Modelling Deception Detection in Text

by

Smita Gupta

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
November 2007

Your file Votre référence
ISBN: 978-0-494-36986-9
Our file Notre référence
ISBN: 978-0-494-36986-9

Abstract

As organizations and government agencies work diligently to detect financial irregularities, malfeasance, fraud and criminal activities through intercepted communication, there is an increasing interest in devising an automated model/tool for deception detection. We build on Pennebaker's empirical model which suggests that deception in text leaves a linguistic signature characterised by changes in frequency of four categories of words: first-person pronouns, exclusive words, negative emotion words, and action words. By applying the model to the Enron email dataset and using an unsupervised matrix-decomposition technique, we explore the differential use of these cue-words/categories in deception detection. Instead of focusing on the predictive power of the individual cue-words, we construct a descriptive model which helps us to understand the multivariate profile of deception based on several linguistic dimensions and highlights the qualitative differences between deceptive and truthful communication. This descriptive model can not only help detect unusual and deceptive communication, but also possibly rank messages along a scale of relative deceptiveness (for instance from strategic negotiation and spin to deception and blatant lying). The model is unintrusive, requires minimal human intervention and, by following the defined pre-processing steps it may be applied to new datasets from different domains.

Acknowledgments

Many thanks:

To my supervisor and 'guru', Dr. David Skillicorn, for the initiation he has provided me into the art of taking an academic idea from its seed to implementation and finally presentation at international conferences.

To my teachers and professors: from whom I have learned many valuable things.

To Queens University: for the chance to pursue my graduate studies here – for the scholarships and full-funding and for rendering open to me its vast learning resources.

To my competent colleagues: for their encouragement and support.

To my friends: for being plain wonderful. This journey would neither be interesting nor so rewarding without them.

To my family: for their love and for sharing with me the wisdom that the final goal is to be a better human-being with every passing day.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter we describe the motivation behind this research effort, the objectives that we set for our project and the overall contributions of this work. The chapter ends by laying out the organization of this thesis documentation.

1.1 Motivation

There is an increasing interest in deception detection in the context of intelligence services, law enforcement and intra-organizational monitoring. Though it is expected that human capability to indulge in deception would be balanced by the ability to detect deception, it has been borne out in studies [31, 37, 43] that humans are not naturally good in deception detection. In fact, even trained personnel correctly identify deception at rates only little better than chance (52 percent accuracy) [37].

In contrast, software systems based on number-crunching algorithms and correlation analysis often provide surprisingly accurate results. This is especially true in real-time applications where computerized systems can quickly and efficiently parse huge datasets, flagging objects of interest that can then be sent for final analysis and scrutiny by human judges.

Psychological studies [28, 29, 43] in the area of deception detection have shown that changes in behavior such as the body posture, facial expressions, speech rhythm and pitch are closely correlated with deception. Markers of deception that are under conscious control (like the content of the deceptive 'story') may be modified by those who wish to deceive, but fortunately most of these cues are a result of subconscious processes and therefore even awareness of vigilance and scrutiny does not make deception any easier.

In the context of deception in text, though communication is stripped to its essentials and no non-verbal deception-cues are transmitted, the linguistic manifestations of deception remain consistent across most domains. It has been empirically shown [28, 37, 45] that deception leaves a linguistic signature caused largely due to the high demands that indulging in deception generates on a person's cognitive capabilities.

Many research groups in the field of psychology [26, 44, 45] have constructed sets of linguistic markers for deception detection in text using different methodologies. James Pennebaker *et al.* at the Department of Psychology, University of Texas [37, 38] have constructed an empirical deception model based on cue-word usage-frequency profiles. Though this linguistic signature of deception is not easy for humans to detect directly, it is easily detected by software. According to the model, deception in text is marked by:

- Decreased frequency of first-person pronouns (I, mine, myself *etc.*) – a subconscious attempt by the author to disassociate from the deceptive content;

- Decreased frequency of exclusive words (or, but, without *etc.*) – to keep the content simple, concrete and without abstractions in order to avoid faltering while being repeatedly interrogated;

- Increased frequency of negative emotion words (anger, abandon, hate *etc.*) – a reflection of the subconscious feeling of guilt involved with being deceptive;

- Increased frequency of action verbs (move, run, lead *etc.)* – as a form of distraction to keep the 'story' moving while the basic content remains simple and insignificant.

Being based on word-usage patterns and word-counting strategies makes the Pennebaker empirical deception model available for use in computerized text analysis and for being implemented and tested in an automated environment.

As email and rich media become ubiquitous in everyday life as a means of communication, there is an accompanying interest in the possibility of using this communication for deception detection. Also, with organizations and government agencies investing heavily in monitoring systems to detect financial irregularities, malfeasance and criminal activities, there arises a pertinent need to devise an automated, unobtrusive and cost-effective model for deception detection in text.

1.2 Objective

Our objective is to build on the Pennebaker empirical deception model to devise an automated textual deception-detection model. To begin with, we formulate a normalization process so that the Pennebaker model can be applied to new datasets from different domains.

Certainly, from the four categories described in the empirical model (first-person pronouns, exclusive words, negative emotion words and action words) not all the cue-words or even categories make an equal contribution to the distinction of the model. We work on refining the characterization of the Pennebaker model by studying the relationships between the various deception cue-words/cue-word-categories that contribute to the model. We aim to understand the differential use of the cue-words in deception detection by using unsupervised data mining [7, 24].

By applying the Pennebaker deception model to the Enron email dataset [9, 10, 11] and using a matrix-decomposition technique [21], we attempt to construct a descriptive textual deception-detection model that can help us to highlight how deceptive communication is qualitatively different from truthful communication. We aim to use this automated model not only to capture messages that are deceptive in the conventional sense, but also those where the author attempts to spin the content and words to convey something that they themselves do not quite believe. Capturing such messages, where there seems a dichotomy between the overt meaning of what is being communicated and the mindset of the sender, is important to us as we try to rank (if possible) messages along a scale of relative deceptiveness (say from strategic negotiation to spin to deception to blatant lying).

1.3 Contributions

The contribution of this research is that, by using unsupervised data mining, we have been able to refine the empirical deception detection model provided by Pennebaker and devise an automated descriptive model for deception detection in text.

Using a matrix-decomposition technique [21] has helped us to weigh each of the deceptive-cue attributes to best understand their differential use in deception detection. We have shown that each of the four categories of cue-words in Pennebaker's deception model behaves as a distinct latent factor. Of these the two classes, first-person pronouns and exclusive words, are much more important than the others.

The descriptive deception detection model devised highlights the multivariate profile of deception based on the usage-frequencies of multiple cue-words. Also, being based on word-usage patterns and word-counting strategies makes the automated model unintrusive and cost-effective, requiring minimal human-intervention.

We have investigated the structure of the model when it is applied to a large corpus of organizational email. Additionally, we have been able to formulate a normalization process that can successfully over-ride the external effect of the message/document text-length on its deception-cue-word usage pattern; neutralize the inherent disparities in the usage-frequencies of the deception-cue-words (since some words like 'I', 'walk' are much more common than words like 'wicked', 'lead'); and is able to capture the significant features of a dataset-matrix regardless of how sparse it is. This normalization method makes it possible to apply the textual deception detection model to new datasets.

1.4 Organization of Thesis

We proceed in Chapter 2 by explaining the concepts and techniques needed as background material to understand this research work. In Chapter 3, we detail our research methodology and experiments – describing at length the steps involved at the data pre-processing stage followed by the implementation and refinement of the deception detection model. Chapter 4 describes the results that we have obtained and highlights the accomplishments of our work. The chapter ends with a discussion based on the results achieved. Chapter 5 provides the conclusions that can be drawn from this research and some of its current limitations.

Chapter 2

Background

This chapter explains the concepts and techniques needed as background material to understand our research work – modelling deception detection in text. Deception theory suggests that deception leaves a linguistic signature. To build on empirical studies and analysis of this linguistic signature, we use data mining (in specific, the singular value matrix-decomposition technique) to explore the Enron email dataset. The final goal is to devise an effective, inexpensive and unintrusive method for textual deception detection.

2.1 Deception and Spin

Deception, as defined by the Oxford Dictionary [17] is "the act of deliberately making somebody believe something that is not true (equivalent of deceiving them); a trick intended to make somebody believe something that is not true."

Through this research work we are interested in detecting textual deception. We attempt to detect deception not only in the usual sense as defined above, but also in the wider sense of being able to capture 'spin'. Spin, we believe, encompasses more shades of lying/deceit than deception – it is the attempt by authors to convey something they themselves do not (quite) believe [36]. This refers to speech or text where the apparent meaning is not the true belief of the person saying or writing it. For instance, a politician saying or projecting things as (s)he may think is more acceptable to the target audience at a current time is an example of spin.

Spin may not necessarily entail lying, but rather, the act of cautiously and cleverly manipulating words to achieve an ulterior motive. This is not dissimilar to the act of negotiation as a means of gaining an end; but is more complex. It might not be unfair to say that even deception has its spectrum of shades – ranging from blatant lying; to spin; to finally the art of negotiation, where the deception is more strategic and socially acceptable.

2.1.1 Linguistic Signature of Deception

Deception theory suggests that deception leaves a linguistic signature. It has been empirically qualified by psychological studies [28, 37, 43] that deceptive communication is qualitatively different than truthful communication. This is largely because deception is more cognitively demanding. Forming and telling a story about something that never happened or happened very differently taxes a person's cognitive capabilities. Also, since language generation is a subconscious process it is heavily affected by the emotional states associated with deceiving and the cognitive demands of deception cause performance deficits in other areas [37]. In fact, the awareness of monitoring or scrutiny only tends to further strengthen this subconscious feedback. Omitting and/or substituting words to avoid surveillance may lead to 'excessive blandness' [41] which in itself becomes a substitution signature [32, 33]; making it almost impossible to override the characteristic signature of deception.

To build on this, federal law enforcement officers in the United States of America (USA) while being trained in interviewing techniques [23] are encouraged to keep in mind four fundamental points about human nature – people like to talk; people do not want to lie; people mean exactly what they say; and people's words will betray them if someone is there to listen. It has been noted that often people tell us one thing, but we hear something entirely different. This is due to our tendency to interpret what is being said. It is important that we do not interpret anything that is being said, but only listen objectively, because people mean exactly what they are saying. By paying attention to each word used in a statement, we are able to recognize better what a person is saying and gain additional information from the statement or subject.

To take an example [23], suppose someone were asked the question: *"Did you have a gun when you entered the house?"* and he answers: *"I never had a gun"*. We can immediately interpret that, if a person never had a gun, then they obviously could not have had a gun when they entered the house. However, this is merely an interpretation. The truth is that the person has not even answered the question. A straightforward affirmative or negative answer to the question has been avoided by making a generic statement. The ambiguity in the answer provided is a part of human intrinsic psychological makeup that allows us to almost effortlessly engage in manipulation and deception, without any apparent pangs of conscience or embarrassment.

Studies in deception detection [27, 28] have revealed that many changes in a person's body posture and behavior are closely correlated with deception. Non-verbal cues like pupil dilation, repeated blinking, shifting eye contact, facial expressions and pitch are often considered the best predictors of deception. However, depending on non-verbal cues for deception detection has some limitations. It requires that the person of interest be physically present. Also, to effectively interpret non-verbal cues for deception it is critical to initially establish the normal body movements of the interviewee. Establishing 'what is normal' is key to reading the red-signal non-verbal cues. Lastly, as one concentrates on non-verbal cues there is a possibility of missing-out/mis-interpreting what is being said. Thus, though non-verbal cues should be used, whenever possible, they are sometimes not the most effective mode for deception detection.

Everyone has lied and everyone has sometimes been lied to. This fact not only generates a lot of interest in deception detection, but also makes it an important and practical problem.

2.2 Deception Detection

Earlier deception detection was mainly explored in a forensic context. Now, there is an increasing interest in this field in the context of intelligence, law-enforcement and even organizational management.

Legislation like the U.S. Sarbanes-Oxley act [20] mandates that organizations enforce proper management control structure, procedures to monitor financial reporting and be responsible for any financial irregularities. Additionally, organizations may also want to be aware of employee activity, in order to reduce illegitimate usage of company resources, detect harassment, fraud, criminal issues and industrial espionage by insiders.

Many countries already have in place intelligence frameworks that intercept and analyze communication messages. The largest of these is the Echelon [1]. Run jointly by Canada, USA, UK, Australia and New Zealand, this intelligence system captures all types of communication messages such as radio, satellite, telephone, fax, emails, *etc.* All messages are checked for keywords; those that do not use the predefined keywords are discarded and the ones that do are passed on for further human analysis.

For our research we intend to explore, specifically, email communication. Email is now widely accepted as a primary means of communication within and without most organizations. It has become almost ubiquitous because of its ease of use, fast delivery and cost effectiveness. Surprisingly, over the space of the last decade, when email has gained much popularity, white-collar crime in North America has also increased both in frequency and magnitude [25]. Certainly, there is no direct correlation between these two phenomena, but this does highlight the possibility of using email as a tool in combating some forms of white-collar crime.

Email text is much like spoken language – characterized by an informal style and sentence fragments produced in real-time. Email text, linguistically, lies somewhere in between spoken language and written communication. It is true, that unlike spoken language, one does have the option to edit an email, but empirical evidence suggests that this rarely happens [26, 27]. Though some markers of deception that are under conscious control may be altered, most are subconscious and even knowledge of these give-way traits and/or the existence of monitoring systems does not make deception any easier.

2.3 Pennebaker Deception Model

Our research builds on the empirical deception model suggested by James Pennebaker *et al.* at the Department of Psychology, University of Texas and implemented in the Linguistic Inquiry and Word Count (LIWC) Program [37, 38]. According to this empirical model, deceptive writing is characterized by reduced frequency of first-person pronouns and exclusive words, and increased frequency of negative-emotion words and action-verbs. Not using personal pronouns (I, me, my, *etc.*) is seen as an attempt towards dissociating oneself from the ownership of a statement. The decreased usage of exclusive words (but, except, without, *etc.*) is to keep the story cognitively simple, concrete and without abstractions – easy to remember and easy to repeat. Negative emotion words (hate, anger, greed, *etc.*) reflect the inherent feeling of guilt associated with deceiving. The increased usage of action words (go, carry, run, *etc.*) keeps the story moving without revealing much; almost as if it were a plaintive narration of events or a 'to-do list'.

Psychological studies in fraud detection [26, 43] have revealed that it is often not the content of what people say, but how they say it (the little words they use and the usage frequencies of these words) that unfolds the true feelings and mental state of the person. Liars can, to some degree, control the content of their stories, but not their style of language usage.

Certainly, it is not difficult to understand why liars feel more anxious, nervous and guilty (use of negative emotions); try to avoid taking responsibility (reduced use of first-person pronouns); and find it difficult to differentiate between what happened and what did not (reduced use of exclusive words). Interestingly, truthful communication on the other hand is generally far less emotionally negative and uses more self-references and exclusive words [8, 16].

The Pennebaker deception model has been significantly successful in deception detection, with an average accuracy rate of 67 percent as opposed to 52 percent (almost equivalent to chance) by trained human judges [37].

The LIWC model was further put to test when Gary Bond [16], a psychology doctoral student at New Mexico State University, used the model to analyse the transcribed speeches of inmates in six prisons in the states of New Mexico, Kansas and Mississippi. The felons were asked to lie or tell the truth about a video they had just watched. Bond and his colleagues were able to replicate the accuracy rate of the LIWC in the field (in fact, they did slightly better then 67 percent accuracy rate).

The *Pennebaker Deception Score* for a document can be calculated by first counting the usage-frequencies of the individual cue-words (first-person pronouns/exclusive words/negative-emotion words/action-verbs) in the model. Then the frequency counts for the first-person pronouns and exclusive words are adjusted (multiplying by -1) so that, as required, low frequencies of these cue-words are captured as high frequencies and vice versa. Finally, the sum of the cue-words' frequencies forms a single score. Note that the Pennebaker score weighs each attribute in the cue-list equally. It is important to remember, however, that each cue-word might not be equally significant (we explore this further in our work).

We use a subset of words from the Pennebaker deception model [35, 38]. Being based on frequency-count profiles of cue-words rather than content and semantics makes the model available to be implemented and tested in an automated environment.

2.4 Related Work

Previous work in the field of automated textual deception-detection has been done primarily in supervised environments [38, 44, 45]. The aim in these studies was to compare documents, where it was already known that the subject was lying, with those where the subject was telling the truth, and determine the differences in their word usage. Since these studies were conducted in controlled environment(s), they are unable to provide a true representation of real-time communication and facilitated a very narrow, tunnel-vision predictive approach instead of an inductive and descriptive approach to understanding deception and its detection.

As far as studying deception in an unsupervised environment [7, 24] is concerned, Parambir Keila [35] suggested the use of an automated textual deception detection model as a tool to capture malfeasance and organizational dysfunction. The model suggested by Keila, required refinement and he outlined two basic approaches to improve its deception-detection ability. The first approach suggested successively excluding deception cue-words that are found to be least useful in classifying emails as deceptive. These would generally be words that, through correlation analysis, are found to be too confounded with other words and thus serve a minimal individual purpose. The objective is to reduce the total number of words/attributes that serve as pointers for deception. The second approach, on the other hand, aims to increase the count of words that indicate deceptiveness. This method would involve a human reader who would manually parse emails that have been categorized as deceptive and intelligently pick words that appear in maximum number of deceptive emails and/or are in themselves obvious indicators of deception (*eg.* “believe me”, “no doubt”).

2.5 Technique

Data-mining algorithms help in the non-trivial extraction of implicit, previously unknown information from voluminous datasets and databases. Their goal is to find unsuspected patterns, relationships and "summarize the data in novel ways that are both understandable and useful" [34]. Data mining serves as a second mode of science – a method of theory validation for systems where controlled experiments cannot be or are difficult to use [42].

Singular Value Decomposition(SVD) an unsupervised data mining technique [21, 22, 40], is particularly effective in cases where each entry in the dataset arises as a result of a number of real-world processes. The decomposition makes the analysis easier by revealing the underlying structure in the data, and by systematically reducing the size of the dataset to highlight hidden attributes and their correlations.

The SVD of a matrix A with n rows and m columns is described as

$$A = USV'$$

where U is a $n \times m$ matrix; S is a diagonal matrix with non-increasing entries (the singular values) which indicate the importance of each dimension; V is a $m \times m$ matrix and the superscript dash indicates transposition. Additionally, both U and V are orthogonal matrices such that $UU' = I$ and $VV' = I$.

There exist several interpretations of the SVD decomposition. According to the *factor interpretation*, the rows of V highlight the hidden/latent factors in a dataset and therefore provide a better representation of the data. Here, how one must mix these latent factors to get the values observed in A is described by the rows of U.

The *layer interpretation* views matrix A as the sum of k outer product matrices;

such that each outer product matrix A_i is the product of the i^{th} column of U, the i^{th} row of V and the i^{th} diagonal element of S.

Though all interpretations can be used to understand a dataset, the factor interpretation is, generally, used in the Social Sciences and the geometric interpretation is extensively employed in the Sciences.

According to the *geometric interpretation*, the m rows of V are considered as axes in a transformed space and the rows of matrix U are seen as coordinates in this m-dimensional space. While the rows of matrix A do have a natural interpretation in the m-dimensional space, the transformed space provides the most faithful representation of the underlying relationships in these m-dimensions – such that as much variation as is possible is captured along the 1^{st} axis, as much as of what remains along the 2^{nd} axis and so on. Thus with the information that the maximum structure is represented in say any suitable 'd', the decomposition's right-hand-side matrices can be truncated at these 'd' dimensions. For instance, choosing 'd' $=$ 3 means being able to study the maximum amount of structure in 3 dimensions.

Importantly, in the Singular Value Decomposition (SVD), the columns of matrix V represent orthogonal and independent factors. Also, the correlation between objects (rows) is proportional to the dot product (cosine of the angle) of their position as vectors from the origin. Uncorrelated objects have a dot product close to zero. A large and positive dot product indicates that the two objects are highly correlated, whereas a large and negative dot product indicates negative correlation. Note that there exist two ways for the dot product between vectors to be close to zero. The vectors can be placed orthogonal to each other. However, since typically m is less than n, there exist fewer directions in which vectors can point orthogonally than there are vectors. As an alternative therefore, the uncorrelated vectors must have small dot products (can not be orthogonal) and the corresponding objects are placed close to the origin in the transformed space.

In the SVD transformed space, distance from the origin indicates interesting correlation with the other objects. Objects that are correlated either with nothing or everything (uninteresting objects), because of the constraint that they must have small dot products with almost all the other objects, are placed close to the origin. In contrast, objects that are correlated with only few other objects (interesting objects) are located far from the origin – generally there exist different clusters for these objects in different directions.

Another important and useful property of SVD is that multiplying row(s) or columns(s) by a scalar effectively changes their influence on the entire decomposition due to the 'pulling' effect of the up-weighted points. Multiplying by a scalar value greater than one not only moves the corresponding up-weighted points away from the origin but, as a side effect, pulls the points correlated to them away from the origin as well – thus, highlighting their correlation with the up-weighted points.

Additionally, SVD is completely symmetric

$$A' = VSU'$$

This means that, while using SVD, the same properties that apply to objects also apply to attributes. Therefore, the SVD technique can be used to visualize and understand relationships among the objects; among the attributes; and among the objects and the attributes.

2.6 Choosing a Dataset

Counting the cue-word frequencies in the four categories (first-person pronouns, exclusive words, negative-emotion words, action words) the deception model can be applied to a single document, but it is difficult to see increased or decreased frequency patterns in it. Thus, it is important to consider a set of similar messages in order to establish norms for the set; rather than working based on a hypothetical threshold. The deceptive messages become more visible in contrast to the background provided by the ordinary messages. Also, it is difficult to manipulate deceptive messages to make them look ordinary, for there are a range of ordinary messages that they are likely to be compared to. Altering frequency patterns in any possible way would lead to 'unusual substitution' ratios that will in themselves appear inappropriate [32, 33].

For our experiments we decided to use the Enron email corpus [9, 10, 11], which comprises of messages received and sent by Enron employees over a period of three and a half years just before the company's collapse in 2001.

2.6.1 Enron

Named by Fortune as "America's Most Innovative Company" for six consecutive years from 1996 to 2001, Enron [9, 10] at its peak was one of the world's leading electricity, natural gas, pulp and paper, and communications companies, with claimed revenues of USD 111 billion in 2000 [9]. After its bankruptcy filing, however, Enron is best known only as a symbol of harsh corporate culture, failed management-control structures, accounting fraud and wilful corporate corruption.

The clever idea behind Enron's initial spectacular success was that of establishing a 'gas bank' [12, 13] which traded electricity as a hard commodity. Initially praised from all quarters for its pioneering and extremely effective management, the collapse of the company revealed that its reported robust financial status was largely the product of systematically planned and manipulative accounting techniques. Senior financial executives at Enron, along with accountants at Arthur Andersen [2], set up various independent special-purpose entities (SPEs) accounts to keep a large number of liabilities off the Enron balance-sheet and smooth out profits between quarters. Though SPEs are supposed to have at least three percent equity stake at risk by outside investors, far from being independent, most of these SPEs were funded either by Enron shares or had senior Enron executives as their primary investors. Though the rising recorded profits made Enron look attractive to investors, in reality most of the company's recorded assets were inflated, or even wholly fraudulent and non-existent [9]; just like its fictitious independent SPEs.

The bubble burst for Enron, when the questioning of its financial accounts and the consequent revelations led to a sudden fall in investor confidence and severe crash in its share price. Unable to sustain their farce anymore, the Enron management filed for bankruptcy protection in late 2001. The collapse of Enron, with its estimated loss of 85,000 jobs and millions of dollars of investors' money, has become a saga of absolute ethical and financial corruption in corporate America.

Chapter 3

Experiments

This chapter details our work's research methodology and experiments. We begin from James Pennebaker's empirical textual deception detection model (see Section 2.3). The Pennebaker model is based on word-frequency counts and does not take into account sentence structure and semantics. According to the model, deception leaves a linguistic signature which is characterised by decreased frequency of personal pronouns (dissociation from the content); decreased frequency of exclusive words (reduced cognitive complexity); increased frequency of negative-emotion words (sublimated guilt); and increased frequency of action-verbs (distraction). Being based on word-frequency count profiles renders the model available for being tested, analysed and improved in an automated environment.

In our experiments we use Singular Value Decomposition (see Section 2.5), a matrix decomposition technique for unsupervised data mining. In addition to the Pennebaker model, which just sums up the frequencies of the words and gives each attribute in the word list an equal weight, Singular Value Decomposition (SVD) analyses the data in a more in-depth way. By computing the SVD for the data, we essentially include correlation information to weight each of the attributes to best reflect their differential use in deception detection.

3.1 The Dataset

We used the Enron email dataset (see Section 2.6) for our experiments. First made public as the result of the investigation into the corporation's collapse, this email corpus consists of three and a half years of emails to and from Enron personnel. These messages include organizational issues and personal communications of different kinds. Since most Enron personnel would never have imagined that their emails would become public, their email text is often markedly informal. These email messages exchanged by this large group of people, who do not form a closed group, is thus a valuable representative sample of communication both inside an organization, and in a larger context.

The version of the Enron email dataset that we have used in this research comes from William Cohen at the Carnegie Mellon University [11]. The dataset contains 517,413 emails from the mail folders of 150 ex-Enron employees including senior management executives like Kenneth Lay (ex-chairman and CEO) and Jeffrey Skilling (ex-CEO). These emails range from the late 1990's up to and including the bankruptcy filing by Enron. Each email contains the sender and receiver(s) email addresses, the timestamp, the subject line and text. Attachments have not been included and some messages have been deleted as part of a "redaction effort due to requests from affected employees" [11]. As the emails from top executives who have been convicted or are being currently prosecuted are included, it seems reasonable to believe that some shades of deception do exist in this dataset.

3.2 Creating the Email-Deception-Cue-Matrix

From the Enron email dataset we created an email-deception-cue-matrix in which each email from the dataset is represented by a matrix row; and each column represents one of the 86 words from the Pennebaker deception model [35, 38]. Each entry in the matrix counts the frequency of its respective column cue-word in the email represented by its respective row. Although a few words in the deception model are relatively common (like the words 'I', 'without', 'move'), most of the deception model words do not occur in most emails – making this resulting matrix quite sparse.

3.3 Normalization

Normalization is the mathematical process that adjusts for differences among data from varying sources in order to create a common basis for comparison [6]. It adjusts the fluctuations in data values so that the final normalized result exhibits the most representative pattern. The normalization process employed in any particular scenario is generally formalized with consistent rules applied systematically to maintain data integrity while removing redundant attributes, keys, relationships and external factors.

The standard method of normalization of a data matrix involves changing each frequency entry to a z-score. This z-score is calculated by subtracting from the entry the mean value of its respective column, and then dividing by the standard deviation of the values in the column. Subtracting the column mean makes the set of values in each column centered around the origin, and dividing by the standard deviation makes the range of values for each column roughly the same.

In our case, since SVD is a numerical technique the magnitudes of the entries in the data matrix are particularly important. As mentioned earlier, the frequencies of the words that make up the deception model are quite different – some words, such as 'I' are quite common; while others such as 'wicked', 'lies', 'lead' are rare. Due to these large differences in frequencies, even normalizing by computing standard z-scores renders a data matrix that is dominated by the use of a few very common words. To override this, from the raw frequency data matrix we took logarithms of the frequency values, and then normalized using the standard z-scores. The log transformation used implicitly asserts that doubling the frequency of occurrence of a word makes it only incrementally more significant instead of twice as important. This produced a matrix in which the values in each column ranged, at most, from -20 to $+20$. This helped to mitigate the effect of the wide variations in the usage frequency profiles of the words that make up the deception model.

Also, since the email-word-cue matrix contains many zero entries and is very sparse it is important not to allow the zero entries, which convey no information, to influence the normalization of the non-zero values; which the standard normalization scheme would do. For this reason, we computed the mean and standard deviation of only the non-zero entries in each column. We then normalized the matrix by computing z-scores only for each non-zero entry. The zero entries remained unchanged. This normalization method is, thus, able to capture the interesting non-zero structure of a matrix – regardless of how sparse it is.

Clearly, frequency is not a perfect surrogate for significance. If a content word (a noun, say) occurs more often in one document than in another, we might expect that it will be more significant in that one. However, this is probably not the case with most of the non-content words – and most of the words in the deception model are non-content words.

Figure 3.1 shows a SVD plot, with one point corresponding to each email, and viewed in two dimensions. Figure 3.2 shows the same plot, but from an angle that shows the fan shape of the points. These figures show the basic relationship amongst the emails with respect to their use of the words in the deception model. The basic structure is a fan in 3-dimensions with the origin close to, but not exactly at, the narrow end.

3.3.1 Eliminating the external effect of email message length

A previous investigation [35] of the structure present in the Enron email dataset based on word-use frequency profiles had revealed that there is a strong differentiation between short messages using rare words (eg. 'despise', 'lied'), and long messages using more typical words (eg. 'sorry', 'walk'). The study had also highlighted that the characteristic length of the emails with the most interesting correlative structure was generally relatively long.

In order to eliminate any external effect of the email text length on its cue-word-use frequency profile, we decided to divide each entry in the frequency data matrix by the total wordcount of its respective email. Figure 3.3 shows the same SVD plot as in Figure 3.2, but now with the raw frequency count entries divided by the wordcount of their respective email. Studying the emails from this U matrix space revealed that these spiked projections were the result of the dominating effect of extremely short email messages (wordcounts of 3-10) which contain one or more occurrences of the deception cue-words. The type of cue(s) found in the message led to the relative placement of the email message in a specific spike on the plot.

Thus it is clear that, though at a glance the solution to eliminate the external effect of the email wordcount on its cue-word usage frequency profile may appear to be to simply divide the latter by the former, this does not eliminate the external effect of the message length. In fact, it leads to a rather lopsided effect where the data matrix is now dominated by short-length messages.

Figure 3.4 shows a plot of the email wordcount versus the total cue-words frequency count for each email in the data matrix. It appears that, though the message length and the total cue-word usage are linearly related till about the 280 wordcount boundary, the steadily rising slope seems to fade way and flatten out thereafter.

Our aim was to achieve a consistently flat, uniform distribution of the cue-words frequency counts regardless of the email wordcount. Though it appeared obvious to leave emails with wordcounts greater than 280 untouched, something more intuitive was required for emails with wordcounts less than approximately 280. In Figure 3.5, we can see that dividing the total cue-word count for an email (wordcount less than 280) by its total wordcount leads to the same pumped-up effect for short messages as was observed in Figure 3.3.

Note that in Figure 3.4 the left-hand end of the count curve is approximately logarithmic. Therefore dividing the total cue-word count for an email (wordcount less than 280) by the logarithm of its total wordcount would lead to a result of approximately 1 (modulo some scaling). Keeping this in mind we re-plotted the SVD graph. Figure 3.6 is the same as Figure 3.4, but it incorporates this new normalization method. Here the emails with wordcount more than 280 have been left untouched, but for emails with wordcount less than 280 the total cue-word frequency count has been divided by the logarithm of the email wordcount (modulo scaling by 7 in this case).

Clearly, Figure 3.6 illustrates a more uniform distribution than Figure 3.4. This normalization process is able to successfully eliminate the external effect of email message length on the cue-word use frequency profiles used by the deception-detection model.

3.4 Applying the Pennebaker Model

We applied Pennebaker's deception model to the normalized form of the email-deception-cue-matrix. As mentioned earlier, the model is characterized by decreased frequency of personal pronouns (dissociation from the content); decreased frequency of exclusive words (reduced cognitive complexity); increased frequency of negative-emotion words (sublimated guilt); and increased frequency of action words (distraction). To enable the cue-matrix to best capture deception, we multiplied the first person pronouns and exclusive words columns by (-1) so that high-frequency values in these were transformed into high negative values. This alteration is important because SVD, being a numerical technique, interprets increased attribute values as a sign of increased importance. Negating the high frequencies for the first-person pronouns and exclusive words reiterates the fact that low values of these are significant to us.

Figure 3.7 shows the same plot as seen in Figure 3.2 – the new normalization method and the Pennebaker deception detection model have been included in this representation. This plot illustrates the basic relationships amongst the emails with respect to the usage of words in the deception model. The distinct spikes here represent cue-words/cue-word categories which the SVD technique finds most useful in positioning an object. Notice that the basic structure, that of, a fan in 3-dimensions with the origin close to, but not exactly at the narrow end – remains consistent throughout our experiments. It can be seen in Figure 3.1, Figure 3.2 and now in Figure 3.7.

Figure 3.8 shows a plot of the singular values as described in matrix S (Section 2.5) – a diagonal matrix with non-increasing values which indicate the importance of each dimension.

Since SVD is completely symmetric (see Section 2.5) it can be used to study relationships between the objects (emails) and the attributes (deception-cue-words); those amongst the objects; and those amongst the attributes.

Figure 3.9 shows a SVD plot for the attributes, that is for the cue-words. It depicts the relationships among the different words used by the deception model (see Table 3.1 in Section 3.5). Proximity of points in the plot shows similarity in the way the words are used. The distance from the origin highlights the interesting correlation structure among the words.

Figure 3.10 depicts the points corresponding to the cue-words (in red) overlaid with the points corresponding to the emails (in blue). Though not to scale, the figure after SVD has a physical meaning – messages (objects) are being drawn towards cue-words (attributes) for which they have affinity. Messages with high frequencies of personal pronouns lie close along the bottom edge of the fan and those with high frequencies of exclusive words are found higher up the fan-shaped structure. This plot served just as a starting point and we continued to investigate the data matrix.

3.5 Refining the Deception Model

Keila [35] recommended two main ways to improve the deception model (see Section 2.4). One approach involved increasing the number of deception cue-words and the second involved reducing the set of words. The first approach involves a human reader who manually parses emails that have been categorized as deceptive by the current model and intelligently picks out words that appear in maximum number of such deceptive emails and/or are in themselves obvious indicators of deception. These manually selected words are then added to the cue wordlist. This helps in increasing the count of words that help in deception detection.

According to the second approach, after applying the matrix decomposition technique and obtaining a plot like the one shown in Figure 3.9 we can exclude words that we find are least useful in classifying emails as deceptive. These would generally be words that are found to be too confounded with other words and thus serve a minimal individual purpose. We can walk through this procedure in an iterative fashion; successively excluding words in each iteration. This approach aims to reduce the total number of words that serve as pointers for deception.

Keila [35] had mentioned that both the approaches of refining the Pennebaker deception model, that is increasing the number of cue-words that indicate deception or reducing the number of cue-words that serve as pointers of deception, are equally valid.

To explore the first approach of increasing the number of words in the cue list, we included the word 'we' – an obvious plural extension of the commonly used first-person pronoun word 'I' – in our deception cue wordlist. We were curious to find out whether the role of such manually selected similar words is actually different in context of the model. In Figure 3.11 we can see that even after log-scaling and normalization the two words, 'I' and 'we' dominate the entire set of words. Notice the distance between the points representing 'I' and 'we' and their orthogonality. Even after normalization, their roles are indeed very different. It is clear that the words 'I' and 'we' tend not to occur in similar contexts. This finding is also not surprising from an organizational point of view. In a business setting, as is the case with Enron, certainly team work is prized, but still individuals think very differently about their individual and team accomplishments. The deception model, therefore, is correct to exclude the word 'we' from the cue-word list.

Keeping in view Figure 3.9, we then moved on to explore the second approach to enhance the model by trying to eliminate cue-words that serve a minimal individual purpose in deception detection. As mentioned earlier, the distance from the origin is an indicator of the significance of the word; as to how unusual its correlation is with the other words and indirectly with the other emails. The list of 86 words from the deception model sorted by distance from the origin is shown in Table 3.1 (on page 30). This table provides us a high-level view of the overall importance of each word as a marker for deception. Since some words like those in the table's first column play very little role in determining whether a message is deceptive, such words can be removed from the model without diluting its deception-detection capability.

As a starting point, we decided to exclude words that have a total frequency count of less than 35 in the entire Enron dataset. This is a group of 13 words, as listed in Table 3.2 (on page 31). Our intuition was that words with such low total frequency counts were less likely to lend significant advantage in deception detection. Notice that most of the words in Table 3.2 lie in the first column of Table 3.1 – that is they

terrible	inferior	hate	drive
hated	boring	carrying	without
hates	lied	fear	walk
lame	whereas	nor	move
vain	lonely	lies	action
wicked	disappointing	unless	lead
fled	flew	weak	i'd
jerked	greed	tragedy	i've
ignorant	lie	moved	look
inadequate	crazy	myself	i'll
worthless	anger	rather	follow
jerk	enemy	although	take
disappointed	driven	goes	i'm
dumb	stressed	devil	go
bastard	suspicious	bitch	going
anguish	followed	sorry	my
agony	arrived	bringing	me
afraid	f-word	taking	but
abandon	however	carry	or
dislike	except	run	I
despise	mine	loss	
besides	arrives	arrive	

Table 3.1: Table of deception model words sorted by distance from the origin (least interesting at top left, most interesting at bottom right)

are least interesting in the list of cue-words sorted based on their distance from the origin. This observation further reiterated our decision to delete these words from the cue-word list.

Figure 3.12, the SVD plot for the attributes obtained after excluding the 13 words from Table 3.2, looks strikingly similar to the earlier Figure 3.9. The trend is also maintained in the zoomed in Figure 3.13 (as compared to Figure 3.15) and Figure 3.14 (as compared to Figure 3.16). This similarity in the plots before and after removing

Cue Word	Count
vain	10
wicked	34
fled	03
jerked	13
ignorant	31
inadequate	00
worthless	22
bastard	23
anguish	04
agony	14
dislike	29
despise	02

Table 3.2: Table of cue-words that have a total frequency count of less than 35 in the entire Enron dataset. Their respective total frequency counts are listed.

the words with low total frequency counts, again highlights the fact that the excluded cue-words do not hold much individual significance. In fact they only have a diluting effect on the model, if any, and may also lead to wastage of useful time and space resources. Thus, reducing the number of cue-words not only sheds further light on the deception model itself, but also helps us to reduce the amount of work that would be required to analyse new datasets.

3.5.1 Working with cue-word categories

Since the Pennebaker deception model is based on four kinds of words, an important question is whether these should be thought of as four separate factors. The beginning of an answer to this vital question is provided by Figure 3.15 which shows the SVD plot for the deception cue-words (same as in Figure 3.9), but a region closer to the origin. The point corresponding to 'I' is now outside the plotted region and so not visible. Notice here, that the exception words (‘but’ and ‘or’) tend to form a linear cluster reaching out towards the top of the figure. First-person pronoun words tend to form a similar cluster but reaching out towards the lower left (which is also the direction of the point representing ‘I’ although this can’t be seen in the figure). Words from the other two categories remain clustered near the origin.

Figure 3.16 shows the same plot as shown in Figure 3.15, but of the region close to the origin in even more detail. The first-person pronoun cluster continues to be visible in the plot in the words ‘I’d’ and ‘I’ve’. Since in SVD plots, the distance from the origin highlights the interesting correlation structure among the words and both the action words and negative emotion words are clustered so close to the origin – it is clear that they are not good signals for deceptiveness. Note that the action words are generally farther from the origin than negative emotion words.

The SVD plots for the individual cue-word categories can be seen in Figures 3.17 to 3.19. We see in these figures that the four categories of words form four individual clusters. In Figure 3.17, though, ‘I’ is the most significant member by a large margin, the cluster of first-person pronouns is, on the whole, weak. In Figure 3.18 there are members of the cluster on both sides of the origin, suggesting that some of these words signal deception, while others do not. Nonetheless, the cluster of exclusive words can be clearly seen. Figure 3.19 shows, that ‘go’ and ‘going’ are the most significant action words. As can be seen throughout the plots, action words are not as strongly predictive of deception as the first-person pronouns and exclusive words. Figure 3.20 depicts the negative emotion words, and again shows a cluster with words on both sides of the origin. Notice that the scale of this plot is also much smaller than of the previous ones, indicating that this class of words has little effect on detecting deception. Though there are plenty of examples of negative emotion words in the Enron corpus, these words probably occur in the context of personal rather than business emails; so their correlations with other words are also probably different. In a business context, many negative emotion words are regarded as inappropriate. This may also be the reason why, in the SVD plots, these words are found to be so confounded with each other and clustered around the origin.

It is worth highlighting here that, within the first-person pronouns the word ‘I’ seems to have the most significant and distinct effect (is in the same direction as the other personal pronouns, but far apart in magnitude). In the context of the exclusive words, though all the words seem to lie in a single cluster, the words ‘or’ and ‘but’ seem to dominate the scene.

We wanted to investigate if any sub-clusters exist within these individual clusters formed by each of the four cue-word categories; that is while excluding the words ‘I’ ; ‘or’; and ‘but’. For this purpose we employed data-clustering techniques [4, 5], and ran the k-means [15]; hierarchical(single and complete) [14]; and random forest [18] clustering algorithms on each of the individual word categories. Running these clustering algorithms on the individual four clusters verified that no further unintuitive sub-clustering was possible.

It seemed fair to consider the entire deception cue wordlist as if it were divided into seven unique driving forces – namely, the first-person pronoun word ‘I’; the remaining set of first-person pronouns; the exclusive word ‘or’; the exclusive word ‘but’; the remaining set of exclusive words; action words; and negative emotion words. Each of these seven categories plays a distinct and significant role in deception detection – that is, they work as seven different latent factors.

The SVD plot in Figure 3.21 depicts the first-person pronouns as two separate factors – the first-person pronoun word 'I'; and the remaining set of first-person pronouns. The large distance between the points in the plot reflects how these factors act as distinct driving forces in the context of the deception-detection model.

Figure 3.22 shows a similar SVD plot for the exclusive words category. The exclusive words are shown as three separate factors – the exclusive word 'or'; the exclusive word 'but'; the remaining set of exclusive words. The figure highlights the importance of understanding these three factors as distinct driving forces.

Figure 3.23 is the same plot as shown in Figure 3.9. It explicitly depicts the relationships among the words used by the deception model as if they were divided into seven unique driving forces. Each of these seven categories works as distinct latent factors in the deception detection model. Based on this deduction, we created a new version of the email-deception-cue-matrix.

In this matrix each row still represents an email, but there are only seven columns – each corresponding to a category of word(s) associated with deception. The seven columns contain, for each email, the frequencies of:

- First-person pronoun 'I';
- The remaining set of first-person pronouns (mine, myself *etc.*);
- Exclusive word 'or';
- Exclusive word 'but';
- The remaining set of exclusive words (besides, however *etc.*);
- Action verbs (move, run, lead *etc.)*;

- Negative emotion words (anger, abandon, hate *etc.*);

We used SVD to decompose this new matrix version and from the resultant plots (as discussed in the next chapter) continued to study the Enron email dataset for the purpose of deception detection.

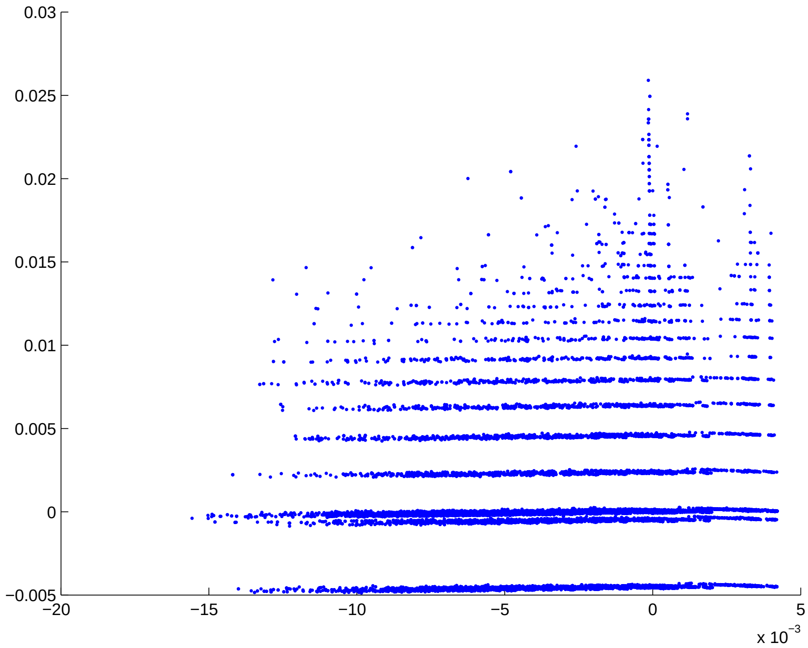

Figure 3.1: SVD plot in 2 dimensions, with one point for each email. Depicts the basic relationship amongst emails with respect to their use of the deception-cue-words.

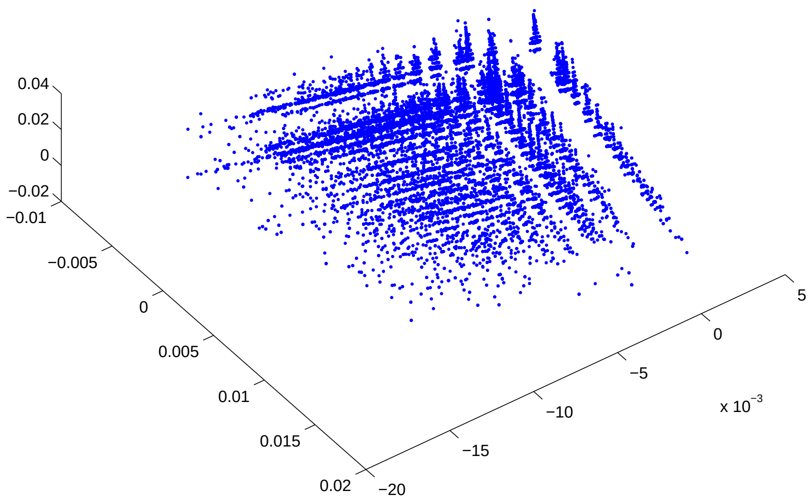

Figure 3.2: Another view, SVD plot in 3 dimensions with one point for each email.

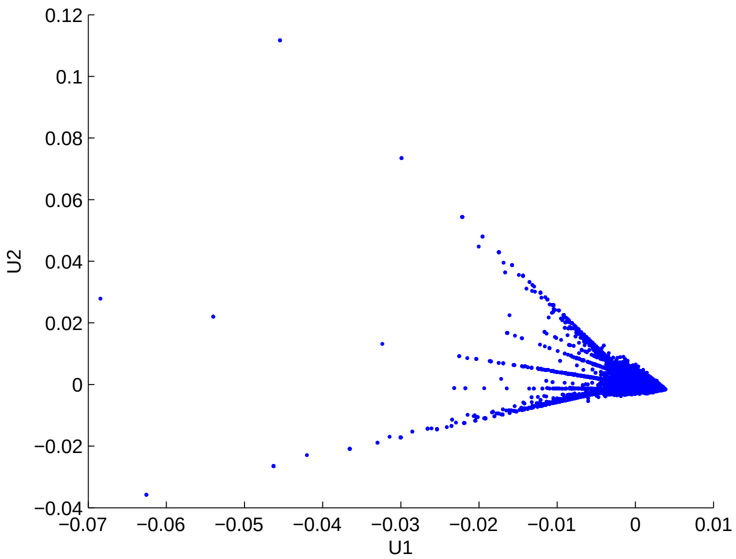

Figure 3.3: SVD plot in 2 dimensions with one point for each email. The cue-word frequency counts have been divided by their respective email wordcount as an attempt to eliminate the external effect of the email text-length on its cue-word-frequency profile. However, this leads to a lop-sided effect where extremely short messages, that contain one or more occurrences of the cue-words, dominate the entire data matrix space.

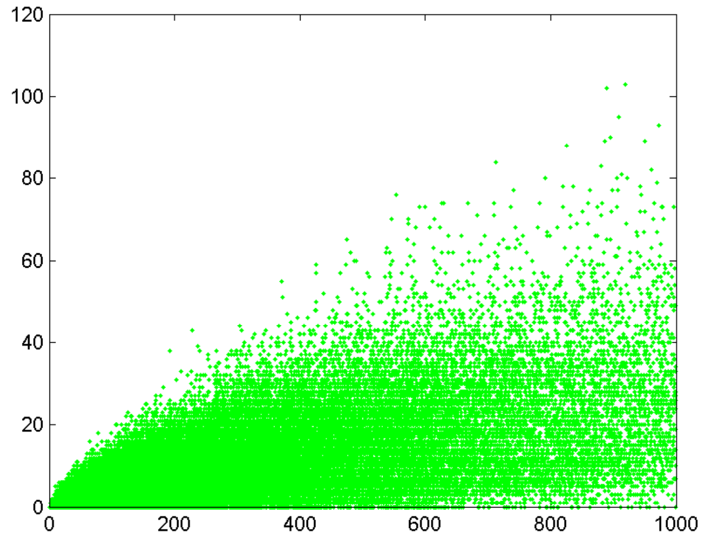

Figure 3.4: Plot of total cue-words frequency count(y-axis) against total wordcount(x-axis); with one point for each email. Notice, that though the message length and the total cue-word usage appear to be linearly related till about the 280 wordcount boundary, the steadily rising slope seems to fade away and flatten out thereafter.

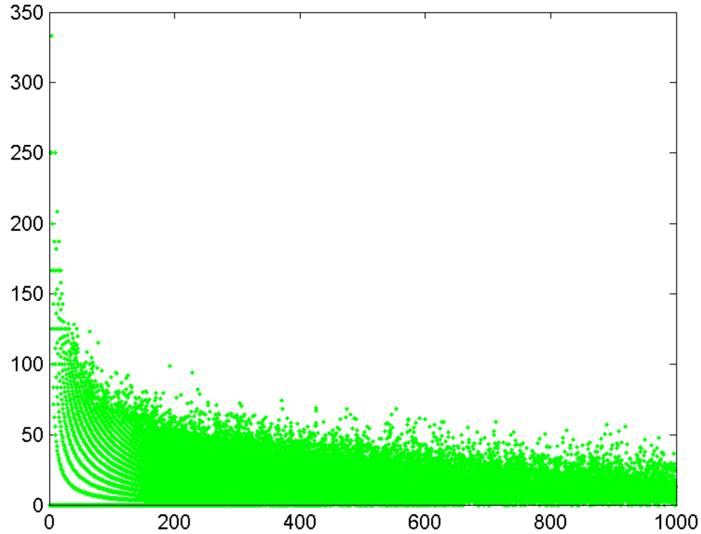

Figure 3.5: Total cue-words frequency count along y-axis and total wordcount along the x-axis. For emails with wordcount less than 280 the total cue-words frequency count has been divided by their total wordcount (then scaled by 500).

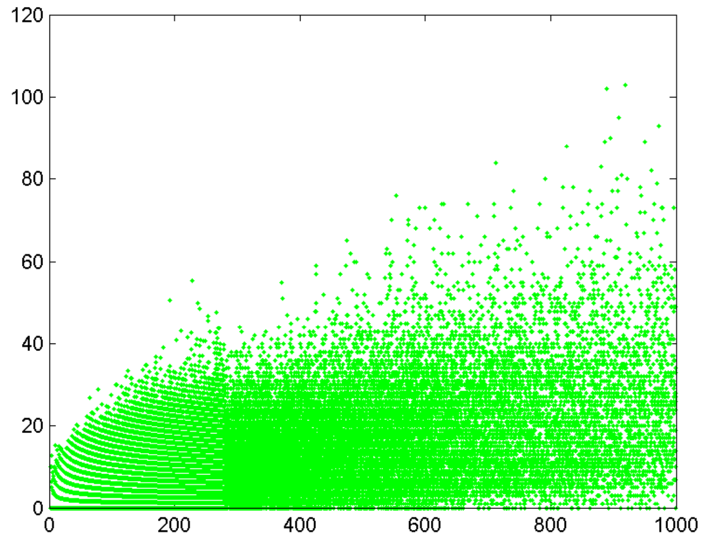

Figure 3.6: Total cue-words frequency count along y-axis and total wordcount along the x-axis. For emails with wordcount less than 280, the total cue-words frequency count has been divided by the logarithm of their wordcount.

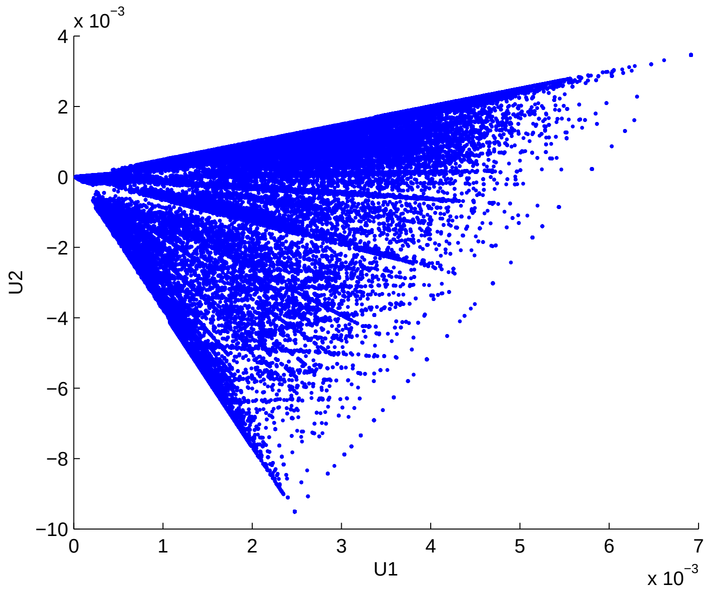

Figure 3.7: SVD plot in 2 dimensions with one point for each email, after normalization and applying the Pennebaker Deception Model.

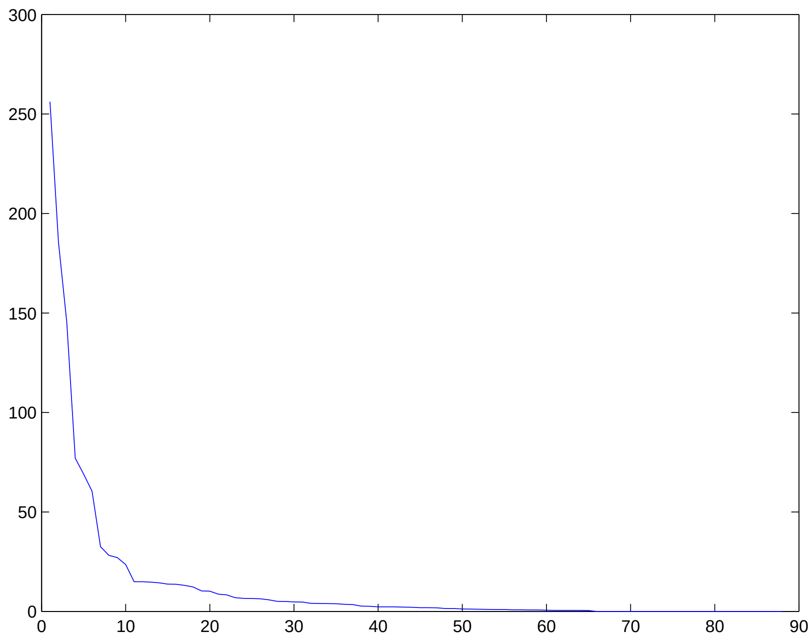

Figure 3.8: Plot of the singular values as described in matrix S – a diagonal matrix with non-increasing values which indicate the importance(y axis) of each dimension(x axis).

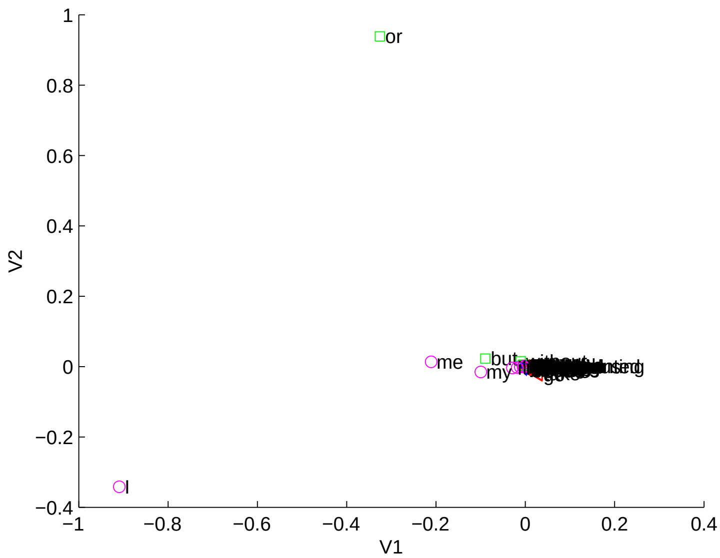

Figure 3.9: SVD plot in 2 dimensions, with one point for each word in the deception model: (purple) circles = first-person pronouns; (green) squares = exclusive words; (blue) triangles = negative emotion words; (red) triangles = action words. Proximity of points shows similarity in the way the words are used and distance from the origin highlights the interesting correlation structure among the words.

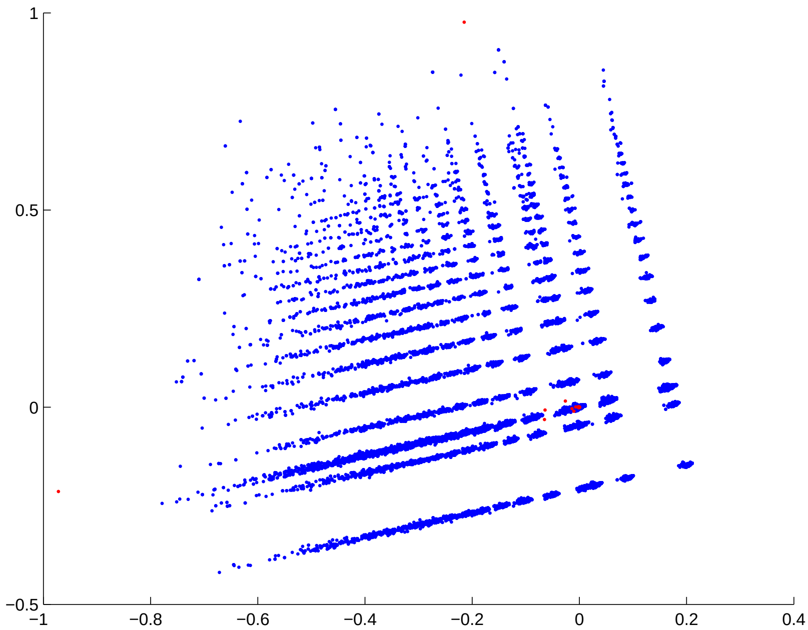

Figure 3.10: SVD plot in 2 dimensions, points corresponding to the cue-words (in red) overlaid with the points corresponding to the emails (in blue). Though not to scale, the figure after SVD has a physical meaning – messages (objects) are being drawn towards cue-words (attributes) for which they have affinity.

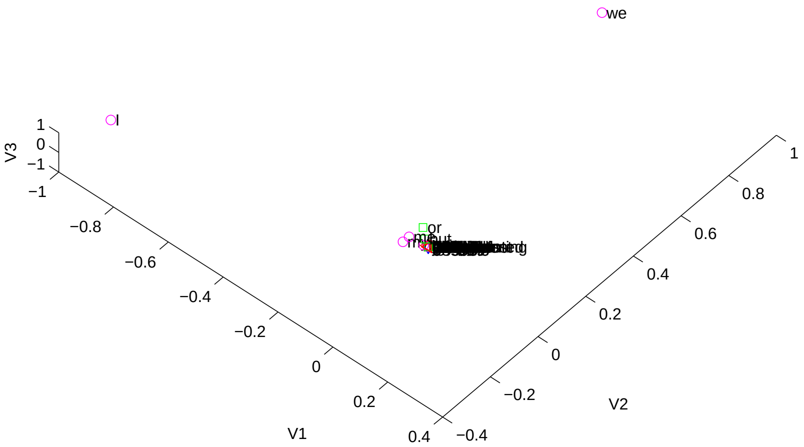

Figure 3.11: SVD plot in 3 dimensions, with one point for each word in the deception model: (purple) circles = first-person pronouns; (green) squares = exclusive words; (blue) triangles = negative emotion words; (red) triangles = action words. Additionally, the word ‘we’ has been included. Notice the distance between the points representing ‘I’ and ‘we’ and their orthogonality. Even after normalization, their roles are indeed very different.

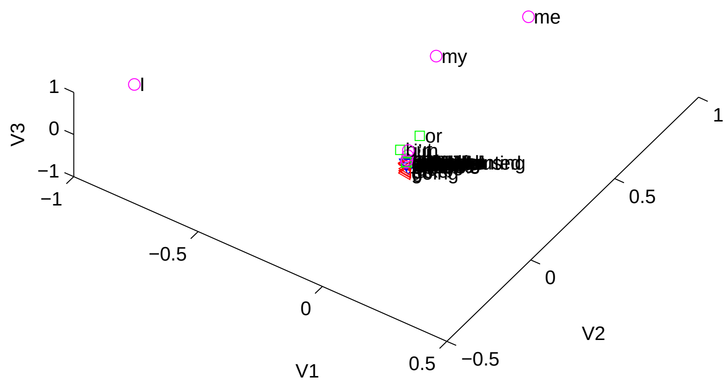

Figure 3.12: SVD plot in 3 dimensions after excluding the 13 cue-words, with one point for each word in the deception model: (purple) circles = first-person pronouns; (green) squares = exclusive words; (blue) triangles = negative emotion words; (red) triangles = action words.

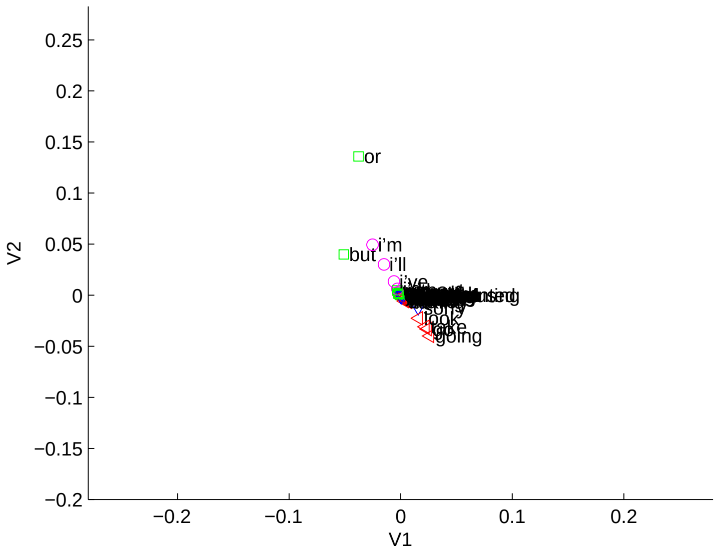

Figure 3.13: SVD plot in 2 dimensions, with one point for each word. The figure is the same as Figure 3.12, but of a region closer to the origin.

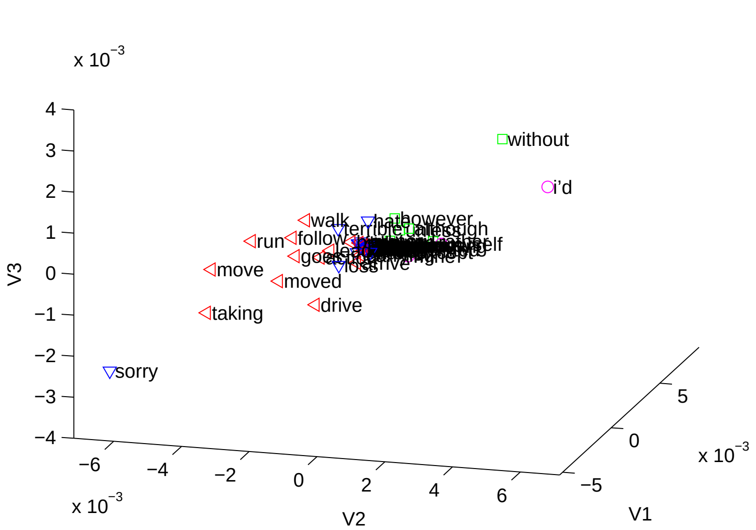

Figure 3.14: Zoomed in: Close-up of plot in Figure 3.13. SVD plot in 3 dimensions, with one point for each word.

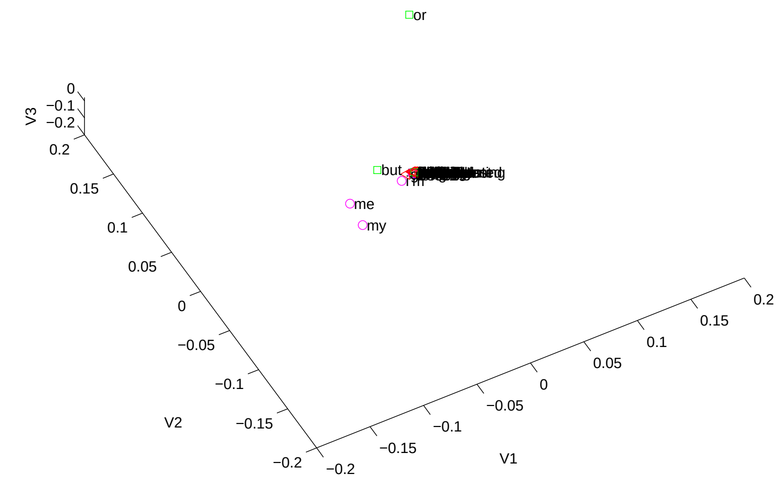

Figure 3.15: SVD plot in 3 dimensions, with one point for each word. The figure is the same as Figure 3.9, but of a region closer to the origin. Exclusive words form a cluster reaching out towards the top of the figure. Personal pronouns form a similar cluster but reaching out towards the lower left.

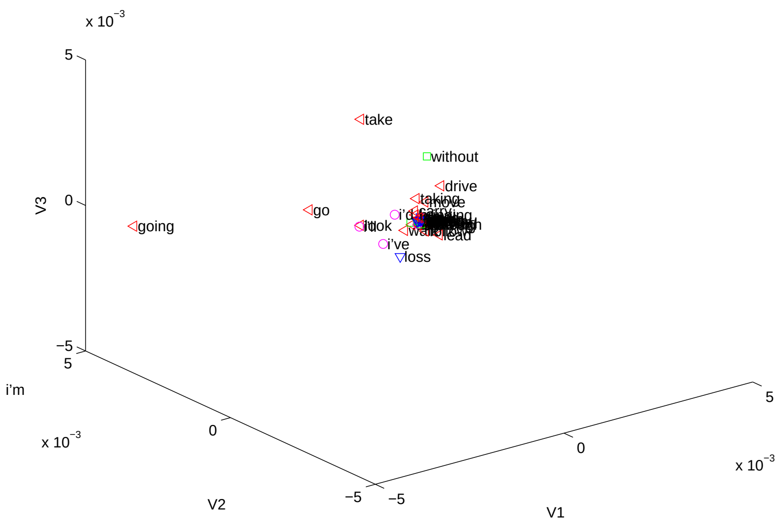

Figure 3.16: Zoomed in: Close-up of plot in Figure 3.15

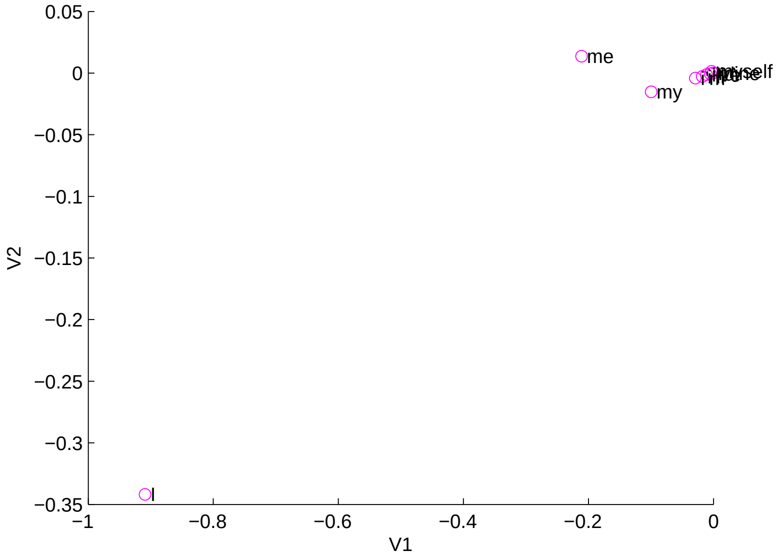

Figure 3.17: SVD plot in 2 dimensions for the individual cue-word category – personal pronouns. Though, 'I' is the most significant member by a large margin, the cluster of first-person pronoun words is, on the whole, weak.

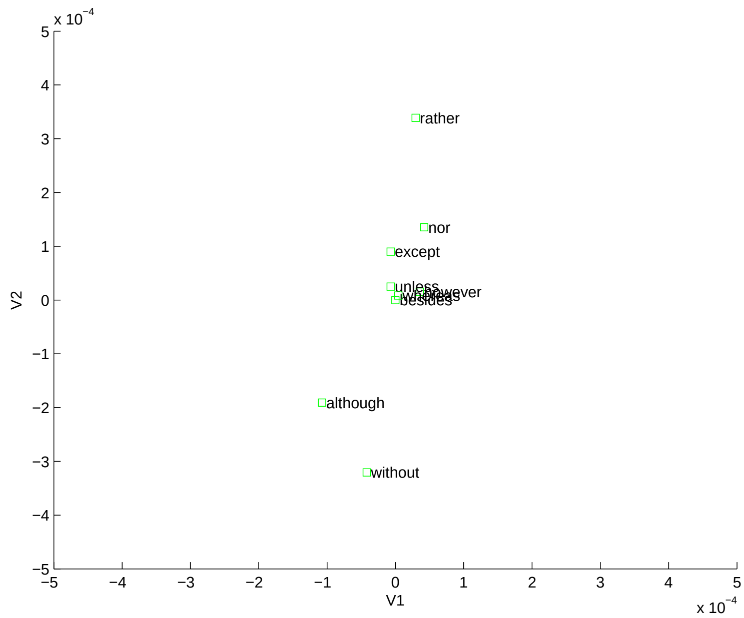

Figure 3.18: SVD plot in 2 dimensions for the individual cue-word category – exclusive words. The cluster of exclusive words can be clearly seen.

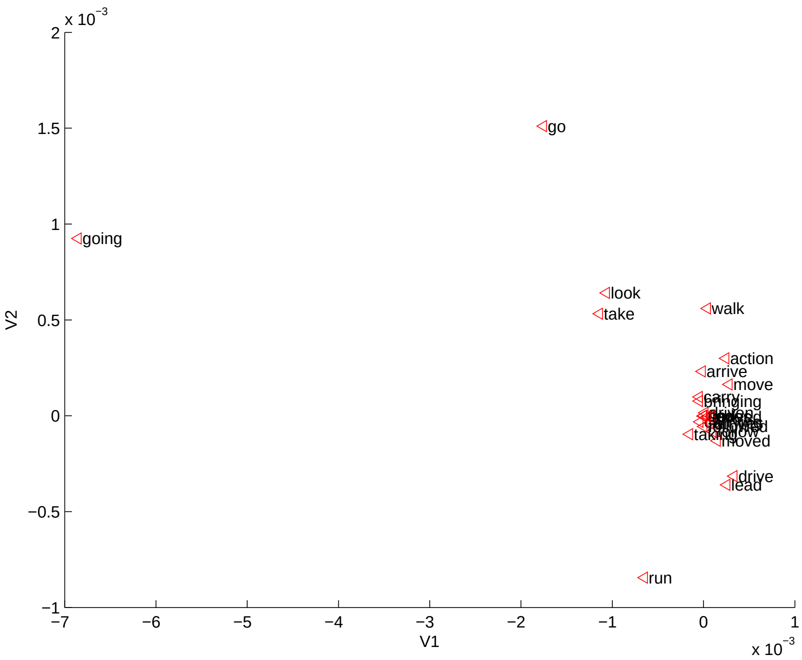

Figure 3.19: SVD plot in 2 dimensions for the individual cue-word category – action verbs. The words ‘go’ and ‘going’ are the most significant action words.

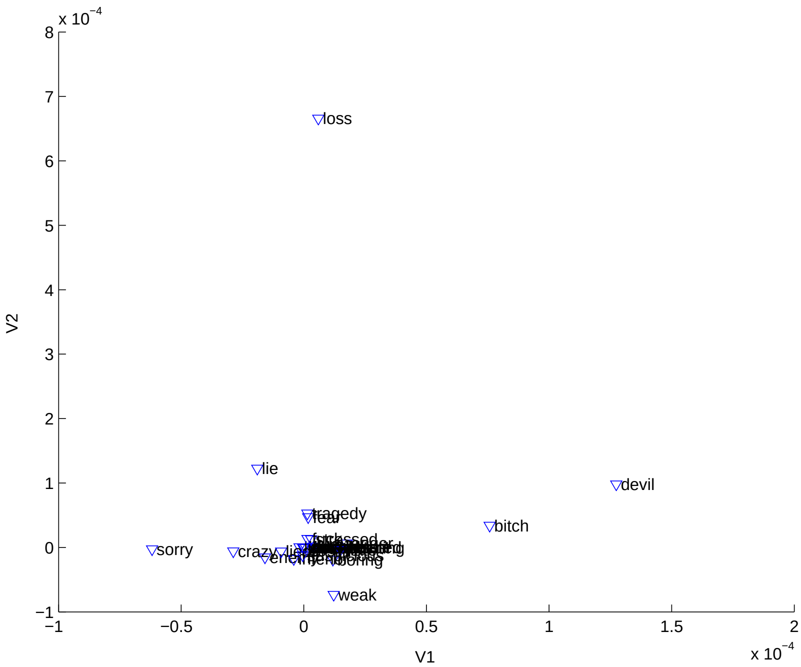

Figure 3.20: SVD plot in 2 dimensions for the individual cue-word category – negative emotion words.

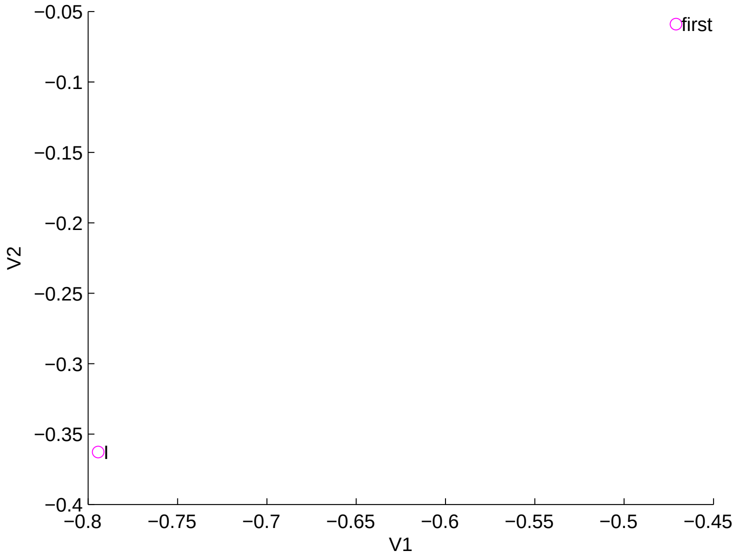

Figure 3.21: SVD plot in 2 dimensions of personal pronouns shown as two separate factors – the first-person pronoun word ‘I’; and the remaining set of first person pronouns. Within the first-person pronouns, the word ‘I’ seems to have the most significant and distinct effect (lies in the same direction as the other personal pronouns, but is far apart in magnitude).

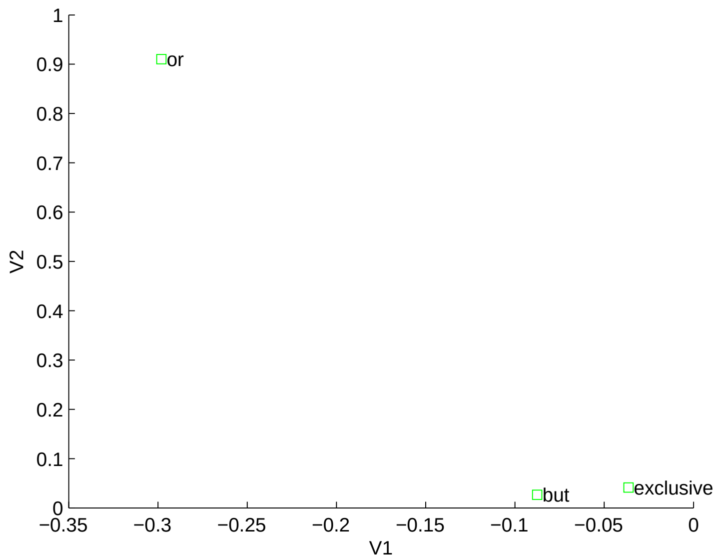

Figure 3.22: SVD plot in 2 dimensions of exclusive words shown as three separate factors – the word ‘or’; the word ‘but’; the remaining set of exclusive words.

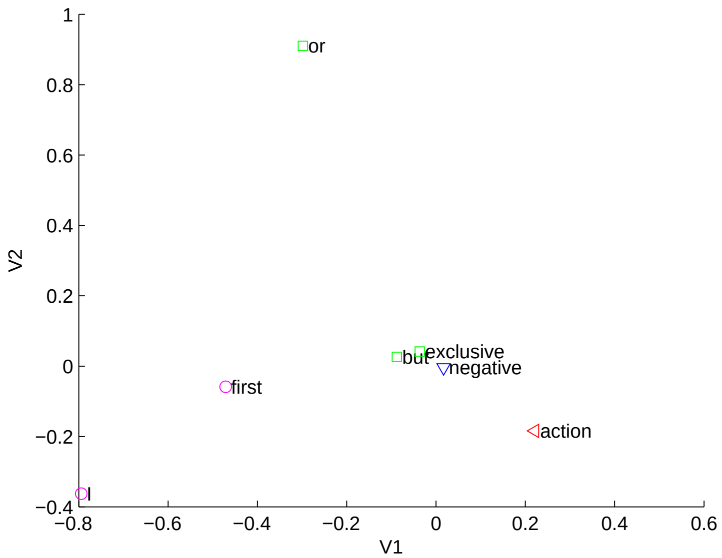

Figure 3.23: SVD plot in 2 dimensions, with cue-words depicted as the 7 different latent forces in deception detection: (purple) circles = first-person pronouns; (green) squares = exclusive words; (blue) triangles = negative emotion words; (red) triangles = action words.

Chapter 4

Results

In this chapter we detail the results that we have obtained. Starting with the normalized email-deception-cue-matrix with only seven columns, one for each of the latent factors for the deception model (see Section 3.5.1), we explore the Enron email dataset to devise a descriptive model for textual deception detection. We compare our results with those achieved by the empirical model provided by James Pennebaker (Section 2.3) and highlight the accomplishments of our work. The chapter ends with a discussion based on the results achieved.

4.1 Deception Detection in the Enron Email Dataset

We used the Singular Value Decomposition (SVD) technique (see Section 2.5) to explore the normalized version of the email-deception-cue-matrix version with only seven columns, one for each of the seven distinct latent factors of the deception model (Section 3.5.1) – namely, the first-person pronoun word 'I'; the remaining set of first person pronouns; the exclusive word 'or'; the exclusive word 'but'; the remaining set of exclusive words; action words; and negative emotion words. Figure 4.1 and Figure 4.2 show SVD plots in two and three dimensions respectively, with each email represented by a single point. The plots are the same as shown in Figure 3.7, only in this case the email-deception-cue-matrix version has only seven columns.

These plots highlight the interesting correlation structure amongst the plotted points – that of each email with other email messages and indirectly with the seven unique driving forces of the deception model. The distinct spikes here represent cue-words/cue-word categories which the SVD technique finds most useful in positioning an object. The basic structure, that of a fan in 3-dimensions with the origin close to, but not exactly at the narrow end, continues to be seen throughout our experiments in Figure 3.1 and Figure 3.2; Figure 3.7; and now in Figure 4.1 and Figure 4.2.

Data that can take on only a discrete set of values, in this case word frequencies, typically forms the discrete clumps of points seen in these figures. As we only count the frequencies of a small subset of words, many messages whose content is completely different look very similar based on this subset. Notice that as we continue to refine the deception-detection model, the plots just become more clear and classified into distinctive spikes.

On studying the emails in this U matrix space we realized that how much an email contains reduced frequencies of first-person pronouns and increased use of exclusive words primarily decides the position of the email in this plot. Messages that contain predominately one of these markers are close to one of the edges of the fan, whereas messages in the middle of the fan contain, in differing proportions, a mixture of both markers.

In Figure 4.1, on investigating closely we found that the three prominent spikes seen in this plot going from the bottom-most to top-most (from the region labelled A to C) correspond to messages that contain (1) the first-person pronoun word 'I' – along A; (2) the exclusive word 'or' – along B; and (3) the exclusive word 'but' – along C, respectively. These spiked projections represent cue-words which the SVD technique considers most useful in positioning an object in the context of deception detection.

Email messages with a high frequency (before multiplying with (-1)) of first-person pronouns and/or high frequency of exclusive words are thrown outwards towards the periphery of the fan (the right-hand edge).

The frequency counts of the personal pronouns in the messages decrease as one moves away from the fan's periphery towards its origin (from point A to D); and away from the bottom-most spike towards the top-most spike (from point A to C).

The exclusive-word frequency counts, also, decrease as one moves away from the fan's periphery and towards its origin (from point C to D). However in contrast to first-person pronouns (where the frequency decreases from bottom to top-most spike), the frequency counts of exclusive words in the messages decrease from the top to bottom-most spike (from point C to A).

The messages close to the top-most spike (along C) are also more emotionally charged in their content. Greater complexity in terms of increased usage of action words and negative emotion words is seen in messages as one moves away from the fan's periphery and towards its origin.

Note that along the top-most spike (along C) moving from the fan's periphery towards its origin we observe decreased usage of first-person pronouns and exclusive words, and an increased usage of action words and negative emotion words. These characteristics are most evident in messages plotted close to the origin. Thus, based on Pennebaker's theory of deception and according to our study of the U matrix space, the most deceptive messages in Figure 4.1 should ideally lie close to the fan-shaped structure's origin and perhaps along its top-most spike close to the origin.

On studying the emails plotted close to the origin of this fan-shaped figure, we found that these messages were mostly some form of formal and often legal correspondence – like a list of things to keep in mind while signing contracts; instructions from people of authority *etc.* Messages found elsewhere in the plot (apart from the origin) are informal and sometimes profane. For sample email messages refer to Appendix *A*. We have at times listed only an excerpt from the email because the actual email message is too long. The objective is to give the reader an idea of the 'type' of email content.

4.2 Correlating our results with the Pennebaker Deception Score

We were intrigued to find out whether the most deceptive messages according to the *Pennebaker Deception Score* (Section 2.3) matched our findings based on applying SVD to the normalized email-deception-cue-matrix.

Note, that our email-cue-matrix had already been tuned to capture deceptive writing as characterized by the empirical model provided by Pennebaker (Section 3.4) – that is, low usage-frequencies of first-person pronouns, exclusive words and increased frequencies of action and negative emotion words are considered interesting and significant. Therefore, our findings as described in the above section (Section 4.1) are based on applying the SVD technique beyond the Pennebaker Deception Model. Our aim, now, was to compare our results found on investigating the U matrix space with the results found independently based on the *Pennebaker Deception Score*.

Figure 4.3 shows a SVD plot in three dimensions (after normalizing, applying the Pennebaker Deception Model to the email-cue matrix), with each email represented by a single point. The most deceptive messages according to the *Pennebaker Deception Score* are plotted in black. Figure 4.4 shows the same plot as Figure 4.3, but from a different angle. According to our description of the U matrix space in the above section (Section 4.1), we would expect the most deceptive messages to be plotted close to the origin and perhaps along its top-most spike close to the origin. As expected, the most-deceptive messages have been plotted close to the origin in Figure 4.3 and Figure 4.4.

However, closer observation of just the most deceptive messages, as shown in Figure 4.5 and Figure 4.6, reveals that there is a clear streak of messages that are being pulled away from the origin of the fan-shaped structure and towards its periphery; along the spike which we had earlier identified as representing the exclusive cue-word 'or' (see Section 4.1). Since SVD uses cue-words/cue-word categories it finds most useful in deception detection to position and cluster points, these figures make us realize that the cue-word 'or' is a particularly effective pointer to deception.

This is a further step forward towards understanding the differential use of cue-words in detecting deception. While the words 'I', 'or', 'but' (as shown in Figure 4.1 and Figure 4.2) are important pointers, in the exclusive words category the cue-word 'or' is probably a more effective and reliable marker. From here, we can also learn that increased use of the word 'I' or of first-person pronouns in general; and the increased use of exclusive words indicates uninterestingness in terms of deception.

Our results find resonance in the work of James Pennebaker's group [37] which independently found that among the four main categories of deception cue-words (first-person pronouns/exclusive words/action words/negative emotion words) exclusive words are the most reliable in defining deception.

Further we decided to plot the entire dataset as if they were bands of relative deception – after applying the SVD technique we ranked email based on the *Pennebaker Deception Score*. The resultant SVD plots are shown in Figures 4.7, 4.8, 4.9, 4.10.

It is important to note at this point that, though our results do seem to find some resonance with those found independently based on the *Pennebaker Deception Score*, there is one vital difference. The Pennebaker Score as seen in Figure 4.9 ranks messages right along the top-most spike (refering to the cue-word 'but') as slanted slabs – from most to least deceptive as one moves away from the origin. In this case, therefore, messages found in the yellow band of Figure 4.9 close to the fan's right-hand edge appear more deceptive than those found, say, in the preceding blue band along the top-most spike of the fan. This can be misleading and untrue.

This can be yet better understood by seeing Figure 4.11 where the least deceptive messages according to the *Pennebaker Deception Score* are marked in black. The *Pennebaker Deception Score* plots these points only where we observe the least usage-frequency of the actions words, negative emotion words (along the top-most spike that represents the cue-word 'but' in the figure and close to the fan's periphery). However, careful study of emails in the U matrix space reveals that messages all along the fan's periphery fall in the same category – least deceptive and therefore uninteresting. So based on the ranking provided by the *Pennebaker Deception Score*, one might waste resources looking for deceptive messages in the yellow band close to the fan's right-hand edge in Figure 4.9; these messages are probably the least interesting in the context of deception detection.

The Pennebaker Deception Score weights each attribute in the deception-cue-list equally and simply sums up the counts of the cue-words' frequencies into a single score. This is also the reason why the score ends up ranking messages along the top-most spike as slanted slabs (most to least deceptive as one moves away from the origin).

Recall that in Figure 4.1, along the bottom-most spike the increased usage-frequency count of personal pronouns is counter-balanced by the decreased frequency count of exclusive words. Similarly along the top-most spike, the increased usage-frequency count of exclusive words is counter-balanced by the decreased frequency count of personal pronouns (Section 4.1). With the other two cue-word categories' counts being almost counter-balanced, along the top-most spike we observed an increased usage of action and negative emotion words moving from the fan's periphery towards its origin. Thus, the *Pennebaker Deception Score*, which simply sums up the counts of the cue-words' frequencies into a single score, decides on the deceptiveness of a message based just on the increasing count of the action words and negative emotion words.

The main difference between our descriptive model and the Pennebaker model is that by computing the SVD for the data, we include correlation information to weight each of the attributes to best reflect their differential use in deception detection. Our analysis based on the SVD technique shows that, though messages along the top-most spike of the fan may be more deceptive than those along the bottom-most spike (due to increased usage-frequencies of action and negative emotion words), deception is best depicted as a set of concentric bands (rather than slanted slabs tilted against the top-most spike) starting from around the fan's origin and moving towards its periphery; depicting most to least deceptive messages as one moves away.

We also decided to highlight Kenneth Lay and Jeffrey Skilling's (see Section 3.1) emails on the U matrix space in red and black respectively. The resulting plots are shown in Figure 4.12 and Figure 4.13. Though these plots do not seem to define anything conclusively, it is not hard to see a clogging of messages close to the fan-shaped structure's origin – exactly where we have earlier predicted that the most deceptive messages should appear.

4.3 Discussion

In context of the normalization process (Section 3.3), by calculating the column mean and standard deviation based on only the non-zero entries and then computing the z-scores only for the non-zero values, we have been able to capture the interesting features of the sparse email-deception-cue-matrix. Unlike the standard normalization process, the zero entries (which convey no information) have no influence on this normalization. By this method we are able to capture the significant non-zero features of a matrix; regardless of how sparse it is.

The log transformation used during the normalization process implicitly asserts that doubling the frequency of occurrence of a word makes it only incrementally more significant instead of twice as important. This transformation helps us override the wide variations that exist in the usage frequency profiles of the words that make up the deception model.

Additionally as described in Section 3.3.1, a functional solution to removing the external effect of the email text-length on its cue-word usage profile has been devised. Though, at a glance, the solution to eliminate the external effect of the email text-length on its cue-word frequency profile may appear to be to simply divide the latter by the former this may lead to a lopsided result where the short messages that use one or more of the cue-words will dominate the entire matrix structure.

Previous research work [35] discussed that both the approaches to refining the deception detection model, that is reducing the number of deception cue-words or increasing the number of cue-words, are equally good. However, by trying to include the word 'we' into the deception-cue-word list we realised that the approach of increasing the number of cue-words can be tricky and sometimes misleading. Manually selecting common, oft-repeated and/or similar words from messages flagged as interesting by the current deception model, and including them in the cue wordlist requires significant human intervention. As noted in empirical studies in the field of psychology [31, 37, 43], human capability to distinguish a deceptive message from a non-deceptive one is probably around 50 percent; which is not much higher than just plain chance or the prediction result based on the toss of a coin. Indeed, it has been borne out in studies that even trained law-enforcement personnel detect deception at rates little better than chance.

Our experiments have shown that reducing the number of cue-words used by the model is a more effective approach to improving deception detection. This method involves minimal human intervention. It focuses on removing words that are too confounded with other cue-words in the SVD plots and thereby serve little individual purpose as pointers of deception.

Certainly, all the cue-words described by Pennebaker do not make an equal contribution to the distinctions of the deception detection model. However, one needs to be careful while removing words from the cue wordlist. Our decision to remove the 13 words shown in Table 3.2 from the cue wordlist was based on two well-grounded reasons – (1) that they have a total frequency count of less than 35 in the entire Enron dataset; (2) that they were some of the least interesting cue-words based on their distance from the origin in the SVD plot.

Since the SVD plots achieved before and after removing these 13 cue-words looked strikingly similar, we can safely conclude that these words do not hold much individual significance in deception detection. They only have a diluting effect on the model and may also lead to wastage of useful time and space resources in the process. By removing them from the wordlist we have not only been able to shed further light on the model itself, but also reduce the amount of work required to analyse other new datasets.

Having sorted the cue-words based on their distance from the origin in the SVD plot (Section 3.5), we learn that the cue-words of greater importance (placed at the bottom right of Table 3.1) are all very bland words that occur commonly in text, but do not carry much content. Such words tend to be produced by largely subconscious processes. They are not those upon which we focus our conscious attention. This translates into one of the strengths of the deception model, for even with awareness that a deception-detection model may be applied it is hard to consciously change the frequency of such words.

Instead of focusing on building a predictive model or solely testing the predictive power of individual cue-words, we have tried to work on a descriptive model that highlights the qualitative differences that characterize deceptive communication. This descriptive model takes into account several linguistic dimensions in the form of cue-words, but weights them according to their differential use in detection deception. Employing a large corpus of real organizational emails has only further helped in investigating the structure of the model and in understanding the multivariate profile of deception.

This descriptive model technique can help us detect unusual and deceptive communication. Also, it allows us to, rank emails along a scale (for instance from negotiation and spin to deception and plain lying) that reflects their importance in unearthing malfeasance within and without of an organization.

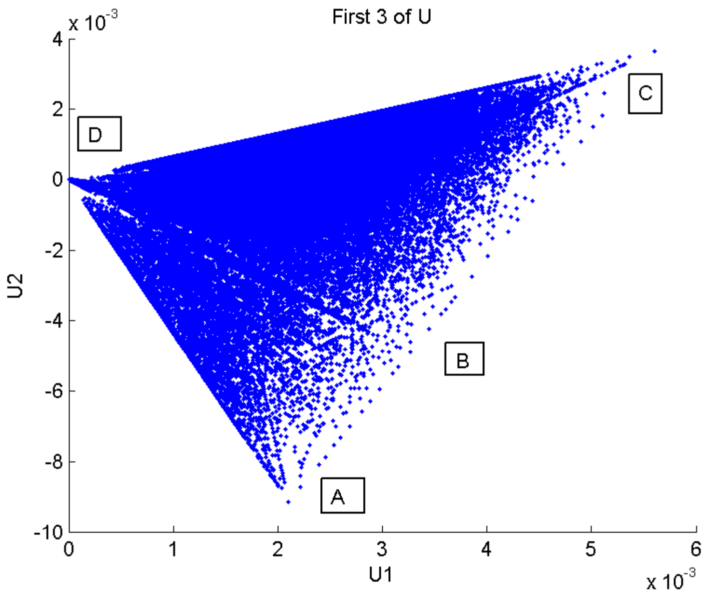

Figure 4.1: SVD plot in 2 dimensions with one point for each email. This email-deception-cue-matrix version had only seven columns; one for each of the seven distinct latent factors of the deception model – namely, the first person pronoun word ‘I’; the remaining set of first-person pronouns; the exclusive word ‘or’; the exclusive word ‘but’; the remaining set of exclusive words; action words; and negative emotion words.

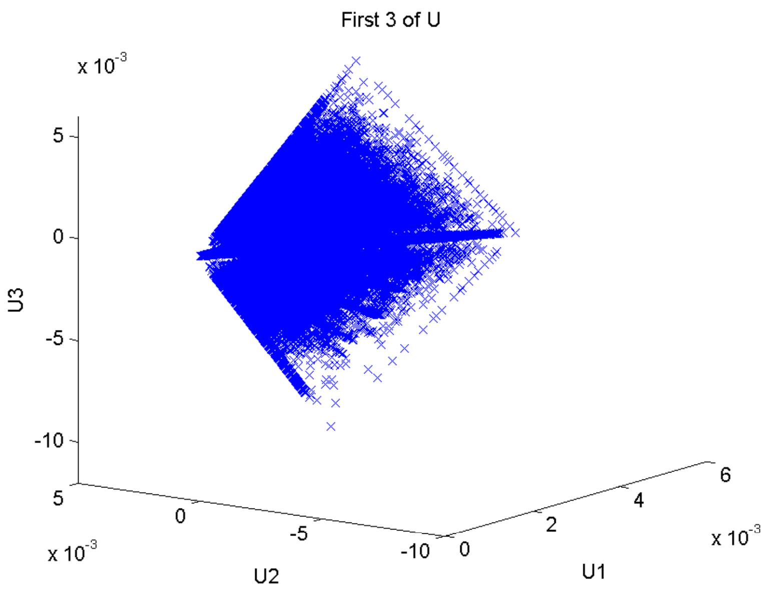

Figure 4.2: Another view: SVD plot in 3 dimensions with one point for each email. This email-deception-cue-matrix version had only seven columns; one for each of the seven distinct latent factors of the deception model.

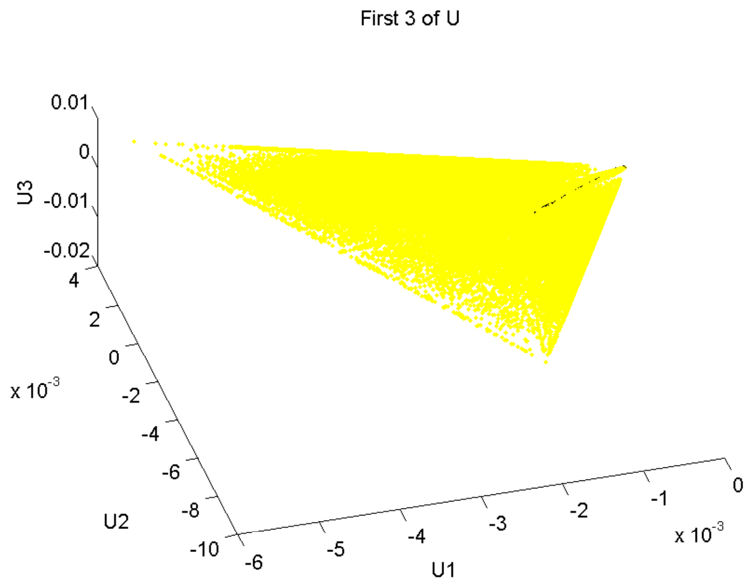

Figure 4.3: SVD plot in 3 dimensions with one point for each email. The most deceptive messages according to the Pennebaker Deception Score are plotted in black.

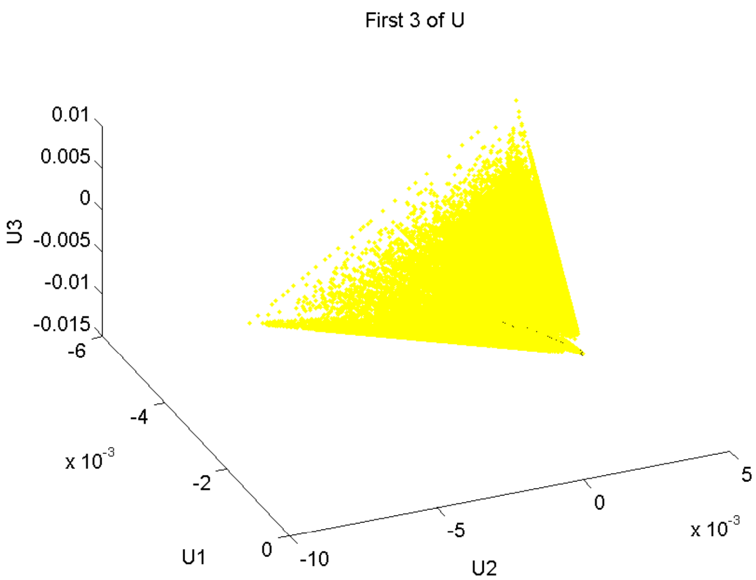

Figure 4.4: Another view: SVD plot in 3 dimensions with one point for each email. The most deceptive messages according to the Pennebaker Deception Score are plotted in black.

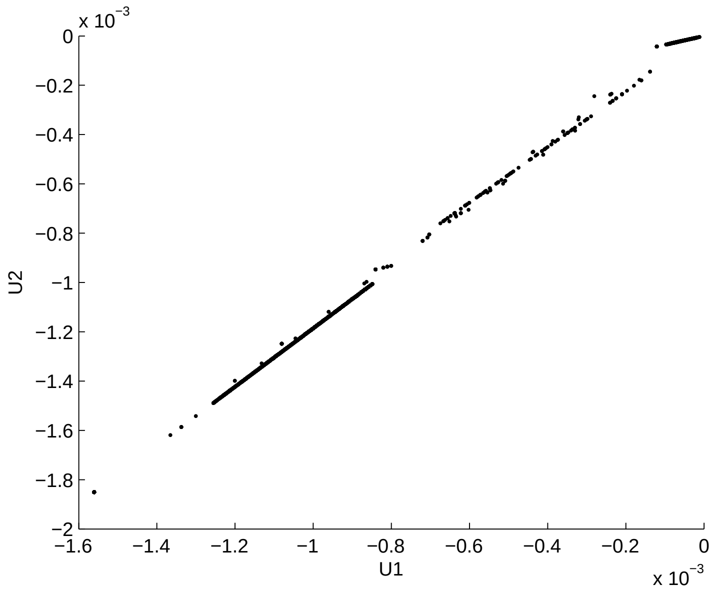

Figure 4.5: SVD plot in 2 dimensions, only the most deceptive messages according to the Pennebaker Deception Score have been plotted.

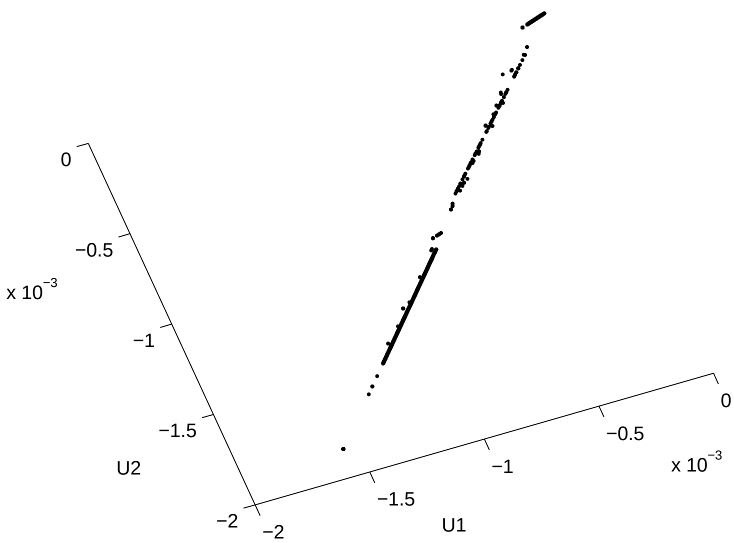

Figure 4.6: Another view: SVD plot in 2 dimensions, only the most deceptive messages according to the Pennebaker Deception Score have been plotted.

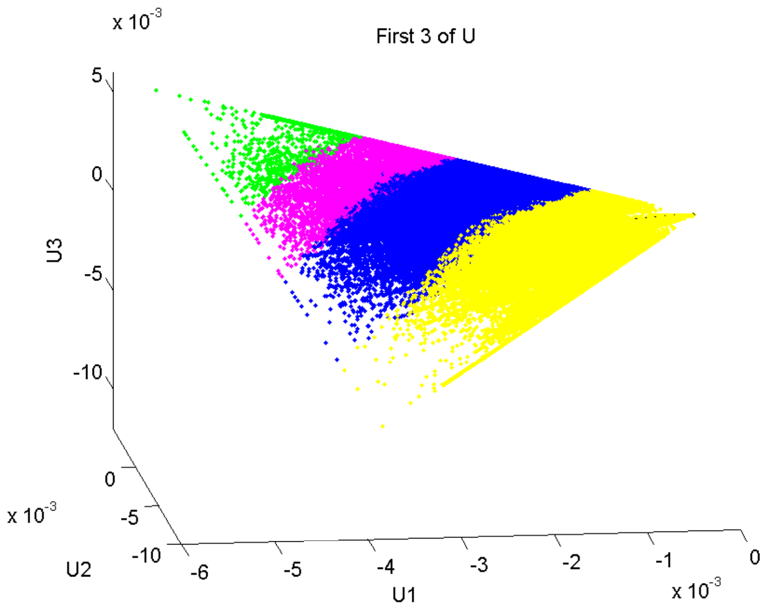

Figure 4.7: SVD plot in 3 dimensions with one point for each email. Messages ranked according to the Pennebaker Deception Score – from least to most deceptive in green, pink, blue, yellow and black respectively.

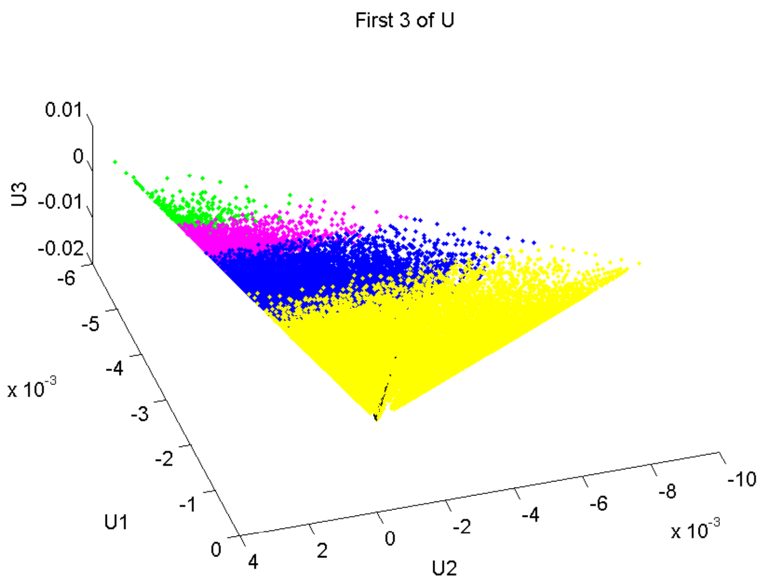

Figure 4.8: View 2: The figure is the same as Figure 4.7, but from another angle. Notice the streak of messages that are being pulled away from the origin of the fan-shaped structure and towards its periphery; along the spike identified as representing the exclusive cue-word 'or'.

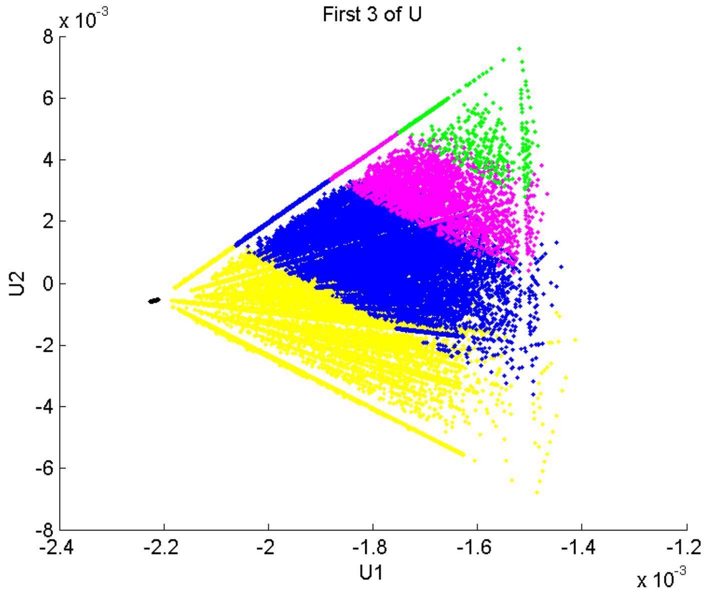

Figure 4.9: View 3: SVD plot in 2 dimensions with one point for each email. Messages ranked according to the Pennebaker Deception Score – from least to most deceptive in green, pink, blue, yellow and black respectively.

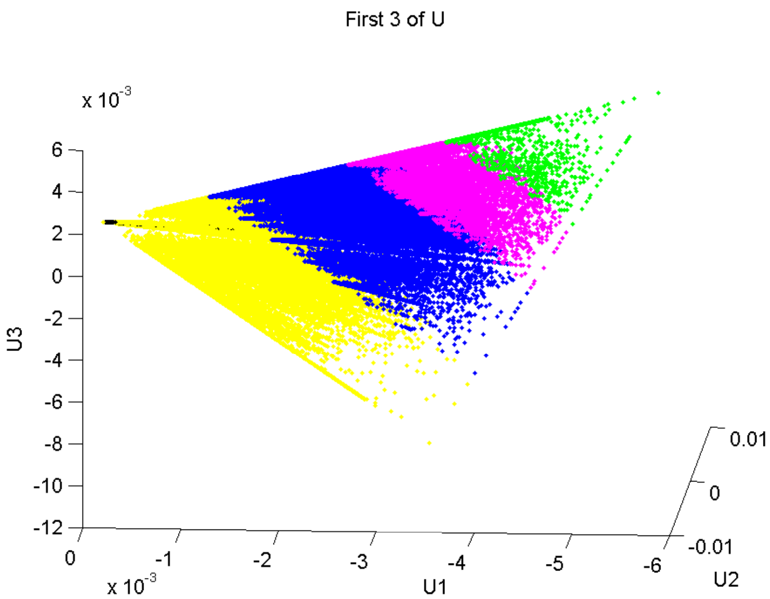

Figure 4.10: View 4: SVD plot in 3 dimensions with one point for each email. Messages ranked according to the Pennebaker Deception Score – the boundaries between the coloured bands remain unclear and fuzzy.

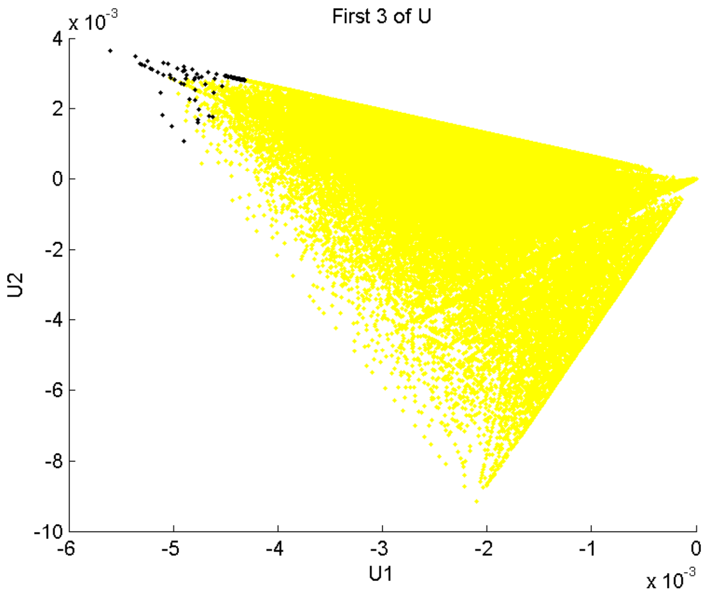

Figure 4.11: SVD plot in 2 dimensions with one point for each email. The least deceptive messages according to the Pennebaker Deception Score are plotted in black.

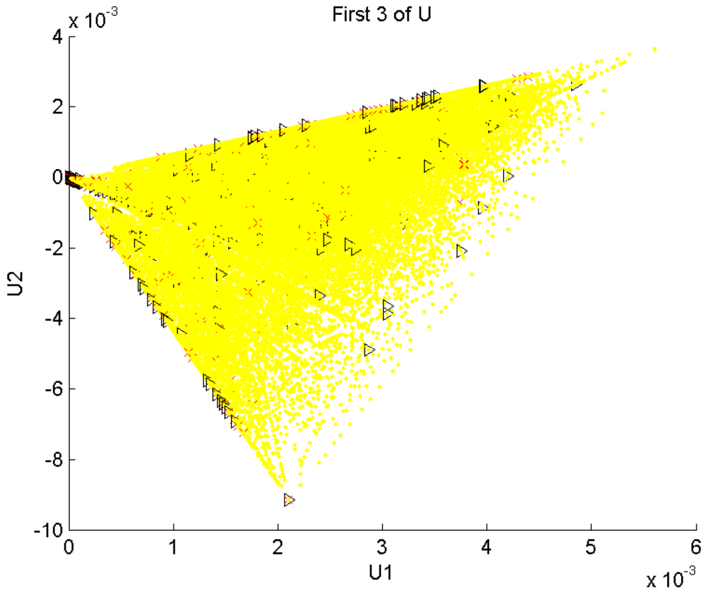

Figure 4.12: SVD plot in 2 dimensions with one point for each email. Kenneth Lay and Jeffrey Skilling's emails highlighted in red and black respectively.

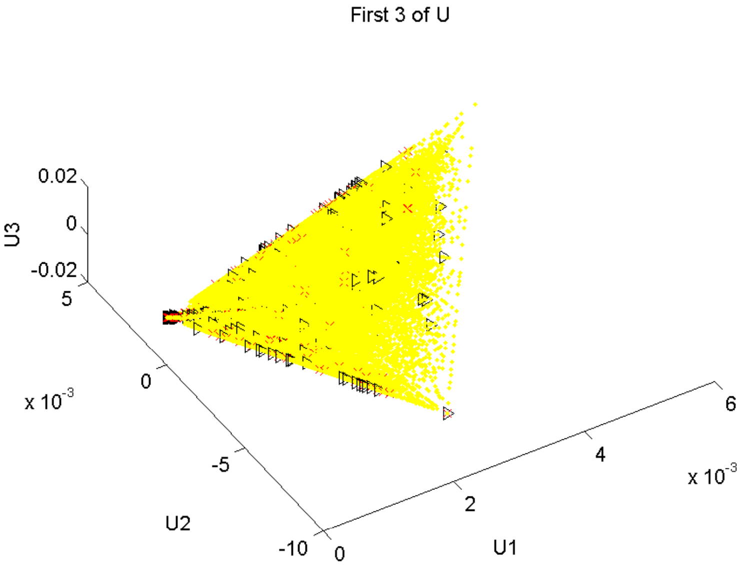

Figure 4.13: Another view: SVD plot in 3 dimensions with one point for each email. Kenneth Lay and Jeffrey Skilling's emails highlighted in red and black respectively.

Chapter 5

Conclusions and Limitations

5.1 Conclusions

Through this research work we have highlighted how unsupervised date mining, in specific the Singular Value Decomposition (SVD) technique, can be used to devise a descriptive model for deception detection in text. The model can be used as an effective tool for flagging unusual and deceptive emails or ranking emails along a scale of relative deceptiveness (say from negotiation and spin to deception and blatant lying). It may be used in intelligence, law enforcement and within organizations in order to detect fraud, white-collar crime or employee malfeasance of various kinds.

As an important pre-processing step to handling new dataset(s) we have detailed a normalization method (Section 3.3) which is able to successfully capture the significant non-zero features of a matrix; regardless of how sparse it is. The log transformation used during the normalization process is an essential step since the usage-frequencies of the cue-words in the deception model are quite different - some words are very common, while others are rather rare. Also, we have been able to define a method to over-ride the external effect of the message text-length on its deception-cue-words usage frequency profile.

After exploring both approaches to refining the deception model (that is increasing the number of deception-cue-words or decreasing the number of cue-words), we conclude that the approach of reducing the cue-words is more effective. By successfully removing 13 words from the Pennebaker cue-wordlist (Section 3.5) we have not only been able to shed more light on the deception model itself, but have also helped to reduce the resources required to analyze other new datasets.

It is clear from our work that each of the four categories of cue-words in Pennebaker's Deception Model behaves as a distinct latent factor. Of these the two classes, first-person pronouns and exception words, are much more important than the others. In fact, the entire deception cue-wordlist can be divided into seven unique driving forces (Section 3.5) – namely, the first person pronoun word 'I'; the remaining set of first-person pronouns; the exclusive word 'or'; the exclusive word 'but'; the remaining set of exclusive words; action words; and negative emotion words.

We have shown how the inductive model, based on applying the SVD technique, helps in weighting each of the cue attributes to best reflect their differential use in deception detection. This is a step forward from the results obtained solely based on the *Pennebaker Deception Score* (PDS) which simply sums up the deception-cue-words' frequencies into a single score and assigns an equal weight to all the attributes in the cue-list. The PDS may give misleading results based on the increased/decreased frequency of a single cue-word. Addition of the SVD correlation analysis is important because not all the attributes in the deception-cue-wordlist are equally significant.

By using unsupervised data mining, we have tried both to understand deceptive communication as it is qualitatively different than truthful communication and highlight the differential use of multiple cue-words/cue-word categories in the process of deception detection. The descriptive model underscores the multivariate profile of deception based on the usage-frequencies of multiple cue-words.

The strength of this deception model lies in the fact that, language generation being a subconscious process, liars can to some extent control the content of their stories, but not their style of language use [30, 37, 43]. Many of the linguistic markers of deception remain consistent across situations. In fact, any attempt to dodge surveillance by omitting or substituting deception-cue-words only leads to a substitution signature of ‘excessive blandness’ [41]; thus making it almost impossible to over-ride the characteristic signature of deception.

Also, since many sensitive issues of privacy need to be kept in mind while dealing with email and other forms of communication, the automated deception model provides an unintrusive and efficient solution. It is cost-effective since no human-reader(s) are required to parse the messages. Human judges are, also, far more likely to make ‘false’ identifications [31, 37] since humans are not naturally good at deception detection.

5.2 Limitations

As mentioned in Section 3.5.1, though the exclusive words cluster can be clearly seen in Figure 3.18 there are members of the category on both sides of the origin indicating that some of these words are good pointers of deceptiveness while others are not. The same is the case for the negative emotion words; which remain clustered close to the origin (Figure 3.20). Clearly, some words in each of these classes are more significant than the others.

These results suggest some refinements to the deception model, including finding the significant cue-words (and/or the deceptive quadrant with reference to the origin) and removing the other exclusive words and negative emotion words from the deception-cue-wordlist. The reason for removing words from the cue-list should be well-grounded for some words might be more relevant in a particular context than in another (for instance, some words might not occur much in ordinary business-related communication).

Certainly, the model can be generalized/customized to be used in any domain – ranging from intra-organizational monitoring to detecting white-collar crime over the internet to government vigilance applications for counterterrorism and smuggling. The customization for a domain may require slight fine-tuning of the base linguistic model to be more suitable for the specific field (some words are more likely to be used in business, financial fields – for instance, the word 'loss').

A good application might be to integrate the textual deception detection model with the Echelon [1] intelligence network to intercept communication of interest. Though majority of the messages would undergo only limited inspection by the automated system, those flagged as unusual/deceptive/interesting can then be sent for further human analysis and scrutiny.

At this point it is worth mentioning that, despite various studies in the field of psychology, deception detection still remains an inexact science. There exists no universal, fault-proof method for detecting deception. Unfortunately, the Pennebaker empirical deception model (LIWC) [38] completely ignores "syntax and context and linguistic devices such as irony and sarcasm" [39]. There remains the want for more refined, sophisticated and accurate language analysis programs to understand the linguistic manifestations of people's mental state(s) and what exactly they are thinking or feeling.

Computerized text analysis can help us to quickly and efficiently summarise varied and complex language samples. This can not only aid psychologists in better understanding natural language use, but also provides an excellent gateway to explore and correlate the psychology of natural language usage across different domains in everyday life.

Bibliography

[1] European Parliament Temporary Committee on the ECHELON Interception System. Final report on the existence of a global system for the interception of private and commercial communications (echelon interception system), 2001.

[2] Arthur Anderson: Dishonesty, Greed and Hypocrisy in Corporate America, retrieved November, 2007. http://www.commondreams.org/views02/0712-02.htm.

[3] Dabhol Power Company, retrieved November, 2007. http://en.wikipedia.org/wiki/Dabhol_Power_Company.

[4] Data Clustering: A Review, retrieved November, 2007. http://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf.

[5] David Corney's DoClustering Script, retrieved October, 2007. http://www.oup.com/oald-bin/web_getald7index1a.pl.

[6] Definition of Normalization, retrieved October, 2007. http://www.hud.gov/offices/pih/programs/ph/phecc/definitions.cfm/.

[7] Definition of Unsupervised Data Mining, retrieved October, 2007. http://www.thearling.com/glossary.htm.

[8] Detecting Deception, retrieved October, 2007. http://www.apa.org/monitor/julaug04/detecting.html.

[9] Enron: Company Information, retrieved November, 2007. http://en.wikipedia.org/wiki/Enron.

[10] Enron: Company Website, retrieved November, 2007. http://www.enron.com/corp/.

[11] Enron Email Dataset, retrieved October, 2007. http://www.cs.cmu.edu/~enron/.

[12] Enron: From Collapse to Convictions – A Timeline, retrieved November, 2007. http://www.cbc.ca/news/background/enron/.

[13] Enron: Gas Bank, retrieved November, 2007. http://www.riskglossary.com/link/enron.htm.

[14] Hierarchical Clustering, retrieved November, 2007. http://www.resample.com/xlminer/help/HClst/HClst_intro.htm.

[15] K-Means Clustering, retrieved November, 2007. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html.

[16] NMSU 'psychological sleuth' works to Detect Deception, retrieved October, 2007. http://www.nmsu.edu/~ucomm/Releases/2004/July/deception_detection.html.

[17] Oxford Dictionary, retrieved October, 2007. http://www.oup.com/oald-bin/web_getald7index1a.pl.

[18] Random Forests, retrieved October, 2007. `http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm`.

[19] Rebecca Mark: former head of the Enron International Division, retrieved November, 2007. `http://en.wikipedia.org/wiki/Rebecca_Mark-Jusbasche`.

[20] Sarbanes-Oxley Act, retrieved October, 2007. `http://www.sarbanes-oxley.com/`.

[21] Singular Value Decomposition, retrieved November, 2007. `http://en.wikipedia.org/wiki/Singular_value_decomposition`.

[22] Singular Value Decompositon Technique, retrieved October, 2007. `http://www.kwon3d.com/theory/jkinem/svd.html`.

[23] Statement Analysis, retrieved October, 2007. `http://www.statementanalysis.com/`.

[24] Unsupervised Data Mining, retrieved October, 2007. `http://en.wikipedia.org/wiki/Unsupervised_learning`.

[25] A.D. Beresford. WCC Issue: Securities/Investment Fraud. Technical report, National White Collar Crime Centre, 2003.

[26] D.P. Biros, J. Sakamoto, J.F. George, M. Adkins, J. Kruse, J.K. Burgoon, and Jr. J.F. Nunamaker. A quasi-experiment to determine the impact of a computer based deception detection training system: The use of Agent 99 trainer in the US military. In *Proceedings of the 38th Hawaii International Conference on Systems Science*, volume 1, 2005.

[27] R.J. Boyle and C.P. Ruppel. The Impact of Media Richness and Suspicion and Perceived Truth Bias on Deception Detection. In *Proceedings of the 38th Hawaii International Conference on Systems Science*, 2005.

[28] B.M. DePaulo, J.J. Lindsay, B.E. Malone, L. Muhlenbruck, K. Charlton, and H.Cooper. Cues to Deception. In *Psychology Bulletin*, pages 9:74–118, 2002.

[29] P. Ekman. Why Lies Fail and What Behaviors Betray A Lie. In *J.C. Yuille (Ed.) Credibility Assessment*, pages 71–81, 1989.

[30] P. Ekman and M.G. Frank. Lies that Fail. In *M. Lewis and C. Saami (Eds.) Lying and deception in everyday life*, pages 184–200, 1993.

[31] P. Ekman, M. O'Sullivan, and M. Frank. A Few Can Catch A Lair. In *Psychological Science*, pages 10:263–266, 1999.

[32] SW. Fong, D.B. Skillicorn, and D. Roussinov. Detecting Word Substitution in Adversarial Communication. In *Workshop on Link Analysis, Counterterrorism, and Security, SIAM International Data Mining Conference*, 2006.

[33] SW. Fong, D.B. Skillicorn, and D. Roussinov. Measures to Detect Word Substitution in Intercepted Communication. In *IEEE Conference on Intelligence and Security Informatics, ISI 2006*, pages 190–200, 2006.

[34] D.J. Hand and P. Smyth H. Mannila. *Principles of Data Mining.* MIT Press, 2001.

[35] P.S. Keila. Knowledge Discovery and Deception Detection in the Enron Email Corpus. MSc thesis, Queen's University, Kingston, ON, Canada, January 2006.

[36] P.S. Keila and D.B. Skillicorn. Detecting Unusual Email Communication. In *CASCON 2005*, pages 238–246, 2005.

[37] M.L. Newman, J.W. Pennebaker, D.S. Berry, and J.M. Richards. Lying Words: Predicting Deception from Linguistic Styles. In *Personality and Social Psychology Bulletin*, pages 29:665–675, 2003.

[38] J.W. Pennebaker, M.E. Francis, and R.J. Booth. Linguistic Inquiry and Word Count (LIWC). Technical report, Erlbaum Publishers, 2001.

[39] J.W. Pennebaker and A. Graybeal. Patterns of Natural Language Use: Disclosure, Personality, and Social Integration. *Bulletin of the American Psychological Society*, 2001.

[40] D.B. Skillicorn. Applying Matrix Decomposition to Counterterrorism. Technical report, Queens University, 2004.

[41] D.B. Skillicorn. Beyond Keyword Filtering for Message and Conversation Detection. In *IEEE International Conference on Intelligence and Security Informatics, ISI 2005*, pages 231–243, 2005.

[42] D.B. Skillicorn. *Understanding Complex Datasets: Data Mining with Matrix Decompositions.* CRC Press, 2007.

[43] D.P. Twitchell, Jr J.F. Nunamaker, and J.K. Burgoon. Using Speech Act Profiling for Deception Detection. In *Intelligence and Security Informatics: Second Symposium on Intelligence and Security Informatics, ISI 2004*, pages 403–410, 2004.

[44] L. Zhou, J.K. Burgoon, Jr J.F. Nunamaker, and D.P. Twitchell. Automating Linguistic-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communication. In *Group Decision and Negotiation*, pages 13:81–106, 2004.

[45] L. Zhou, D.P. Twitchell, T. Qin, J.K. Burgoon, and Jr J.F. Nunamaker. An Exploratory Study into Deception Detection in Text-Based Computer Mediated Communication. In *Proceedings of the 36th Hawaii International Conference on Systems Science*, 2003.

Appendix A

Sample Email Messages

With reference to Figure 4.1, emails plotted close to the origin of the fan-shaped structure represent mostly some form of formal and often legal correspondence – like a list of things to keep in mind while signing contracts; instructions from people of authority *etc.* Messages found elsewhere in the plot (apart from the origin) are informal and sometimes profane.

We have at times listed only an excerpt from the email because the actual email message is too long. The objective is to give the reader an idea of the 'type' of email content. Sender/recipient names have not been included since there are many sensitive issues that need to be taken care of with regard to the privacy of the people; most of whom were certainly not involved in any of the actions which precipitated the Enron Investigation.

Email (1):
With reference to Figure 4.1, this message is from the region close to the origin of the fan-shaped structure. The message is formal/technical/business-oriented in its content, language use and style.

——Subject: ECA Hearing Summary——
The new Assembly Committee on Energy Costs and Availability(ECA) met today and heard two bills: ABX 5 (Keeley), relating to the ISO and PX governing boards, and ABX 6 (Dutra/Pescetti), relating to utility retention of assets. Both measures passed with only one dissenting vote on ABX 6 (Asm. John Campbell).
Although ABX 5 passed in its current form, possible amendments to be taken in the Senate would include: increasing the number of governing board members, increasing the length of the terms, and addressing the issue of Governor appointees on the EOB confirming Governor appointees on the ISO and PX. Testifying in support of the measure were the Consumers Union, EOB, CalPIRG, TURN, ORA and the California Utilities Employees. The ISO testified in opposition based on a past FERC letter which opposes an appointed Board.
The ABX 6 debate was over the conflict in the Public Utilities Code regarding utility asset retention (Sections 851 and 377). The debate, however, was not limited to this one point. Asm. Leonard started the ball rolling when he noted a problem when some generation in the state is regulated (the utility generation) and some isn't (the new generators). This issue was picked up by Asms. Steinberg, Jackson, Oropeza, and Diaz, who suggested that there may need to be legislation to regulate all the generation in the state. Asm. Wright did a good job pointing out that in doing so, the state would have to offer generator owners something in return the state isn't able to afford.
When Loretta Lynch spoke in support of the bill she began with the Stage 3 alert, noting there was 15,000 megawatts off-line today and generators were refusing to allow state inspectors into their plants. Although the only company she named was AES, she made every committee member believe it was everyone. In response to a question about long-term contracts, she said that since August this CPUC has allowed it and the utilities are the ones who refuse to enter into contracts.
Pete Conaty from labor repeated his statement that when the new generators purchased these plants, they significantly reduced the labor force, ran the plants harder than when the utilities ran them and that is why the plants are breaking down.
Both measures will be debated on the Assembly floor next when they go into session tomorrow morning.

Email (2):
With reference to Figure 4.1, this message is from the region close to the origin of the fan-shaped structure. Keeping in mind, the fact, that Enron [9] filed for bankruptcy protection in late 2001, we can only wonder if this email 'truthfully' describes the status of the company in October 2001.

——Subject: Message Points on Current Status of Company——
Message Points on Current Status of Company
A. Our core businesses are strong, performing well and customers continue to demonstrate their confidence in us. B. EnronOnline experienced above average levels for number of transactions Wednesday (10/24/01), Enron recorded more than 8000 transactions with 1399 external users for a gross notional value of approximately dollar 4 billion We continue to be the market maker of choice in wholesale gas and power C. Our gas and power numbers indicate that our customer base is not withdrawing, closing out positions, nor reducing transaction levels with us. Indeed, customers are transacting at higher levels as they prepare their portfolios for the upcoming winter and protect themselves from price volatility.
Media Q and A
What did you receive from the SEC? Enron received a letter and subsequent phone call requesting that we voluntarily provide information regarding certain related party transactions.
When did you receive the request from the SEC? We received a fax on Wednesday afternoon, October 19, 2001, and a phone call on Thursday afternoon, October 20, 2001.
What did you do about the request from the SEC when you received it? An informal inquiry is not a material event; however, because of the high visibility of Enron and the recent public spotlight on the transaction, we convened a meeting of our Board as soon as possible. We recommended that we announce the SEC request; the Board agreed, and we made the release prior to the market opening on Monday, October 22, 2001.
What are you going to do about the request? We are cooperating fully with the SEC and see the request as an opportunity to put this issue behind us.
These message points will be constantly changing and we will update you as the changes occur.
Also, attached below is the Enron link for the Press Room that contains "Frequently Asked Questions".
http://www.enron.com/corp/pressroom/faq.html

Email (3):
With reference to Figure 4.1, this message is from the region close to the origin of the fan-shaped structure. The message is formal/technical/business-oriented in its content, language use and style. The reference to 'Restrictive and Unfair Trade Practices' in India may be interesting, considering former head of the Enron International division, Rebecca Mark's [19], infamous venture with the Dabhol power plant in India [3]. Note that, "Allegations of bribery and human-rights abuses (never proven) against Enron led to riots and the cancellation of the plant in August 1995. Enron spent USD 900m on Dabhol and lost most of it" [19].

——Subject: Restrictive and Unfair Trade Practices.——
Here is the response from India, regarding the need to seek competition authority approval. As you will see, Jane's response focusses on registering agreements, but her conclusion is that the focus would be at the Foreign Investment Board. Let me know if you need more. mcs —————— Forwarded
–Subject: Restrictive and Unfair Trade Practices.
Nothing that I see on basic market power. I think they would focus on the FIPB requirements.
S.33 Registrable agreements relating to restrictive trade practices. (1) Every agreement falling within one or more of the following categories shall be deemed, for the purposes of this Act, to be an agreement relating to restrictive trade practicews and shall be subject to registration.
(a) agreement which restricts or is likely to restrict, by any method the persons or classes of persons to whom goods are sold or from whom goods are bought. (b) any agreement requiring a purchaser of goods as a condition of such purchase to purchase some other goods. (c) any agreement restricting in any manner the purchaser in the course of his trade from acquiring or otherwise dealing in any goods other than those of the seller or any other person. (d) any agreement to purchase or sell goods or to tender for the sale or purchase of goods only at prices or on terms or condtions agreed upon between the sellers or purchasers. (e) any agreement to grant or allow concessions or benefits, including allowances, discounts, rebates or credit in connection with, or by reason of dealings. (f) any agreement to sell goods on condition that the prices to be charged on resale by the purchaser shall be the prices stipulated by the seller unless it is clearly stated that prices lower than those prices may be charged.....*etc. etc. etc.*

Email (4):
With reference to Figure 4.1, this message is from the region close to the origin of the fan-shaped structure. The message is formal/technical/business-oriented in its content, language use and style.

—–Subject: CAISO Notice – Congestion Management Reform Proposal Posting——
Market Participants:
The CAISO has posted its draft Congestion Management Reform (CMR) recommendation on its web site at http://www.caiso.com/clientserv/congestionreform.html ¡http://www.caiso.com/clientserv/congestionreform.html. This document recommends changes to the CAISO's current Congestion Management process and related features of its business practices and operations. This is a draft document, albeit one that we believe reflects significant effort and thought on the part of both the stakeholders who provided the initial input and the interdisciplinary design team that drafted this document. This CMR recommendation represents an essential milestone in the broader Congestion Management Reform Project. The first stage involved soliciting stakeholder input regarding the problems to be solved, alternative solutions to address these problems, and criteria for evaluating reform proposals. The second stage involved the CAISO's crafting an integrated, internally consistent reform package utilizing the ideas developed with and/or by Stakeholders in stage one. This document contains the CAISO's draft recommendation. We emphasize that, although we believe this recommendation package to be a necessary milestone in the Congestion Management Reform Project, it is not intended to predispose the final design. We are actively soliciting Stakeholders' comments and critiques concerning this recommendation over the coming weeks (stage three of the CMR project).......*etc. etc. etc.*

Email (5):
With reference to Figure 4.1, this message is from the region close to the origin of the fan-shaped structure. The message is formal/technical/business-oriented in its content, language use and style.

—–Subject: Fielder needs to cancel Our Meeting——
Called Fielder to confirm and he said that he needed to cancel. As you know, the Legislature closed without doing anything with Edison's bailout legislation. The Governor has said that the will order them back into session in two weeks. That has Fielder and everyone else at Edison scrambling to continue their attempts to stay out of bankruptcy, which means that Fielder won't be focused, and can't focus on, anything else for the next couple of weeks. Let's regroup. Dave, perhaps we should just go ahead and start with PGandE. Thoughts?
Best, Jeff

Email (6):
With reference to Figure 4.1, this message is from the region close to point A. The message is informal/personal in its content, language use and style. Notice the high usage-frequency of first-person pronouns; especially the personal pronoun 'I'.

—–Subject: Good Morning—-
Got it, let's talk tonight. What time is good to call you? or you give me a call!
I appreciate all that you have done. There are 7 days remaining. I wish I was leaving tomorrow. I miss him dearly. Three to four years is going to be a challenge. In that time frame I must finish school and both of my daughters would have graduated from high school and 8th grade. The apartment was bothering me on yesterday. It was laid on my heart to call her and let her know everything up front. She had to run a credit report, you know that is not good due to the circumstances. I let her also know why I can not move until March 1. Emotionally I wish I could move Feb.1. Financially I would be set to move and start my new life of March 1. I prayed about on last night and turned it over to JESUS and I am not to worry about it. I see the lady on today. My whole life will be on the move in the month of Feb. If all goes well I should receive the package no later than Tuesday. Keep your head up about the class. If you need to pull Hunter to the side and work with him personally do that. May be L can help? Beside missing RC I am okay.

—–Original Message—– Subject: Good Morning :)
Hey "G", what's up? I know you will be shocked, but I brought the purse and necklace into the office and will be sending them out this afternoon....so there! How's it going this morning?
I have been sitting in a class all morning that Hunter is teaching. It's on options and volatility. Boy, that stuff does not sink in well with me, I find the class a real challenge. I'm so uncomfortable in the class. Pray for me, because I need to get this stuff and be able to chart this stuff. Everyone is real helpful though, but being put on the spot, its hard to come up with the answers quickly. It's painful for me....whew!
Talk with you later.......how many days left now?
Love ya ......*etc. etc. etc.*

Email (7):
With reference to Figure 4.1, this message is from the region close to point A. The message is informal/personal in its content, language use and style. Notice the high usage-frequency of first-person pronouns; especially the personal pronoun 'I'.

—–Subject: Fw: (no subject)——
Michael is the kind of guy you love to hate. He is always in a good mood and always has something positive to say. When someone would ask him how he was doing, he would reply, "If I were any better, I would be twins!"
He was a natural motivator. If an employee was having a bad day, Michael was there telling the employee how to look on the positive side of the situation.
Seeing this style really made me curious, so one day I went up to Michael and asked him, "I don't get it! You can't be a positive person all of he time. How do you do it?"
Michael replied, "Each morning I wake up and say to myself, you have two choices today. You can choose to be in a good mood or...you can choose to be in a bad mood. I choose to be in a good mood. Each time something bad happens, I can choose to be a victim or...I can choose to learn from it. I choose to learn from it. Every time someone comes to me complaining, I can choose to accept their complaining or...I can point out the positive side of life. I choose the positive side of life.
"Yeah, right, it's not that easy," I protested.
"Yes, it is," Michael said. "Life is all about choices. When you cut away all the junk, every situation is a choice. You choose how you react to situations. You choose how people affect your mood. You choose to be in a good mood or bad mood. The bottom line: It's your choice how you live your life."
I reflected on what Michael said. Soon hereafter, I left the Tower Industry to start my own business. We lost touch, but I often thought about him when I made a choice about life instead of reacting to it.
Several years later, I heard that Michael was involved in a serious accident, falling some 60 feet from a communications tower. After 18 hours of surgery and weeks of intensive care, Michael was released from the hospital with rods placed in his back.
I saw Michael about six months after the accident. When I asked him how he was, he replied. "If I were any better, I'd be twins. Wanna see my scars?"
I declined to see his wounds, but I did ask him what had gone through his mind as the accident took place. "The first thing that went through my mind was the well-being of my soon to be born daughter, " Michael replied. "Then, as I lay on the ground, I remembered that I had two choices: I could choose to live or ...I could choose to die. I chose to live." ......*etc. etc. etc.*

Email (8):
With reference to Figure 4.1, this message is from the region close to point A. The message is informal/personal in its content, language use and style. Notice the high usage-frequency of first-person pronouns; especially the personal pronoun 'I'.

——Subject: (no subject)——
These are the collected writings of the Simpsons from the chalkboard exercises that Bart writes during the opening credits.
I will not carve gods. I will not spank others. I will not aim for the head. I will not barf unless I'm sick. I will not expose the ignorance of the faculty. I saw nothing unusual in the teacher's lounge. I will not conduct my own fire drills. Funny noises are not funny. I will not snap bras. I will not fake seizures. This punishment is not boring and pointless. My name is not Dr. Death. I will not defame New Orleans. I will not prescribe medication. I will not bury the new kid. I will not teach others to fly. I will not bring sheep to class. A burp is not an answer. Teacher is not a leper. Coffee is not for kids. I will not eat things for money. I will not yell "She's Dead" at roll call. The principal's toupee is not a Frisbee. I will not call the principal "spud head". Goldfish don't bounce. Mud is not one of the 4 food groups. No one is interested in my underpants. I will not sell miracle cures. I will return the seeing-eye dog. I do not have diplomatic immunity. I will not charge admission to the bathroom. I will never win an emmy. The cafeteria deep fryer is not a toy. All work and no play makes Bart a dull boy. I will not say "Springfield" just to get applause. I am not authorized to fire substitute teachers. My homework was not stolen by a one-armed man. I will not go near the kindergarten turtle. I am not deliciously saucy. Organ transplants are best left to professionals. The Pledge of Allegiance does not end with "Hail Satan". I will not celebrate meaningless milestones. There are plenty of businesses like show business. Five days is not too long to wait for a gun. I will not waste chalk. I will not skateboard in the halls. I will not instigate revolution. I will not draw naked ladies in class. I did not see Elvis. I will not call my teacher "Hot Cakes". Garlic gum is not funny. They are laughing at me, not with me. I will not yell "Fire" in a crowded classroom. I will not encourage others to fly. I will not fake my way through life. Tar is not a plaything. I will not Xerox my butt. It's potato, not potatoe. I will not trade pants with others. I am not a 32 year old woman. I will not do that thing with my tongue. I will not drive the principal's car. I will not pledge allegiance to Bart. I will not sell school property. I will not burp in class. I will not cut corners. I will not get very far with this attitude. I will not belch the National Anthem. I will not sell land in Florida. I will not grease the monkey bars. I will not hide behind the Fifth Amendment. I will not do anything bad ever again. I will not show off. I will not sleep through my education. I am not a dentist....*etc. etc. etc.*

Email (9):
With reference to Figure 4.1, this message is from the region close to point C. The message uses a formal/semi-formal style of writing. The content is more generic and administrative in nature as compared to messages found close to the origin.

—–Subject: Anthrax and other Biological Agent Threats——
The information below is an excellent overview of the procedures our administrative personnel will be using when handling and distributing your mail into mail folders. However, all administrative personnel in the Portland Enron WTC offices have received a hard copy of the guidelines and procedures issued by Oregon Department Administrative Services. This four-page document has detailed guidelines and instructions for mail/parcel handling. If you would like to get a copy of these guidelines, ask your administrative assistant or you can contact Debra Davidson to get a copy. The World Trade Center Mailroom has informed Debra that they are also screening mail for lumps, powdery substances, unusual written comments on envelopes, etc. If they detect a suspicious piece of mail or package addressed to our office, they will notify us as well as the proper authorities. If you have any questions or concerns, please see me or Debra Davidson.
—–Original Message—– Subject: Anthrax and other Biological Agent Threats
We understand that the recent cases of Anthrax contamination and the possibility that biological agents such as Anthrax may be used in a terrorist attack are raising great concerns among Enron employees. Many government and commercial facilities in the United States have already received Anthrax threat letters containing powdery substances. Most of these have, however, been determined to be false alarms. We have no reason to believe that Enron has been or will be the target of an Anthrax attack, but we want to provide all employees with background information on Anthrax and up-to-date guidance for handling any possible Anthrax exposures.
If you have additional questions or concerns, please contact Corporate Security in Houston at (713) 345-2804 or via email at CorporateSecurity@enron.com.
The most important thing to remember is: Do not panic. To infect someone, the Anthrax organism must be rubbed into abraded skin, swallowed, or inhaled as a fine, aerosolized mist. Infection can be prevented after exposure to Anthrax by early treatment with the appropriate antibiotics. Anthrax cannot be spread from one person to another.
Following are guidelines for identifying and dealing with suspicious letters or packages:
What constitutes a suspicious letter or parcel? (Remember, these are only guidelines. Use your best judgment when determining if a letter or package is suspicious.)
– It is marked with the word "Anthrax." – It has a non-identifiable powdery substance on the outside. – It is unexpected or from someone unfamiliar to you. – Is addressed to someone no longer with your organization or is otherwise outdated. – Has no return address, or has one that cannot be verified as legitimate......*etc. etc.*

Email (10):
With reference to Figure 4.1, this message is from the region close to point C. The message uses a formal/semi-formal style of writing. The content is more generic and administrative in nature as compared to messages found close to the origin.

——Subject: Williams Energy News Live – today's video newscast——
Dear Andrew,
Senate Energy Committee Chairman Jeff Bingaman (D-NM) says he is unsure whether an energy bill will reach the Senate floor this year. Washington Bureau Chief Peter Cook says Senator Bingaman is also downplaying the chances that a new Republican energy proposal will alter the debate significantly. The bureau will have the latest from the Hill on Friday.
Also from Washington on Friday, Executive Director of the International Energy Agency Robert Priddle will discuss the world's energy supply within the context of the campaign against terrorism. Priddle is speaking at the Center for Strategic and International Studies tomorrow, and the Washington bureau will bring us all the details.
Energy Markets Washington Editor Jack Belcher will join us from the Washington bureau on Friday. Belcher is stopping by during the 9:00 a.m. ET newscast to discuss the EPA's new boutique fuels proposal.
The Independent Petroleum Association of America will continue a three-day meeting in Houston on Friday. Interior Secretary Gale Norton and Energy Secretary Spencer Abraham are among those attending the meeting. Houston Correspondent Kim Benestante will have details on the meeting from the Houston bureau on Friday.
Also from Houston, Spinnaker Exploration CEO and Founder Roger Jarvis will discuss his company's third-quarter earnings results. Join us during the 1:00 p.m. ET newscast for the Jarvis interview. And catch Simmons and Company International Vice President of Research Mark Meyer as he talks about how oil stocks are doing this quarter. We'll bring you these interviews from Houston on Friday.
From the West Coast bureau on Friday, ENL's Kym McNicholas will bring us a report on the extra security measures being taken to protect the Hoover Dam. The dam has 17 generators, each capable of supplying electricity to 100,000 households. Catch the Hoover Dam report from the West Coast bureau tomorrow.
Keep in mind things are subject to change at a moment's notice. Occasionally guests cancel or change time slots. We'll continue to do our best to keep you updated on future interviews and events.
Be sure to watch our newscasts every business day - 9 a.m. to 5 p.m. ET, at the top of each hour.

Email (11):
With reference to Figure 4.1, this message is from the region close to point C. The message uses a formal/semi-formal style of writing. The content is more generic and administrative in nature as compared to messages found close to the origin.

—–Subject: RE: comparison of definitions - 1st in each pair is from letter agreeement—
Kay, see clarifications below. Thanks for your help.
—–Original Message—–
1. Ancillary Services means those services required by Entergy interconnection agreement with MDEA or Entergy's tariff. OR Use this one below with noted change. Ancillary Services means those services defined in Entergy's interconnect agreement with MDEA or Entergy's tariff.
2. Available Energy means Energy that is available for sale on any given day that is in excess of MDEA's Native Load. OR Available Energy means Energy that is available for sale on any given day that is in excess of (i) MDEA's Native Load, and (ii) the Energy required to be sold under any Existing Transactions, up to the total amount of Energy any day that can be produced from the Facilities.
3. Available Energy means Energy that is available for sale on any given day that is in excess of MDEA's Native Load. OR Available Energy means Energy that is available for sale on any given day that is in excess of (i) MDEA's Native Load, and (ii) the Energy required to be sold under any Existing Transactions, up to the total amount of Energy any day that can be produced from the Facilities
4. Confirmation means a confirmation of a transaction or transactions. OR Use this one below. Confirmation means the document provided for under the MPPSA or the MGPSA and with the corresponding third party under a Back-to-Back Transaction or with EPMI which specifies the Product being bought or sold, the duration of the Transaction and the other terms, including price. A daily report of all hourly (or similarly short term) purchases and sales will be provided to the Customer and shall serve as a Confirmation for those transactions under the MPPSA or MGPSA.
5. Delivery Point means the busbar of the respective Facilities located at the interconnection between CPUC's and YCPSC's respective transmission systems at the 115 kV switching station at the respective interconnections with the Entergy transmission system. Or Use this one below. Delivery Point: Point of Delivery or POD means (a) for power (i) the interfaces located at the interconnection between Clarksdale and Yazoo City transmission systems at the Entergy system, or (ii) the point specified in any Back-to-Back Transaction, EPMI Transaction or Structured Transaction at which Products are to be tendered under a Confirmation; (b) for natural gas(i) for Clarksdale, the point of interconnection between Texas Gas and Clarksdale (ii) for Yazoo City, the interconnection between Southern Natura Gas (Sonat) and Mississippi Valley Gas for the Yazoo City Power Plant (the Sonat Delivery Point) (iii) the interconnection between Mississippi Valley Gas (MVG) and the Yazoo City Power Plant (MVG Delivery Point) *etc. etc.*

Messages found elsewhere in the plot (apart from the origin/top-most/bottom-most spike in Figure 4.1) are informal and sometimes profane. Short-length messages have been selected here only due to reasons of space limitations.

Email (12):
—-Subject: FW: Kim's 30th Birthday——
Kim's Turning the Big 3-0
Next Wednesday, December 26th, will be the last day in Kim Sachtleben's life she will ever be able to say she is twenty-something. Let's welcome Kim to the 30s by having a few drinks after work on that day at the Fox and Hound. Please feel free to invite anyone that I may have left out. Even if you are on vacation that day - please come join us. Below are a few pictures of key moments in Kim's life.

Email (13):
——Subject: Sunshine——
You are my sunshine, my only sunshine you make me happy, when skies are gray you'll never know dear..how much "I love you" please don't take my sunshine away.... Love ya,

Email (14):
——Subject: Register for Flu Vaccine——
Please register me for the flu vaccine. My cost center is 105655 and my company number is 413. My extension is 36544.
Thank you.