apply matrix decomposition techniques to the obtained results and label every email by its sender and mail number. Mails in the deceptive region are checked with the suspicious deceptive communication link from word usage profiles and if there is an association, then a human reader can extract only those deceptive mails. Further, these emails are scrutinized to find the intent of the communication link, from which we could correctly trace out the employees who intentionally cause harm to the proper functioning of the organization.

5. Experimental Analysis

5.1 Experimental Setup

The most frequent communication links between the individuals are identified by working on the word frequency matrix and the email rank matrix. Later the individuals in the communication links emails are categorized using the categorizing email matrix, to know the general behavior of that individual within the organization. The deception cue and the deception model matrices help, to identify the position of an employee emails within the deceptive chart of the whole organizational email dataset. Thus we have constructed the following matrices from the Enron email dataset for our experimental analysis. Word frequency matrix contains individuals as rows and word frequency as columns, aggregated over all of their emails in the dataset. Hence each row captures a characteristic word use pattern for an individual. In this matrix the importance of the interesting individuals can be reflected by weighing the terms used by them. This has the effect of moving them away from the origin and making them seem more important and also causes correlated objects to follow them, which can be done by, adding weights to the individual's rows and/or columns in the raw data. In email rank matrix each row corresponds to an email, and contains the rank, of each of the words in the message, arranged in decreasing order. "Time" is the most common noun in English, so every email that contains the word "time" will have a 1 in the first column. The categorizing email matrix has its rows representing emails within the dataset and its columns are the two categorizing attributes of the emails. The unusual attribute is comprised of the cues that belong to sports, family issues, weekend plans, entertainment and official attribute is comprised of the cues that belong to legal contracts, document collection and government issues. Deceptive Cue Matrix has emails in its rows and four attributes in its columns, where the entries are the frequencies of deceptive cues of each Email. In deception mode matrix we reduce the attributes for each cue and fix a pyramid for both first person pronouns and exclusive words.

5.2 Analysis

The following sections comprise a discussion on the application of the decomposition techniques on the above constructed matrices.

5.2.1 Word Usage Pattern and Communication Link

The similar word usage and frequent contacts between employees results in a communication link between them. Here we apply SVD, SDD and NMF on the word frequency matrix to identify the communication links. Since both SVD and SDD has it is own drawbacks on using them independently, we plan to use both SVD and SDD jointly to prevail most of their advantages. Fig. 1 shows the correlation in word use patterns. SVD of word frequency matrix with the first three columns of the matrix as its axis determines that individuals with similar word patterns will be placed close together. This trend can be observed in Fig. 1. Fig. 2 demonstrates the positions of individuals by their word use. The words used by two executives Lay and Skilling are weighed by a factor of 4 as discussed in section 4.1.

Compared to Fig. 1, where no weighing has been applied, in Fig. 2, the two executives Lay and Skilling move away from the origin. Since the importance of the words used by these two executives is increased, the other individuals who used some of the same words also change their position. On analyzing these two figures, we can observe that many other individuals make an appearance close to these two executives. This may reflect on particular topics about which, these individuals as well as two executives exchanged emails. In Fig. 3a profiles that have similar values are grouped by the SDD technique. Each cluster represents individuals with similar word usage pattern. Fig. 3b shows the clusters of individuals by word use; after the words used by two executives are adopted a weight factor of 4.

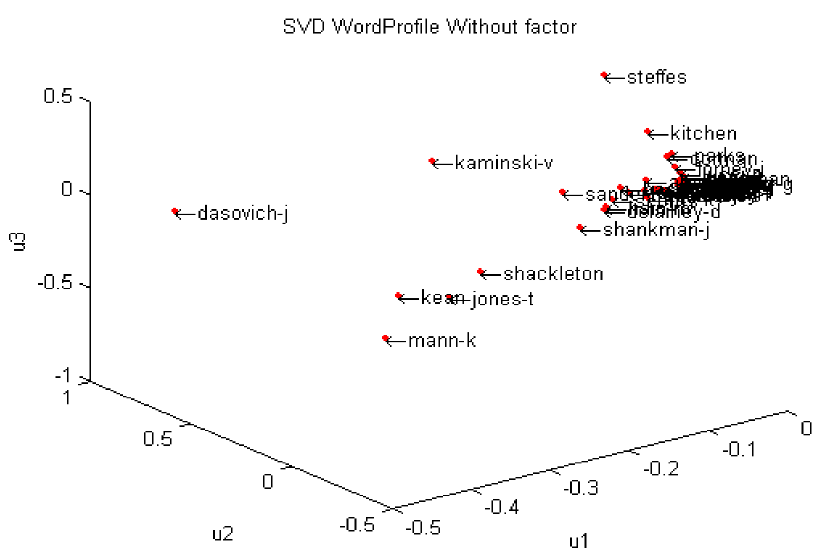

Fig.1. SVD of Word Frequency Matrix

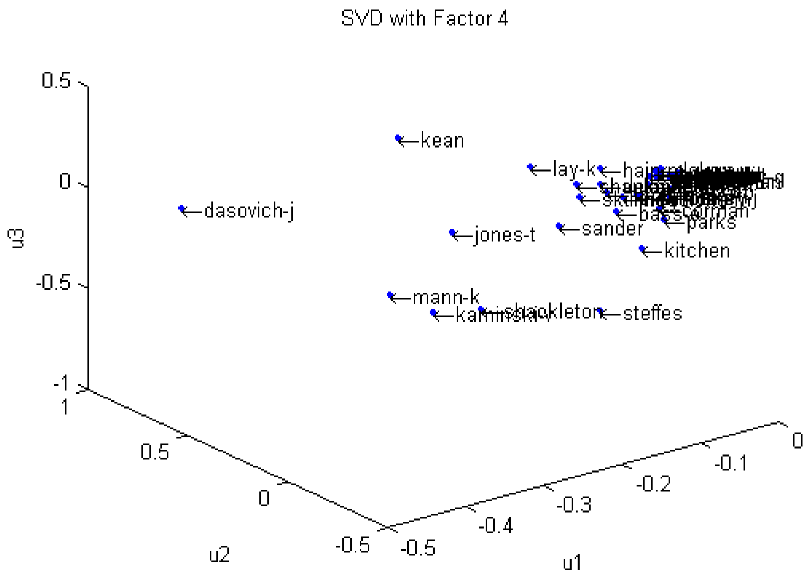

Fig.2. SVD of Word Frequency Matrix with factor 4

Comparing Figures 3a and 3b, we observe that after applying weights, the profiles of two executives Lay and Skilling move to a different cluster. The other individuals who used some of the same words also change their position to a different cluster. Fig. 4a shows that profiles with similar word patterns will be placed close together by NMF technique. On comparing Figures 4a and 4b we infer that, even after increasing the profile values of two executives by a factor of 4, there occurs additive process alone. Only the two executives i.e. Lay and Skilling undergo a change in their appearance while the other profiles remain unchanged. Thus Fig. 4a and Fig. 1 confirms the word usage pattern of the employees by resulting in similar graphs using two matrix decomposition techniques: NMF and SVD.

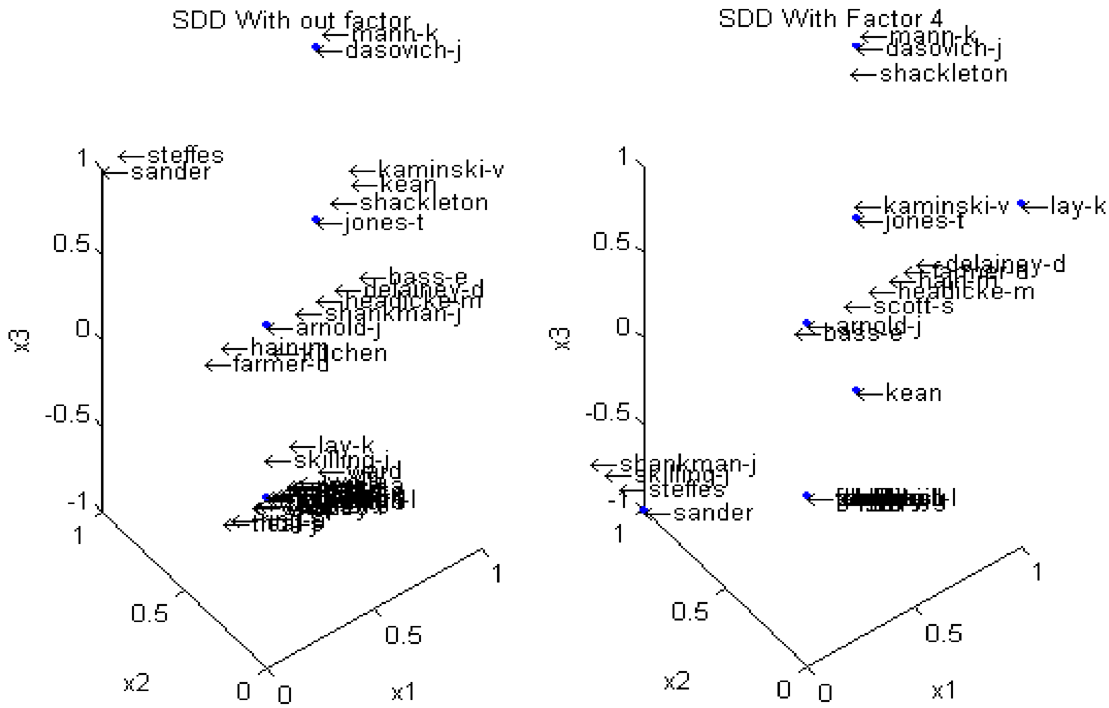

Fig.3a. SDD without factor on Word Frequency Matrix
Fig.3b. SDD with factor 4 Word Frequency Matrix

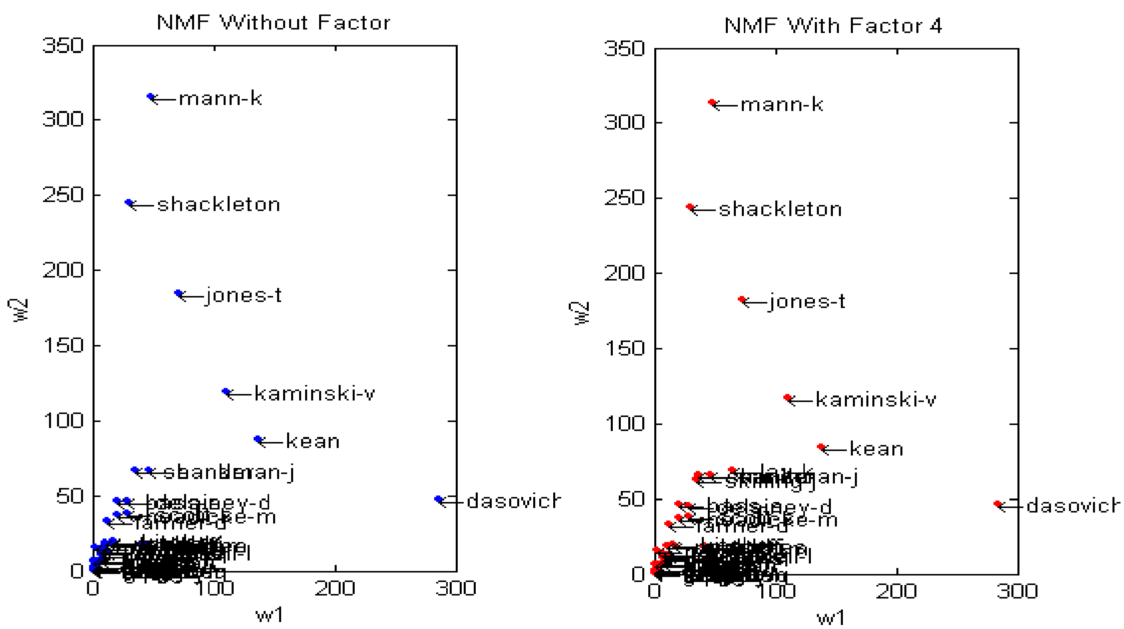

Fig.4a. NMF without factor on Word Frequency Matrix
Fig.4b.NMF with factor 4 on Word Frequency Matrix

In contrast to SVD, techniques based on SDD generate a hierarchical classification allowing the algorithm to predict at least the degree of anomaly of each point [17]. An effective approach to analyze complex datasets is to use SVD and

SDD together as Joint SVD SDD (JSS); Use SVD to position items and SDD to classify them. JSS provides a profound analysis of clusters. From the Figures 1, 3 only a partial knowledge on word usage was gained. On applying JSS, in Fig. 5a we perceive a clear positioning and classification of individuals. Fig. 5b shows the JSS of increased profile weights of two executives. There exists a change with two executives, along with the profiles which have similar word patterns. The profile which undergoes a change in position and classification means that they exchange mails with the two executives Lay and Skilling. Thus these employees form a communication link at the factor 4, which means to be a frequent communication. In the Fig. 6, SVD of the Email rank matrix shows individuals who use less “Enron speak words” (words frequently used within Enron organization). Initially we see that members of this committee produce emails that are short in using Enron speak words, and tend to use rare nouns. Further on continuing the same process, members who use the words of similar ranking will be visible. Thus we will get a communication link at iterations and a few numbers of the communication links are identified.

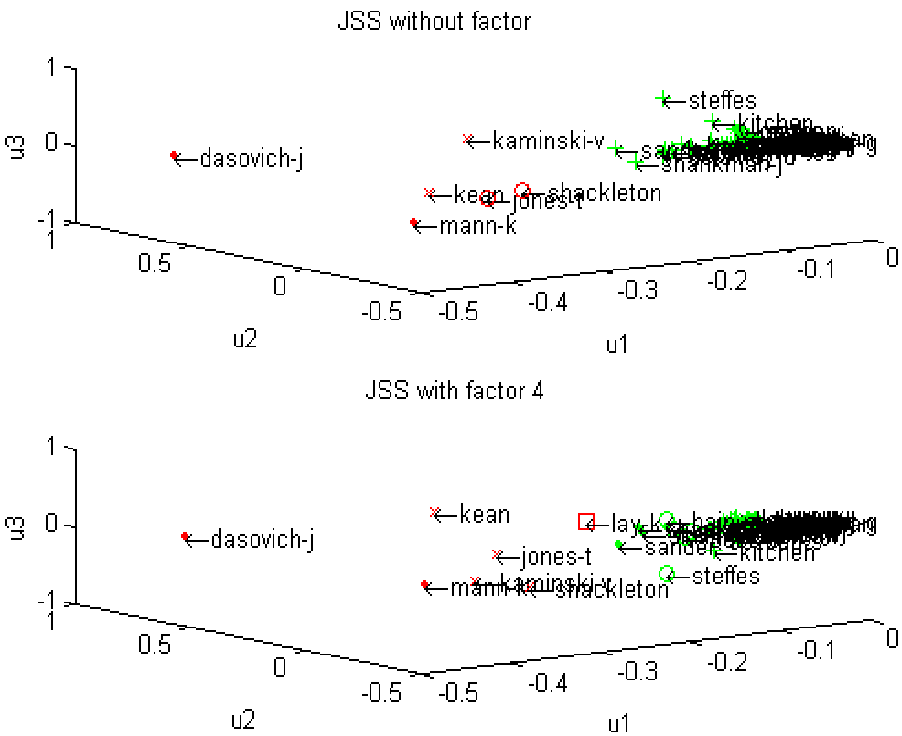

Fig.5a. JSS without factor on Word Frequency Matrix
Fig.5b. JSS with factor 4 Word Frequency Matrix

5.2.2 Categorizer

In this section, we discuss email categorization of employees with samples. Fig. 7 illustrates NMF on categorizing email matrix (unusual and official). Points at the extreme top left are unusual mails related to sports events, personal mails and points at the extreme right bottom are official mails related to legal contracts and document reviews. The points are labeled by the authors and their mail numbers. On analyzing the emails of every individual, their behavior can be determined. If their designation contradicts their behavior then the mails of those individuals are further scrutinized. In Fig. 8, mails of Bass (an employee of Enron) are categorized into official and unusual mails. From the graph we conclude that Bass mails are mostly related to unusual events, so either his designation must be related to sports or he must be misusing the resources at greater extent (However it was found that Bass was the coordinator of a fantasy basketball league within Enron). Similarly, every other communication link member’s profiles are categorized. Thus the general behaviors of the employees within the organizations are exemplified here.

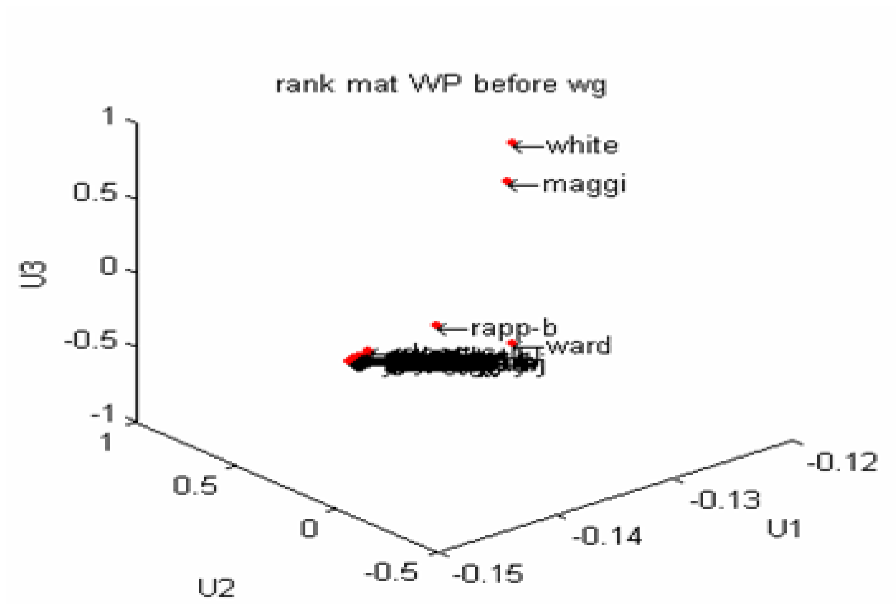

Fig.6. SVD of Email Rank Matrix

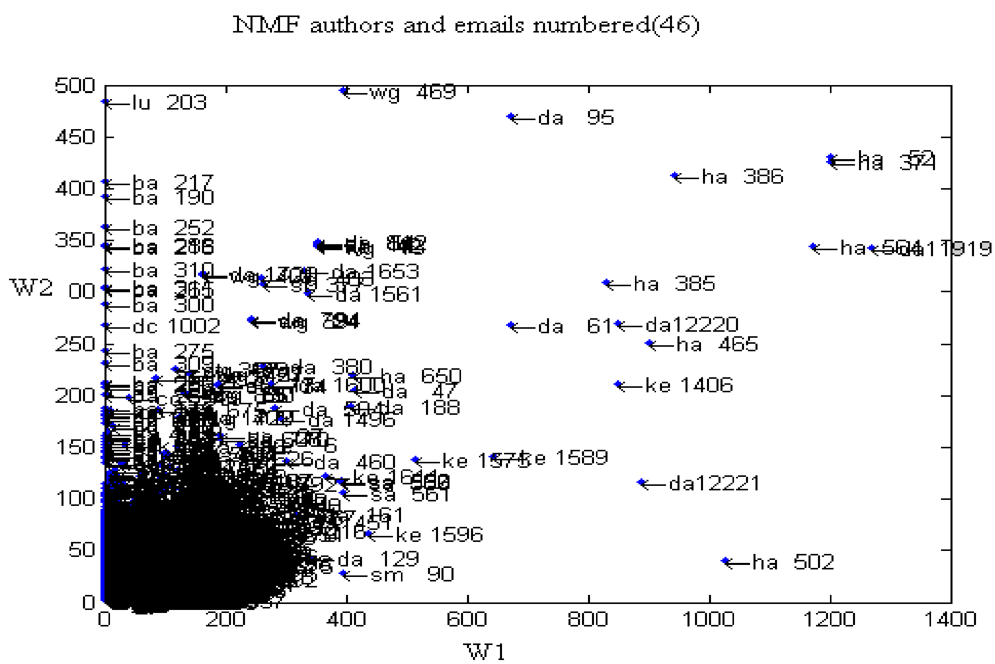

Fig.7. NMF on Categorizing Email Matrix labeled with author and their mail numbers

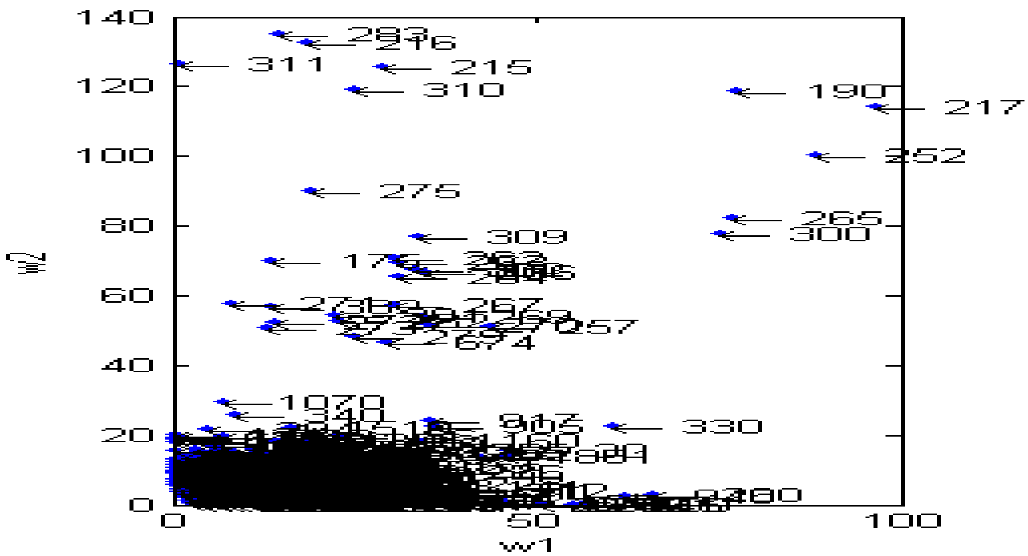

Fig.8. NMF on Categorizing mails of Bass labeling with his email numbers.

5.2.3 Deception Detector

The deception detector includes the analysis on both deception cue and deception model matrices. Initially, to trace out the deceptive employees SVD and JSS techniques are applied on the deception cue matrix. Fig. 9 shows SVD plot of the first 3 columns of U matrix with a row for each mail among 60551 mails and 4 columns representing 4 deceptive cues.

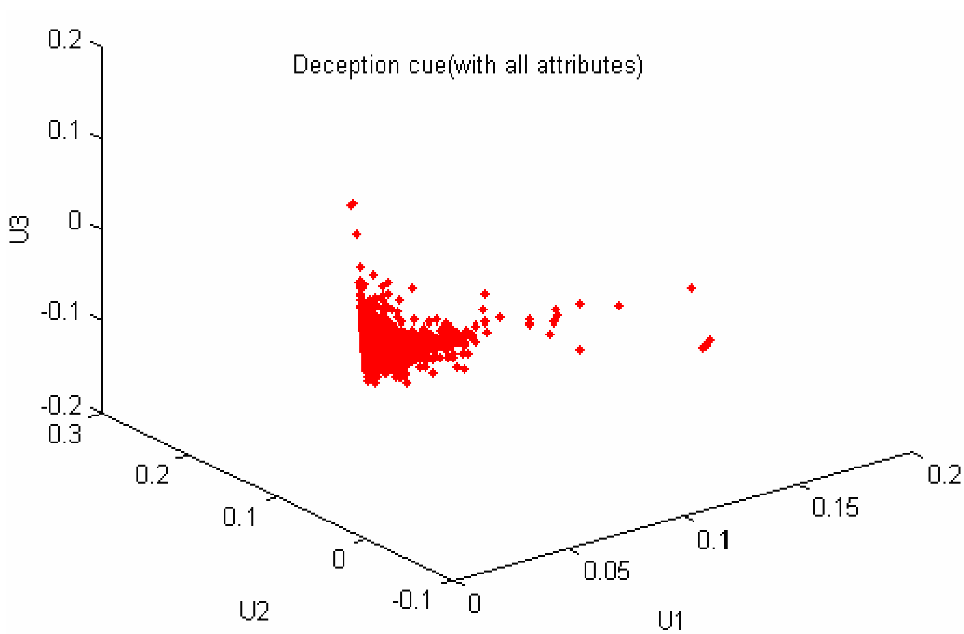

Fig.9. SVD of Deception Cues with all attributes

The extreme messages are of three distinct kinds, shown by the three points of the broadly triangular structure. The top-left corner of the plot corresponds to emails with a large number of exclusive words. The top-right corner corresponds to emails with many first-person pronouns. Points towards the bottom portion of the graph correspond to emails with many action-verbs. From the results of email-deception-cue matrix, there is no single cluster, or area of the graph, that exactly matches the model for deception we are using. Since our current representation makes use of a number of attributes that may exhibit a confusing or dilutive effect, causing a greater variety of messages to be considered deceptive, we refine the deceptive cues to include only those attributes that correctly isolate deceptive messages. Fig. 10 illustrates the V matrix with 84 deceptive attributes in the columns and their global frequencies in the dataset as the corresponding entries and the points in the graph are labeled according to which cue, an attribute represents as discussed in Table 1.

Table.1. Plot specification of JSS plot on V matrix

Plot Specification	Implies - Cues
magenta circle	First-person pronouns
blue diamond	Exclusive words
green square	Negative-emotion words
red triangle	Action words.

While the majority of the points are close to the origin, four points representing three first-person pronouns and an exclusive word are found farther from the origin. These points represent words with maximum global frequencies and are most useful in positioning an object. All the points representing action words and negative-emotion words appear close to the origin. These words occur less frequently in the email corpus than first-person pronouns and exclusive words, and hence SVD finds them less useful in helping to rank the emails.

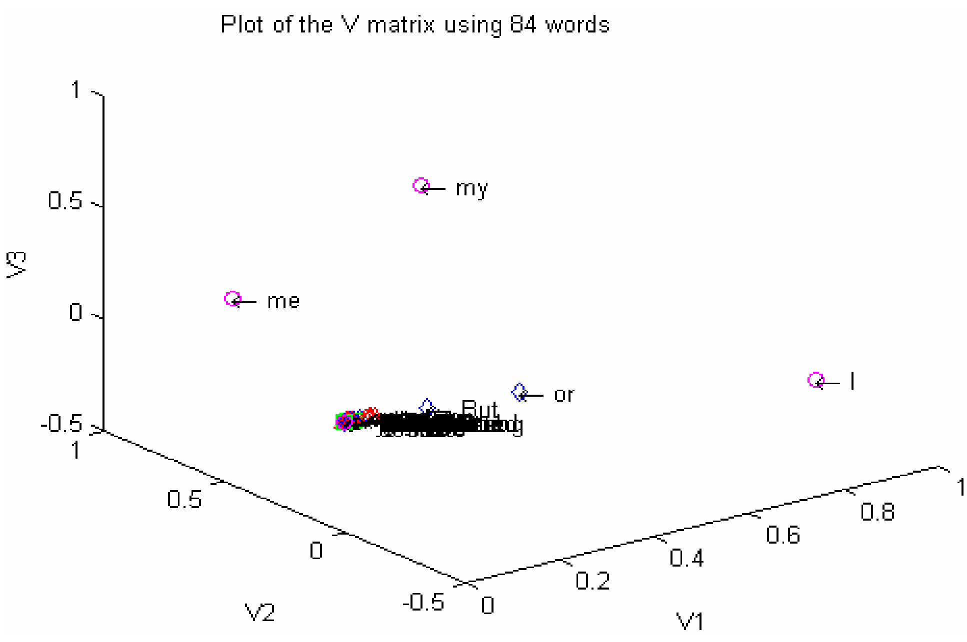

Fig.10. JSS of V matrix (whole dataset x 84 attributes) with all deceptive attributes

Due to the limited number of dimensions, the points appear close to the origin and a number of first-person pronouns and exclusive words are part of the cluster of points found close to the origin, the inclusion of these words may be diluting the model causing a greater variety of emails to be considered ‘deceptive’. By iteratively removing the points which are found away from the origin, and re-running JSS on this new dataset, a new set of points will be found which are away from the origin, and some of these points may represent negative-emotion or action words. Now a ranking of the words which are able to rank emails is created. Then, using only the highest ranked words for each cue, we will have a more discerning model of deception. Fig. 11, presents the graph of V matrix at the 4th iteration, which is obtained by removing the points that are found away from origin after iterations.

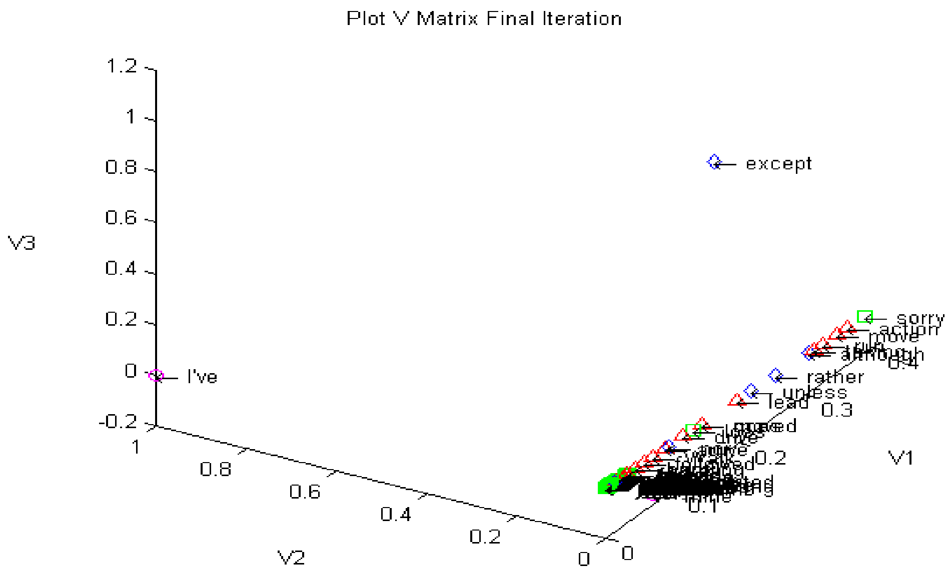

Fig.11. JSS of V matrix (whole dataset x 58 attributes) with all deceptive attributes

Fig. 12, presents, the SVD of the first 3 columns of U matrix with a row for each mail (60551) and 4 columns representing 4 deceptive cues with reduced number of attributes for each cue that are ranked highest. From Figures 9 and 12, we infer that the Fig. 12 shows a few clusters that make the graph quite sparse. Since we are looking for reduced frequency counts of first person pronouns, and exclusive words, we would expect deceptive messages to be found away from the top-right and top-left corner; to be near the origin. Messages with high action-verb frequency counts, which are one of the conditions for our deception model, are found primarily towards the bottom portion of the graph. Searching the emails that are found toward the origin and pointing towards the bottom, we found that there are messages that can be labeled deceptive in nature (like contract negotiations, employees discussing confidential information with one another) in this area of the plot, but there are many more messages on a variety of other topics. Fig. 13 shows the points labeled according to employee designation. This figure shows that the emails from Enron employees are farther from the origin than those originating from outside the organization. The observation is that there are many messages from Enron employees found close to the origin, and specifically, within the region we expect deceptive messages. The significance of this image is to highlight the distinction in writing style between Enron employees and all others. Since SVD is a numerical technique, increased magnitude of attribute values increases the importance of those values. Since we need low values of first person pronouns and exclusive words to be most significant, we alter their raw frequency counts to fix a pyramid.

Fig. 14 shows the SVD on deception model. Points in the lower part of these two clusters match the standard deception model most closely: emails having reduced frequencies of first-person pronouns, exclusive words, increased frequencies of negative-emotion words and action-verbs. Points in the higher part of these two clusters are similar, but do not exhibit increased frequency of negative-emotion words. Note that the external emails in this cluster originate from within Enron organization. Fig. 15 displays each point labeled according to the attributes having high values. This plot can be labeled with the employee and the email number. For clear representation we have avoided labeling the graph. Emails of a particular kind can now easily be selected from such a plot, based on their color coding and their distance from the origin. The corresponding emails can then be retrieved and analyzed by appropriate individuals. Thus our region of focus lies at the lower half of the clusters. This is inferred by referring to the plot specification given in Table 2. Even the employees who were spotted to exist within the communication link can be found in the deceptive area with their email numbers, and then these emails are manually opened to find the objective of that deceptive communication link.

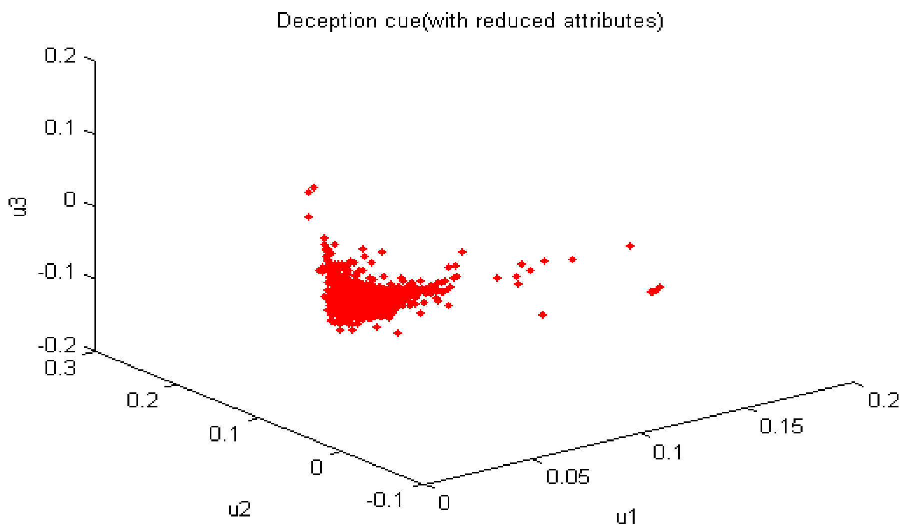

Fig.12. SVD of Deception Cues with reduced attributes

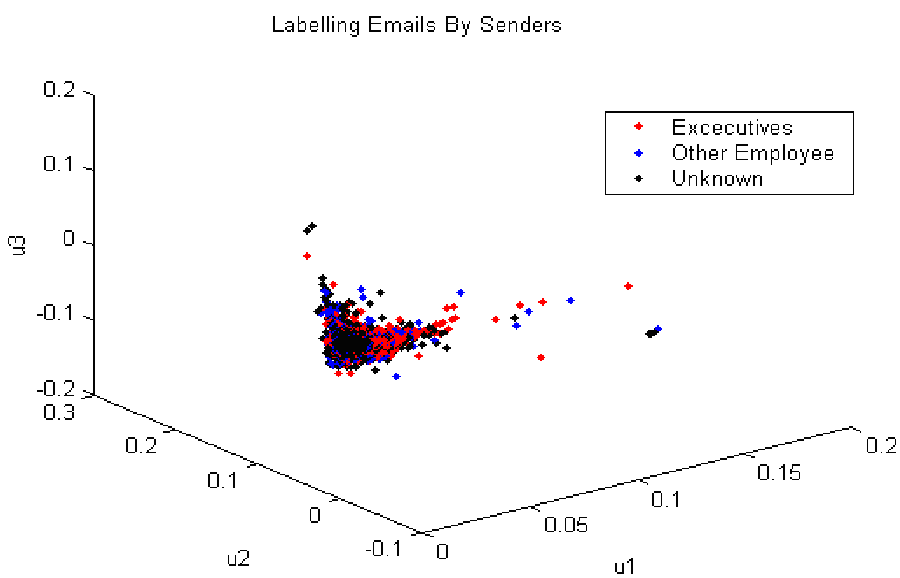

Fig.13. SVD of Labeling Emails by Senders

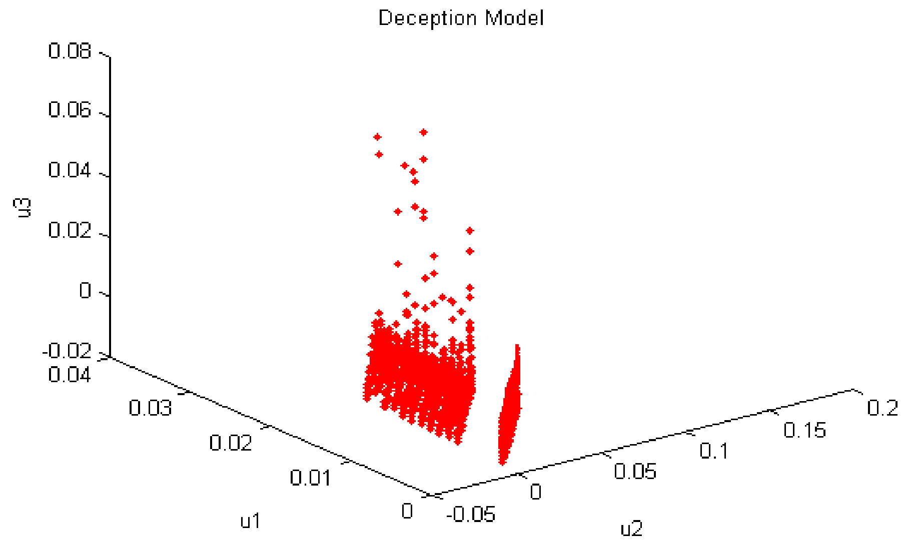

Fig.14. SVD on Deception Model

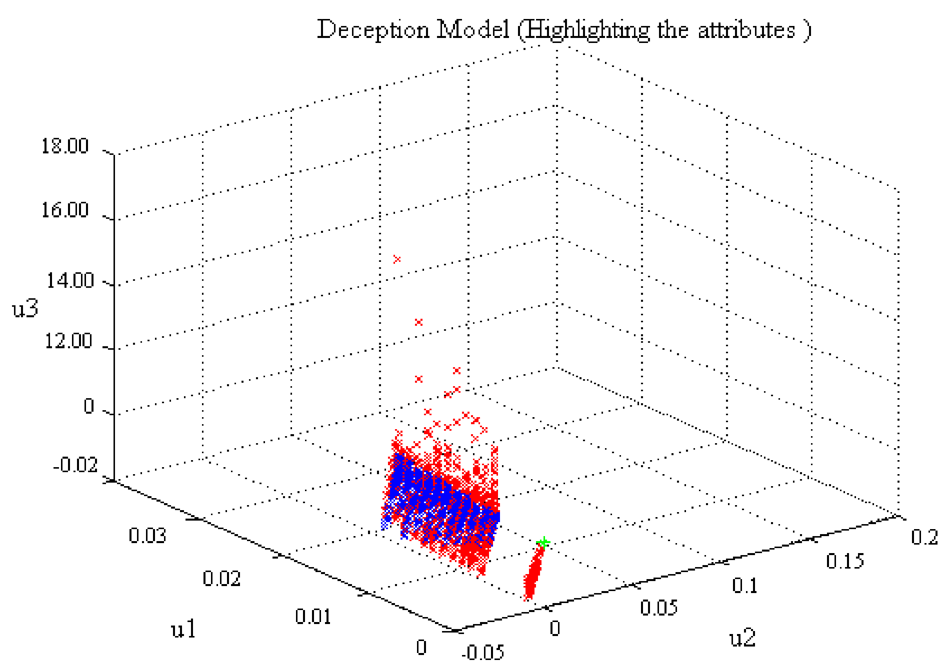

Fig.15. Deception model (Highlighting Attributes)

Finally, emails in the deceptive region are cross checked with the communication link from word usage profiles. If there is an association in the nature of the employees, they are checked using the categorizer. If the email numbers under the deceptive region are found in the official category, then those mails are extracted by the human reader. Similarly a suspicious link is highlighted by ensuring major number of employees in the communication link to be deceptive. Future work can concentrate on improving the scalability as discussed in [11], and exploring other dimensionality reduction methods that are discussed in [23, 25, 27].

Table.2. Plot specification of the deception model

Plot Specification	Implies
Red crosses	low negative emotion, High first person pronoun
Blue color	Moderate negative emotion, high first person pronoun and exclusive words
Green plus	Moderate negative emotion words and exclusive words, Low first person pronoun.
Red stars	High negative emotion, low first person and exclusive

6. Conclusion

Using word-frequency profiles, we found that word use is strongly correlated to an individual's role within the organization. Further applying weights to the commonly used words we extracted the structure of a group of individuals. Also we could establish the existence of a communication link between individuals who would not normally be expected to communicate with one another. The profiles of the interesting members are then categorized to decide their activities within the organization. It is used to isolate the suspected employees with their deceptive emails based on the frequency of the deceptive cues. On implementing various techniques (SVD, SDD, NMF, JSS), we conclude that JSS proffers the finest solution for resolving the employees word usage pattern and their distinguishable communication links and NMF offers better results in categorizing the profiles.

Acknowledgement
One of the authors Ch. Aswani Kumar, gratefully acknowledge the financial support from Department of Science and Technology, Govt. of India under the grant number SR/S3/EECE/25/2005.

References

[1]. Ch. Aswani Kumar, S. Srinivas, "Latent Semantic Analysis using Eigenvalue Analysis for Efficient Information Retrieval", *International Journal of Applied Mathematics and Computer Science*, Vol. 16, No. 4, 2006, pp.551-558.

[2]. M.W Berry, M. Browne, "Email Surveillance Using Non-negative Matrix Factorization", *Technical Report*, Department of Computer Science, University of Tennessee, Knoxville, TN 37996-3450.2006.

[3]. J.R. Carlson, J. F. George, J. K. Burgoon, M. Adkins, C. H. White, "Deception in Computer-Mediated Communication", *Group decision and negotiation*, Vol. 13, 2004, pp. 5-28.

[4]. Enron Email Dataset March 2, 2004. http://www.cs.cmu.edu/~enron/

[5]. P.S. Keila, D. B. Skillicorn, "Detecting Unusual and Deceptive Communication in Email", *Technical report*, School of Computing, Queen's University Kingston, Canada K7L3N6. 8 (2005a).

[6]. P.S. Keila, D.B. Skillicorn, "Structure in the Enron Email Dataset", *Computational and Mathematical organization theory*, Vol. 11, 2005b, pp. 183-199.

[7]. T.G. Kolda, D.P. O'Leary, "A semi-discrete matrix decomposition for latent semantic indexing in information retrieval", *ACM Transactions on Information Systems*, Vol. 16, 1998, pp. 322-346.

[8]. T. Korenius, J. Laurikkala, M. Juhola, "On principal component analysis, Cosine and Euclidean measures in information retrieval", *Information Sciences*, Vol. 177, 2007, pp. 4893 – 4905.

[9]. A.N. Langville, C. D. Meyer, R. Albright, "Initializations for the Nonnegative Matrix Factorization", *Proceedings of 12th ACM SIGKDD Int. Conf Knowledge Discovery and Data Mining* 2006.

[10]. D. Lee, H. Seung, “Algorithms for Non-Negative Matrix Factorization”, *Advances in Neural Information Processing Systems*, Vol. 13, 2001, pp. 556–562.

[11]. D.I Martin, J. C. Martin, M. W. Berry, M. Browne, “Out-of-core SVD performance for document indexing, Applied numerical mathematics”, Vol. 57, 2007, pp. 1230-1239.

[12]. A. McCallum, X. Wang, A.C. Emmanuel, “The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email”, *Journal of artificial intelligence research*, Vol. 30, 2007, pp. 249-272.

[13]. S. McConnell, D.B. Skillicorn, “Semi Discrete Decomposition: A Bump Hunting Technique”, *Proc. 4th Australasian Data mining workshop*, Canberra, 2002, pp. 75-82.

[14]. M.L. Newman, J. W. Pennebaker, D. S. Berry, J. M. Richards, “Lying Words: Predicting Deception From Linguistic Styles”, *Personality and Social Psychology Bulletin*, Vol. 29, No. 5, 2003, pp. 665-675.

[15]. P. Perrin, F.E Petry, “Extraction and representation of context information for knowledge discovery in text”, *Information Sciences*, Vol. 151, 2001, pp. 125-152.

[16]. R. Rajaram, S.Balamurugan, “A Novel Data Mining Approach to Detect Deceptive Communications in Email Text”, *Journal of Computing and Information Technology*, Vol. 15, 2007, pp. 161-169.

[17]. S. Rawat, A. K. Pujari, V. P. Gulati, “On the Use of Singular Value Decomposition for a Fast Intrusion Detection System”, *Electronic Notes in Theoretical Computer Science*, Vol. 142, 2006, pp. 215-228.

[18]. D.H. Shih, H. S. Chiang, C. D. Yen, “Classification methods in the detection of new malicious mails”, *Information Sciences*, Vol. 172, 2005, pp. 241-261.

[19]. D.B. Skillicorn “Applying Matrix Decompositions to Counterterrorism”, *Technical report*, School of Computing ,Queen's University, Kingston, Canada. External Technical Report ISSN-0836-0227-2004-484, 2004.

[20]. J. Tang, H. Li, Y. Cao, Z. Tang, “Email Data Cleaning”, *Proceedings of 11th ACM SIGKDD Int. Conf. Knowledge discovery in data mining*, 2005, pp. 489-498.

[21]. Word List, URL:http://www.giwersworld.org/computers/linux/common-words-req.phtml and http://ogden.basic-english.org/intlword.html.

[22]. L. Zhou, D. P. Twitchell, T. Qin, J. K. Burgoon, J.F. Nunamaker, “An Exploratory Study into Deception Detection in Text-based Computer-Mediated Communication”, *Proceedings of 36th Hawaii Int. Conf. System Sciences*, 2003, pp. 44.2.

[23]. Ch. Aswani Kumar, S. Srinivas, “Concept lattice reduction using fuzzy k means clustering”, *Expert systems with applications*, Vol. 37, 2010, pp. 2696-2704.

[24]. Ch. Aswani Kumar, S. Srinivas, “Mining associations in health care data using formal concept analysis and singular value decomposition”, *Journal of biological systems,* Vol. 18, No. 4, 2010, pp. 787-807.

[25]. Ch. Aswani Kumar, S. Srinivas, “A note on weighted fuzzy k-means clustering for concept decomposition”, *Cybernetics and systems,* Vol. 41, 2010, pp. 455-467.

[26]. Ch. Aswani Kumar, R. Palanisamy, “Comparison of matrix dimensionality reduction methods in uncovering latent structures in the data”, *Journal of information and knowledge management,* Vol. 9, No. 1, 2010, 81-92.

[27]. Ch. Aswani Kumar, “Analysis of unsupervised dimensionality reduction techniques”, *Computer science and information systems*, Vol. 6, No.2, 2009, 217-227.

BIOGRAPHY

R. Divya has received her B.Tech degree in Information Technology from VIT University, Vellore, India.

Ch. Aswani Kumar is Associate Professor, School of Information Technology and Engineering, VIT University, India. He holds a PhD degree from VIT University and Bachelors, Masters Degrees in Computer Science from Nagarjuna University, India. His current research interests are Formal Concept Analysis and Machine Intelligence. He has published 32 refereed research papers so far in various journals and conferences. He is associated with various professional bodies including ACM, CSI. He is reviewer and editorial board member for many international journals and conferences.

S. Saijanani has received her B.Tech degree in Information Technology from VIT University, Vellore, India.

M. Priyadharshini has received her B.Tech degree in Information Technology from VIT University, Vellore, India.