Department of Computer Science
San Marcos, TX 78666

Report Number TXSTATE-CS-TR-2009-12

# Enron Dataset Research: E-mail Relevance Classification

Victoria VanBuren
David Villarreal
Thomas A. McMillen
Andrew L. Minnick

2009-09-25

# ENRON DATASET RESEARCH:
# E-MAIL RELEVANCE CLASSIFICATION

Victoria VanBuren[1]
David Villarreal[2]
Thomas A. McMillen[3]
Andrew L. Minnick[4]

Forensic Systems
CS 4378Y - Spring 2009
Professor Wilbon Davis[5]
Computer Science Department
Texas State University
San Marcos, TX 78666, USA

## ABSTRACT

This paper discusses a probabilistic approach to address the problem of searching through large amount of data to find case-relevant documents. Using a valuable collection of data, e-mail communications from Enron, an actual corporation, we train a Bayes-based text classifier algorithm to identify e-mails known to be case-relevant and those known to be case-irrelevant.

[1] J.D. The University of Texas School of Law, 2006; B.B.A., in Finance, Southern Methodist University, 2003; B.S. in Computer Science Candidate, 2010, Texas State University. E-mail: victoria.vanburen@yahoo.com.

[2] M.S. in Environmental Engineering, Stanford University, 1999; B.S. in Chemistry, The University of Texas at Dallas, 1997; B.S. in Computer Science Candidate, 2010, Texas State University. E-mail: davidv342@gmail.com.

[3] B.S. in Computer Science Candidate, minor in Mathematics, 2009, Texas State University. E-mail: tm1179@txstate.edu.

[4] B.S. in Computer Science Candidate, minor in Chemistry, 2009, Texas State University. E-mail: andminnick@yahoo.com.

[5] The team thanks Professor Davis for his guidance in conducting this research.

## I. INTRODUCTION

With the advent of e-mail came e-discovery. No longer do lawyers have to spend hours searching through boxes filled with of papers in musty, dark warehouses. Paper cuts have become a thing of the past. Now, discovery of electronic records has become central to litigation, particularly in litigation involving review of terabytes of digital data. After all, cases are won or lost based on admissible evidence. But in an ocean of millions of e-mail communications, locating relevant evidence has become a major challenge.

The current search method employed by most lawyers is the Boolean keyword searching. However, a recent study reveals that only 20 percent of relevant documents were found utilizing that methodology. [1] In this project, we will attempt to achieve higher accuracy levels in finding relevant documents taken from a large database. We will utilize a set of corporate e-mails messages, the "Enron Corpus," to test a text classifier filter. The dataset was made available to the public by the Federal Energy Commission ("FERC") during an investigation into Enron's involvement in the manipulation of electricity and natural gas markets. [2]

Part II of this paper provides background details on the Enron Company, describes the origins of the Enron Corpus dataset, and reviews briefly previous work done in the area of text classification. Part III presents the dbacl text classifier filter and examines the mathematical model behind its algorithm. Part IV describes the strategies employed using dbacl and presents the results of the experiments. Finally, Part V concludes the paper and suggests possible improvements to our approach.

## II. ENRON AND RELATED WORK

### *a. Enron: the Company*

Enron was an Oregon public corporation headquartered in Houston, Texas. [3] Prior to its filing of bankruptcy in December 2001, Enron was the seventh largest corporation in the United States. [4][5] In February of 2002, FERC launched a comprehensive investigation of Enron's trading activities in the California electricity markets. [6] According to FERC, EnronOnline ("EOL") gave Enron knowledge of market conditions unavailable to its competitors. [5] Enron's profits from EOL exceeded $500 million in 2000 and 2001. [5] The investigation concluded that many trading strategies employed by Enron violated the anti-gaming provisions of their FERC-approved tariffs for California. [5] Since June 2002, the U.S. Department of Justice brought criminal charges against 30 individuals, including Jeffrey K. Skilling, former President and CEO of Enron, and other top executives and energy traders. [6] The charges included conspiracy, securities fraud, and insider trading. [6]

### b. The Enron Corpus

The original Enron dataset was made public and posted to the web by FERC during its investigation into the 2000-2001 Western Energy Crisis. [2] The dataset was later purchased by Leslie Kaelbling at Massachusetts Institute of Technology where several integrity problems were identified. [7] Shortly thereafter, a team of researchers at SRI International, a non-profit corporation founded by Stanford University, lead by Melinda Garvasio did a major clean up and removal of attachments and sent it to Professor William W. Cohen at Carnegie Mellon University, who posted it on his webpage. [7] A paper analyzing the Enron database presented at a 2004 Conference concluded that the Enron corpus was "suitable for evaluation of e-mail classification methods." [8]

The Enron corpus consists of archived e-mails from Enron employees, mostly senior executives and traders. The version of the dataset that we are utilizing for this project is called the "March 2, 2004 version." This dataset contains 517,431 e-mails organized into 151 folders. The e-mails have no attachments and some e-mails have been removed upon request of Enron employees. [9]

### *c. Related Work*

At least two research studies relating to text classification have been performed on the Enron corpus. One is the automatic categorization of e-mail into folders, done by the Computer Science Department of the University of Massachusetts. [10] The other is related to social networking analysis. Utilizing the Enron corpus and court documents issued by a U.S. Bankruptcy Court, Jitesh Shetty and Jafar Adiby derived a social network constituted of 151 employees from e-mail logs, connecting individuals who have exchanged e-mails. [11]

Additionally, for the past three years, the Text Retrieval Conference Legal Track ("TREC Legal Track') team, lead by Jason Baron and Doug Oard, has been focusing on techniques for large-scale text retrieval. Prompted by the challenge of reviewing millions of e-mails in the tobacco litigation landmark case U.S. v. Phillip Morris, the team has been studying the following search techniques: Boolean, fuzzy search models, probabilistic (or Bayesian) models, statistical methods, (or clustering), machine learning approaches, categorizing tools, and social networking analysis. Researchers at TREC Legal Track found "only between 22 and 57 percent of all relevant documents cumulatively retrieved thought a variety of alternative search methods." [1]

Locating relevant documents in a large database presents unique challenges. George Paul, a Phoenix attorney expert in e-discovery and Legal Track contributor believes that the main problem in e-discovery lies on the language itself. "This is not a computer problem… Words don't stand for behavior, but are elastic and change their meaning depending on the context." [1]

## III. The dbacl Classifier

For this project, we used dbacl, a general purpose digramic Bayesian text classifier released as open source software under the terms of the GNU General Public License (GPL). Dbacl learns to classify based on categorized text documents provided by the user, and then it compares new input with the learned categories utilizing Bayesian statistical principles. By default, dbacl employs a single-word-based tokenization scheme with equal cost weightings for type I and type II errors. [12]

Bayes' well-known mathematical theorem allows the probability of one event to be given if the outcome of another event is known. The theorem is expressed in the form:

$$P(A|B) = P(B|A)\ P(A)/P(B).$$

P(A|B) is the probability of the event 'A' if you know that event 'B' has occurred. This probability is the key measure of the filter as it can make a guess based on this probability that the text is likely to belong in a category. In other words, we can take P(A|B) to give us the probability that a document belongs to the classification A or to the classification B. Research indicates that the Bayes theorem is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. [13]

## IV. Experimental Evaluation

### *a. Methodology and Strategies*

First, for training and testing purposes, we used the entire Enron corpus as a valid pool from which to pull a random sample consisting of 1,000 e-mails. Because the Enron corpus had been cleaned up by previous researchers, our team reasoned that the vast majority of e-mails present in the current dataset version were sufficiently relevant that no pre-processing of the e-mails was necessary for relevancy, including date pre-processing.

However, we decided that only e-mails in the various users' 'Sent Items' (and similar) folders were to be used for analysis. Given that every e-mail sent by an Enron employee to other Enron employee(s) should be present in both, the 'Sent Items' of an employee, as well as the Inboxes of one (or more) employees, it was considered a reasonable simplification to only utilize the 'Sent Items' folders. Likewise, since the goal was to capture the intentions of actual Enron employees, and not non-company officials, the team concluded that removing the 'Inbox' and associated folders would better be able to handle that task. However, once the 1,000 e-mails were selected for training and testing validation, the remaining e-mails in Enron corpus were no longer used.

Next, we classified the e-mails into two datasets:

- Dataset 1: A legal professional reviewed all 1,000 e-mails.
- Dataset 2: A pair of students reviewed all 1,000 e-mails. Ties were broken by the opinion of the legal professional.

In reviewing each e-mail, our objective was to find documents relevant to the issue of whether Enron manipulated electricity and natural gas markets in California and other Western states in 2000 and 2001. Accordingly, each e-mail was manually classified into two categories: relevant or irrelevant. When the reviewers were in doubt, the e-mail was classified as relevant.

Then, we trained the filter with a varying numbers of e-mails to determine an optimal number for training. We used trial scenarios where we trained the filter with 200, 500, and 800 e-mails. The e-mails used initially to train the filter were randomly selected out of the 1,000 e-mails on each run, resulting in differing sets of e-mails for each run. For each trial scenario, we run the program 5 times. The remaining e-mails of each training pile were used for testing.

Once we fed the filter with the training pile of e-mails, we then proceeded to feed the filter with the e-mails designated for testing. We calculated an accuracy level to determine the filter's ability to recognize relevant e-mails as relevant and vice-versa:

- The Relevant Accuracy is the number of known relevant e-mails that are found by the testing program divided by the total number of known relevant e-mails, and

- The Irrelevant Accuracy is the number of known irrelevant e-mails found by the testing program divided by the total number of known irrelevant e-mails.

Thus, a higher "Relevant Accuracy" measurement indicates that the program is better suited to finding more of the e-mails known to be relevant. A lower score, however, indicates that the filter is rejecting more relevant e-mails. Similarly, a higher "Irrelevant Accuracy" measurement indicates that the program is better suited to dismissing more irrelevant e-mails. A lower score indicates that it is including more irrelevant e-mails than desired.

Finally, we experimented with varying the cost weightings for type I and type II errors. Type I errors, or "false positives" are errors where the experiment marks an e-mail "known" (via the previous classification) to be 'relevant' as 'irrelevant'. Type II errors, or "false negatives", are errors where the e-mail classification system marks an e-mail "known" to be 'irrelevant' as 'relevant'. Given that in an e-discovery scenario, it is far more important to keep relevant e-mails than it is to throw out irrelevant e-mails, we experimented with equal weightings up to significantly skewed weightings in favor of relevant e-mails.

### *b. Experimentation Results*

This section presents tables showing the inputs and results for each trial and run, utilizing the two datasets described earlier.

1. Dataset 1: Trial 1

In Dataset 1, Trial 1, we selected randomly 200 e-mails out of the 1,000 e-mails to train the Bayesian filter. We used the remaining 800 e-mails to test the filter's accuracy. The trial was run 5 times; with each run, a different set of 200 e-mails were selected to train the filter. The following tables show the inputs per run (Table 1A) and the results obtained (Table 1B).

Table 1A. Inputs

| Known Relevant | | | | Known Irrelevant | | | |
|---|---|---|---|---|---|---|---|
| total e-mails | random sample | total sample for training | remaining for testing | total e-mails | random sample | total sample for training | remaining for testing |
| 324 | 20% | 65 | 259 | 675 | 20% | 135 | 541 |

Table 1B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 259 | 114 | 145 | 44% | 541 | 139 | 402 | 74% |
| 2 | 259 | 112 | 147 | 43% | 541 | 107 | 434 | 80% |
| 3 | 259 | 120 | 139 | 46% | 541 | 152 | 389 | 71% |
| 4 | 259 | 96 | 163 | 37% | 541 | 85 | 456 | 84% |
| 5 | 259 | 122 | 137 | 47% | 541 | 131 | 410 | 75% |
| | | | average | 43% | | | average | 77% |

### 2. Dataset 1: Trial 2

In Dataset 1, Trial 2, we selected randomly 500 e-mails out of the 1,000 e-mails to train the Bayesian filter. We used the remaining 500 e-mails to test the filter's accuracy. The trial was run 5 times; with each run, a different set of 500 e-mails were selected to train the filter. The following tables show the inputs per run (Table 2A) and the results obtained (Table 2B).

Table 2A. Inputs

| Known Relevant | | | | Known Irrelevant | | | |
|---|---|---|---|---|---|---|---|
| total e-mails | random sample | total sample for training | remaining for testing | total e-mails | random sample | total sample for training | remaining for testing |
| 324 | 50% | 162 | 162 | 676 | 50% | 338 | 338 |

Table 2B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 162 | 100 | 62 | 61% | 338 | 125 | 213 | 63% |
| 2 | 162 | 102 | 60 | 62% | 338 | 115 | 223 | 65% |
| 3 | 162 | 61 | 101 | 37% | 338 | 59 | 279 | 82% |
| 4 | 162 | 61 | 101 | 37% | 338 | 60 | 278 | 82% |
| 5 | 162 | 101 | 61 | 62% | 338 | 111 | 227 | 67% |
| | | | average | 52% | | | average | 72% |

### 3. Dataset 1: Trial 3

In Dataset 1, Trial 3, we selected randomly 800 e-mails out of the 1,000 e-mails to train the Bayesian filter. We used the remaining 200 e-mails to test the filter's accuracy. The trial was run 5 times; with each run, a different set of 800 e-mails were selected to train the filter. The following tables show the inputs per run (Table 3A) and the results obtained (Table 3B).

Table 3A. Inputs

| Known Relevant | | | | Known Irrelevant | | | |
|---|---|---|---|---|---|---|---|
| total e-mails | random sample | total sample for training | remaining for testing | total e-mails | random sample | total sample for training | remaining for testing |
| 324 | 80% | 259 | 65 | 675 | 80% | 541 | 135 |

Table 3B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 65 | 37 | 28 | 56% | 135 | 36 | 99 | 73% |
| 2 | 65 | 25 | 40 | 38% | 135 | 38 | 97 | 71% |
| 3 | 65 | 28 | 37 | 43% | 135 | 25 | 110 | 81% |
| 4 | 65 | 37 | 28 | 56% | 135 | 32 | 103 | 76% |
| 5 | 65 | 44 | 21 | 67% | 135 | 44 | 91 | 67% |
| | | | average | 52% | | | average | 74% |

### 4. Dataset 2: Trial 1

In Dataset 1, Trial 1, we selected randomly 200 e-mails out of the 1,000 e-mails to train the Bayesian filter. We used the remaining 800 e-mails to test the filter's accuracy. The trial was run 5 times; with each run, a different set of 200 e-mails were selected to train the filter. The following tables show the inputs per run (Table 4A) and the results obtained (Table 4B).

Table 4A. Inputs

| Known Relevant | | | | Known Irrelevant | | | |
|---|---|---|---|---|---|---|---|
| total e-mails | random sample | total sample for training | remaining for testing | total e-mails | random sample | total sample for training | remaining for testing |
| 290 | 20% | 58 | 232 | 710 | 20% | 142 | 568 |

Table 4B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 232 | 144 | 88 | 62% | 568 | 186 | 382 | 67% |
| 2 | 232 | 159 | 73 | 68% | 568 | 203 | 365 | 64% |
| 3 | 232 | 124 | 108 | 53% | 568 | 151 | 417 | 73% |
| 4 | 232 | 120 | 112 | 51% | 568 | 129 | 439 | 77% |
| 5 | 232 | 110 | 122 | 47% | 568 | 139 | 429 | 75% |
| | | | average | 56% | | | average | 71% |

5. Dataset 2: Trial 2

In Dataset 2, Trial 2, we selected randomly 500 e-mails out of the 1,000 e-mails to train the Bayesian filter. We used the remaining 500 e-mails to test the filter's accuracy. The trial was run 5 times; with each run, a different set of 500 e-mails were selected to train the filter. The following tables show the inputs per run (Table 5A) and the results obtained (Table 5B).

Table 5A. Inputs

| Known Relevant | | | | Known Irrelevant | | | |
|---|---|---|---|---|---|---|---|
| total e-mails | random sample | total sample for training | remaining for testing | total e-mails | random sample | total sample for training | remaining for testing |
| 290 | 145 | 50% | 145 | 710 | 50% | 355 | 355 |

Table 5B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 145 | 96 | 49 | 66% | 355 | 120 | 235 | 66% |
| 2 | 145 | 87 | 58 | 60% | 355 | 116 | 239 | 67% |
| 3 | 145 | 105 | 40 | 72% | 355 | 157 | 198 | 55% |
| 4 | 145 | 93 | 52 | 64% | 355 | 112 | 243 | 38% |
| 5 | 145 | 96 | 49 | 66% | 355 | 116 | 239 | 67% |
| average | | | | 66% | average | | | 59% |

## 6. Dataset 2: Trial 3

In Dataset 2, Trial 3, we selected randomly 800 e-mails out of the 1,000 e-mails to train the Bayesian filter. We used the remaining 200 e-mails to test the filter's accuracy. The trial was run 5 times; with each run, a different set of 800 e-mails were selected to train the filter. The following tables show the inputs per run (Table 6A) and the results obtained (Table 6B).

Table 6A. Inputs

| Known Relevant | | | | Known Irrelevant | | | |
|---|---|---|---|---|---|---|---|
| total e-mails | random sample | total sample for training | remaining for testing | total e-mails | random sample | total sample for training | remaining for testing |
| 290 | 80% | 232 | 58 | 710 | 80% | 568 | 142 |

Table 6B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 58 | 39 | 19 | 67% | 142 | 67 | 75 | 52% |
| 2 | 58 | 42 | 16 | 72% | 142 | 61 | 81 | 57% |
| 3 | 58 | 45 | 13 | 77% | 142 | 55 | 87 | 61% |
| 4 | 58 | 20 | 38 | 34% | 142 | 19 | 123 | 86% |
| 5 | 58 | 26 | 32 | 44% | 142 | 20 | 122 | 85% |
| | | | average | 59% | | | average | 68% |

## 7. Dataset 1: Trial 4

For this final Trial run, we used the same 500 identical e-mails for each run. But we altered the cost weighting of the algorithm: the cost of misclassifying a 'relevant' e-mail as 'irrelevant' and the cost of misclassifying an 'irrelevant' e-mail as 'relevant.' The following tables show the modified weights (Table 7A) and the results obtained (Table 7B).

Table 7A. Weighs

| run | Relevant | Irrelevant | Notes |
|---|---|---|---|
| 1 | 1 | 1 | equal weights |
| 2 | 1,000 | 1 | 1,000x worse classifying relevant as irrelevant |
| 3 | 1,000,000 | 1 | |
| 4 | 1,000,000,000,000 | 1 | |
| 5 | 1e18 | 1 | |
| 6 | 1e30 | 1 | |
| 7 | 1e45 | 1 | |
| 8 | 1e75 | 1 | |
| | | | |
| 9 | 1 | 10 | 10x worse classifying irrelevant as relevant |

Table 7B. Results

| | Relevant | | | | Irrelevant | | | |
|---|---|---|---|---|---|---|---|---|
| run | known relevant | found relevant | found irrelevant | accuracy | known irrelevant | found relevant | found irrelevant | accuracy |
| 1 | 162 | 80 | 82 | 49% | 338 | 93 | 245 | 72% |
| 2 | 162 | 80 | 82 | 49% | 338 | 98 | 240 | 71% |
| 3 | 162 | 81 | 81 | 50% | 338 | 102 | 236 | 69% |
| 4 | 162 | 84 | 78 | 51% | 338 | 110 | 228 | 67% |
| 5 | 162 | 86 | 76 | 53% | 338 | 122 | 216 | 63% |
| 6 | 162 | 94 | 68 | 58% | 338 | 146 | 192 | 56% |
| 7 | 162 | 99 | 63 | 61% | 338 | 154 | 184 | 54% |
| 8 | 162 | 108 | 54 | 66% | 338 | 188 | 150 | 44% |
| | | | | | | | | |
| 9 | 162 | 79 | 83 | 48% | 338 | 91 | 247 | 73% |

### *c. Discussion*

After running all the trials, it appears that 200 training e-mails are insufficient to properly train the filter. 500 and 800 e-mails both seem to provide the same results. Also, it appears that the filter is better able to remove irrelevant e-mails, when based on the review of one individual (Dataset 1). However, it is better able to capture relevant e-mails based on the training reviews of several reviewers (Dataset 2).

Finally, a cost weighting system does seem necessary if it is more important to catch more relevant e-mails than it is to filter out irrelevant e-mails. However, it cannot be pushed to an extreme. In our case, from the equal weighting to the heaviest weighting differences that we used, relevant e-mail accuracy increased from 49 percent to 66 percent, but also irrelevant e-mail accuracy diminished from 72 percent to 44 percent, significantly impairing the removal of irrelevant documents. The actual cost weights can be adjusted infinitely, but at its heaviest weights, every e-mail would be classified as relevant, rendering it useless. As such, care must be used in evaluating this weighting system.

## V. CONCLUSION

The Enron corpus provides an excellent dataset for experimenting on accuracy of text classifiers. We applied dbacl, a Bayesian text classifier to identify e-mails as case-relevant or case-irrelevant communications. The results show that increasing weight in running the dbacl filter might be more appropriate for e-discovery’s goal of capturing more relevant e-mails rather than weeding out irrelevant e-mails.

Further experimentation could improve the text classifier results. In this project, we only used items from the 'Sent Items' folders. Perhaps, using some additional folders could have improved the numbers further. Also, we collected 1,000 e-mails at random from all employees in the dataset. It is possible that limiting the randomization to some users would render better results. Some alternatives include using the folders with the highest numbers of e-mail communications and taking advantage of social network research to select strategically the folders to be included. Finally, going beyond the 1,000 e-mails selected as a random pool could reduce statistical bias.

## References

[1] Jason Krause, "In Search of the Perfect Search," ABA Journal 95, April 2009 38-43. Available at: http://www.abajournal.com/magazine/in_search_of_the_perfect_search.

[2] Federal Energy Regulatory Commission, "FERC: Information Released in Enron Investigation," available at: http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp.

[3] United States v. Richard A. Causey, Jeffrey K. Skilling, and Kenneth L. Lay, Cr. No. H-04-25(S-2)(S.D.Tex. 2004). Available at: http://www.usdoj.gov/dag/cftf/chargingdocs/skillingindictment.pdf.

[4] In re Enron Corp., 314 B.R. 524 (Bankr. S.D.N.Y. 2004) (discussing Enron's participation in the California Energy Markets).

[5] Federal Energy Regulatory Commission, "Staff Report: Price Manipulation in Western Markets," available at: http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp.

[6] U.S. Dep't of Justice, Press Release (August 5, 2004), available at: http://www.usdoj.gov/enron/.

[7] William Cohen, "Enron Email Dataset," Carnegie Mellon School of Computer Science, available at: http://www.cs.cmu.edu/~enron/.

[8] Bryant Klimt and Yiming Yang, "Introducing the Enron Corpus" Conference on Email and Anti-Spam, Mountain View, CA, 2004, available at: http://www.ceas.cc/papers-2004/168.pdf.

[9] Yingjie Zhou, et at., "Strategies for Cleaning Organizational Emails with an Application to Enron Email Dataset," 5th Conf. of North American Association for Computational Social and Organizational Science (NAACSOS 07), Emory - Atlanta, Georgia, USA. June 7-9, 2007. Available at: http://www.cs.rpi.edu/~magdon/ps/conference/CASOS07_EmailDataCleaningStrategies.pdf.

[10] Ron Bekkerman et al, "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora, Technical Report, University of Massachusetts, available at: http://www.cs.umass.edu/~ronb/papers/email.pdf.

[11] Jitesh Shetty and Jafar Adibi, "The Enron Email Dataset: Database Schema and Brief Statistical Report," Technical Report, Information Sciences Institute, 2004. Available at: http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf.

[12] Laird Breyer, “dbacl – A Digramic Bayesian Filter Classifier,” available at: http://dbacl.sourceforge.net/contents.html.

[13] Harry Zhang, “The Optimality of Naïve Bayes,” University of New Brunswick, available at: http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf.