Evaluating Expertise and Sample Bias Effects for Privilege Classification in E-Discovery

Jyothi K. Vinjumur
College of Information Studies
University of Maryland
College Park, USA
jyothikv@umd.edu

ABSTRACT

In civil litigation, documents that are found to be relevant to a production request are usually subjected to an exhaustive manual review for privilege (e.g, for attorney-client privilege, attorney-work product doctrine) in order to be sure that materials that could be withheld is not inadvertently revealed. Usually, the majority of the cost associated in such review process is due to the procedure of having human annotators linearly review documents (for privilege) that the classifier predicts as responsive. This paper investigates the extent to which such privilege judgments obtained by the annotators are useful for training privilege classifiers. The judgments utilized in this paper are derived from the privilege test collection that was created during the 2010 TREC Legal Track. The collection consists of two classes of annotators: "expert" judges, who are topic originators called the Topic Authority (TA) and "non-expert" judges called assessors. The questions asked in this paper are; (1) Are cheaper, non-expert annotations from assessors sufficient for classifier training? (2) Does the process of selecting special (adjudicated) documents for training affect the classifier results? The paper studies the effect of training classifiers on multiple annotators (with different expertise) and training sets (with and without selection bias). The findings in this paper show that automated privilege classifiers trained on the unbiased set of annotations yield the best results. The usefulness of the biased annotations (from experts and non-experts) for classifier training are comparable.

General Terms

Electronic Discovery, Privilege Classifier Performance, Experimentation.

1. INTRODUCTION

In United States, civil litigation is a legal dispute between two or more parties. Civil lawsuits generally proceed through distinct steps: pleadings, discovery, trial and possibly an appeal. Traditionally, discovery focused on materials that are paper documents. Since most documents today are in electronic format, the meaning of "qualifying evidence" has experienced a definitional change over time. On December 6 2006, the Federal Rules of Civil Procedure amended that the traditional discovery process address the discovery of the Electronically Stored Information (ESI). This enactment resulted in the term "electronic" to precede the word discovery to enable a legal process called electronic discovery or E-Discovery. Since then, identifying and retrieving relevant documents from large collections of electronic records and yet withholding privileged documents during production is a practical process in litigation.

Privilege is a right given to an individual or organization in the lawsuit, to allow protection against disclosure of information. In litigation, there are many types of Privilege namely: Attorney-Client Privilege or Legal Professional Privilege, Public Interest Privilege, Without Prejudice Privilege, Privilege Against Self-Incrimination, etc. Attorney work-product is a doctrine that protects from discovering materials prepared by the attorney or attorney's representative in view of litigation [13].

Since the 2006 amendments, the task of withholding documents on the basis of attorney-client privilege alone, has faced multiple challenges in litigation cases [14, 18]. The attorney-client privilege is aimed to protect the information exchange between "privileged persons" for the purpose of obtaining legal advice. Privileged persons include [13]:

- the client (an individual or an organization),
- the client's attorney,
- communicating representatives of either the client or the attorney, and
- other representatives of the attorney who may assist the attorney in providing legal advice to the client.

Apart from people information, privilege strongly depends on the context of the communication. Thus privilege is a property of a communication that happened between two or more privileged people about the topic of litigation. Even when the communication between the entities has been made in confidence for the purpose of obtaining legal advice, the existence of privilege can be waived due to the involvement

ICAIL '15, June 08 - 12, 2015, San Diego, CA, USA.
Copyright 2015 ACM 978-1-4503-3522-5/15/06 ...$15.00.

Table 1: Adjudication Categories

Category	Sample	Adjudicated	Sampling Rate	Non-Adjudicated
Team Appeal (A)	237 of 6,766	237 of 237	1.00	6,230[1]
Assessor Disagreement (D)	76 of 730	76 of 76	1.00	
Random Sample (R)	6,529 of 6,529[2]	223 of 6,529	0.16[3]	

of a third party [2] or even due to inadvertent disclosures.

In practice, inadvertent disclosures appear at greater frequency [1, 13]. Such accidental disclosures of privileged information cause litigators greater anxiety, since the possibility of failing to protect the attorney-client privilege may potentially lead to lawsuit on unrelated topics. To avoid privilege to be waived due to inadvertent disclosures, dependence on human to review each and every responsive electronic document is adopted. Thus, in e-discovery, the cost of privilege review process is dominated due to the process of having human reviewers review the documents that the classifier predicts as responsive.

To facilitate reduction in privilege review cost, in this paper, we take a first step to study the effect of using automation by training privilege classifiers on a smaller set of manual privilege judgments. We build and evaluate the privilege classifiers by training the model on seed set annotated by subject matter experts (TA) and non-experts (assessors). We analyze which set of training documents derive better model predictions. We build three models; the first two models respectively consider only network and content information as features, while the third joint model exploits both the network and content information as features to generate privilege predictions on the held-out test-set.

The remainder of this paper is organized as follows. In section 2 we review the related work discussing the initial and current state-of-the-art technology-assisted e-discovery approaches currently employed in practice. We describe the test collection used in our paper in section 3. Section 4 details the research questions. In section 5 we explain the experimental design with results. Section 6 concludes the paper.

2. RELATED WORK

The first effort at creating an avenue for e-discovery domain research and evaluation was initiated by the TREC Legal Track after the 2006 revisions to the Federal Rules of Civil Procedure. The principal goal of the TREC Legal Track was to develop multiple ways of evaluating search technology for e-discovery [5]. Keyword search approach was one of the initial attempts taken to help the lawyers manage the enormous amounts of documents [6]. Each document matching the query term in the keyword approach would be subjected to a linear manual review. The idea of using keyword search approach was to filter the number of documents to be reviewed by human annotators. A study of large scale review for both responsiveness and privilege which was performed with 225 attorneys, revealed that an average of 14.8 documents were annotated per hour per attorney [23]. Some extensions to keyword search approach called concept search are employed to extend the search terms to include context information [17]. However, as corporate collections have continued to grow, filtering by keywords have left huge document sets to be linearly reviewed [8] making linear review procedure insupportable [22]. Thus use of automated classifiers with a higher degree of technological assistance using machine learning techniques are currently being studied in e-discovery domain [15].

As more and more litigators today are familiar with the use of technology and automated classifiers, the effectiveness and evaluation of such automated classifiers has gained the interest of not only E-discovery vendors but also the courts [21]. Although many types of ESI documents could be important in e-discovery, emails are of particular interest because much of the activity of an organization is ultimately reflected in the emails sent and/or received by its employees. In addition to its prominence, since email collection is a great avenue to search for communications that could be withheld on the grounds of attorney-client privilege, we utilize the privilege judgments obtained from TREC 2010 Email Test Collection (refer section 3) in our experiments.

Prior work on email collections has shown promising results in classifying emails using features by isolating unstructured text (fields like subject & body) and the semi-structured text (categorical text from "to", "from", "cc" and "bcc") [12, 19]. Shetty et al study the pattern of email exchanges over time between 151 employees in Enron during the height of the company's accounting scandal [24]. McCallum et al took an initial step towards building a model that captures actor roles and email relationships using dependencies between topics of conversation [20]. Since then, several other generative models have been proposed [27, 32]. Identifying key nodes or individuals in email communications has become an essential part of understanding networked systems, with applications in wide range of fields like; marketing campaigns [16], litigation [9], etc. Since such social network and textual content features have shown to uncover interesting communication patterns in emails, we attempt to exploit the benefits of isolating meta-data information and the email content information to build features for our classification system.

To evaluate such classifier's effectiveness, availability of reliable annotated data is required. However, the process of gathering reliable annotations are fraught with multiple problems. In the e-discovery domain, one such problem is the requirement for skilled legal annotators for review who make the review process more expensive. The cost further depends on the expertise of the annotator. Previous work has demonstrated that training a system on assessments from non-expert assessors leads to a significant decrease in

[1] 6,230 [NAS] = (6766 - (237+76+223)[AS]).

[2] 6,529 = (6766-237). Since random sample is obtained after team appeal draw without replacement.

[3] Average of the sampling probabilities of all strata.

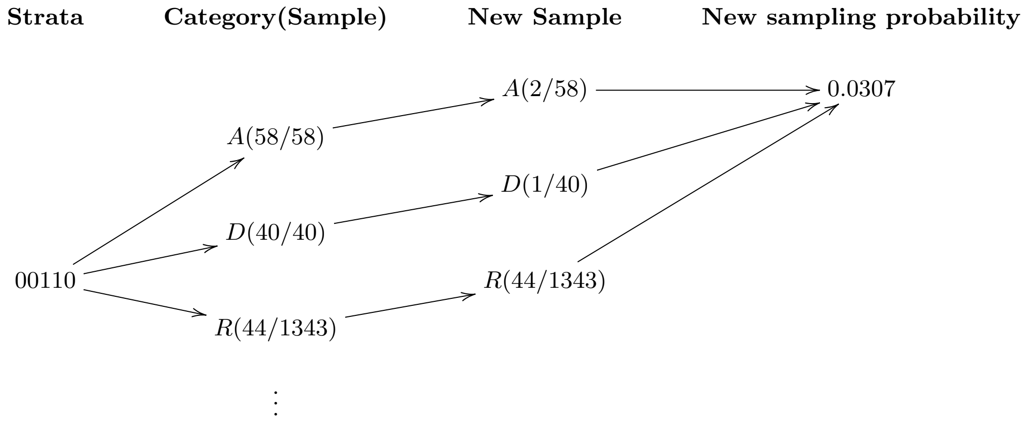


Figure 1: Re-sampling Procedure

reliability of the retrieval effectiveness while evaluated on expert judgments [31] and empirical findings have shown that annotations from experts would lead to better classifier accuracy [4]. However, Cheng et al describes the benefits of utilizing noisy annotations to enhance classifier performance in a multi-annotation environment [10]. Thus it is reasonable to accept that many factors like sampling, annotator expertise, etc., affect the process and quality of gathering relevance assessments. As it is not realistic for human annotators to be infallible [26, 30, 28], this work aims to study the effect of annotator expertise and document selection bias on privilege classifier training.

3. TEST COLLECTION

In the 2010 TREC Legal Track, the document collection used for all Interactive tasks (including the privilege task) was derived from EDRM Enron Collection, version 2, which is a collection of Enron email messages. The privilege task was to retrieve "all documents or communications that are subject to a claim of attorney-client privilege, work-product, or any other applicable privilege or protection". The items to be classified were "document families". A family was defined as an email message together with all of its attachments. For the TREC Legal Track 2010 privilege task, two teams (CB and IN) submitted their results [11]. Team CB submitted four runs and Team IN submitted one run. Each of the five runs retrieved a set of document families that were privileged. Thus the entire test collection was stratified according to the intersecting sets of documents returned by the five submissions creating 32 strata, each defined by a unique 5-bit binary vector [11]. Most families in the collection belonged to the 00000 stratum which contained families there were not retrieved by any of the teams. To obtain relevance assessments, a random sample of families were drawn from each of the strata, with the sampling rate for the 00000 stratum (0.8%) far sparser than for any other stratum (average at 6.1%). Assessors were provided with annotation instructions written by a senior litigator, who was the topic originator (Topic Authority (TA)), to obtain their judgments.

In [11] assessor's judgment on any family could be escalated for TA adjudication under three conditions[4] as shown in Table 1; (1) in good faith, a team could appeal the decision of an assessor to the TA; A total of 237 appeals out of 6,766 total annotated families were received; (2) 730 assessor annotated families were sampled for dual assessment which could create disagreements among assessors. Families that were in disagreement, were escalated for adjudication. (3) a sample of 223 assessor-annotated families were independently drawn at random from each stratum, excluding the 237 families that were appealed. This resulted in a smaller stratified sample of the full collection (creating selection bias due to sampling without replacing the 237 appealed families); Thus the adjudication process resulted in a total of 536 families to be adjudicated by the TA creating a set called the Adjudicated Set (AS). The remaining 6230 assessor annotated families make the Non-Adjudicated Set (NAS).

This paper utilizes these relevance judgments for building and evaluating our classifiers. Since the families in the AS are biased due to the presence of the families appealed by the team and the families that were in disagreement between assessors, to create an unbiased set of adjudicated families for evaluation, we need to eliminate the selection bias by re-sampling from the biased adjudicated categories. Figure 1 shows our graphical re-sampling procedure for a single stratum 00110. In 00110 stratum, 40 dual assessed families that caused disagreement among the assessors were adjudicated along with 58 families that were appealed. To maintain the sampling probability at the rate of 0.03[5], we randomly draw 2 families from appeal (A) category and one from assessor disagreement (D) category and include these families in the test-set. This procedure is repeated for each stratum creating an unbiased stratified sample of 252 families across all strata. To reduce the impact of measurement error on the classifier evaluation, we use TA judgments (on the unbiased 252 families in the held-out test-set) as gold standard [25]. The remaining families in the AS and NAS are used for training our classifiers.

Although the relevance judgments obtained during 2010 TREC Legal Track represent labels for both privilege and attorney work-product, in this paper, we concentrate on modeling and predicting for privilege only on the grounds of attorney-client privilege.

[4] Making three disjoint sets of adjudicated families.

[5] This is the sampling probability of the random category.

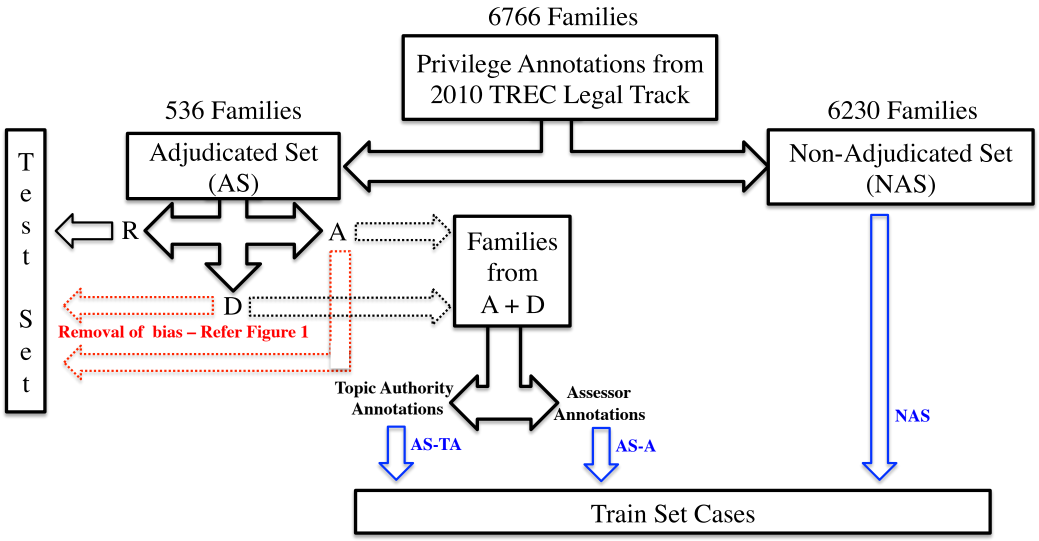


Figure 2: Train-Set and Test-Set Split Procedure

4. PROPOSED WORK

In this paper, we study the effect of training privilege classifiers on two sets of families. Figure 2 graphically explains the process of selecting families for training and testing our classifiers. The 6,766 family annotations from the 2010 TREC Legal Track are utilized to create an unbiased (Section 3 explains our re-sampling procedure to remove the selection bias) test-set. Although the families in the held-out test-set have assessments from both the assessors and the TA, we use the TA judgments on the 252 families in the test-set as gold standard for evaluation [25]. The remaining families in AS annotated by the TA $(AS-TA)$ and the assessor$(AS-A)$, along with annotations from the NAS, create the three training cases. Table 2 shows the privilege class prevalence and the number of privileged and not-privileged families in each of the three training cases.

We build three different classifiers for each of these three training sets. The classifiers differ in their feature set as explained in section 5. Thus, the 9 (3 different models trained on 3 different train-sets) automated classifiers allow us to study the influence of (1) annotator expertise and (2) selection bias on the training families. We build supervised classifiers using labeled families from the two disjoint sets. One set utilizes the families in AS for training while the other utilizes an equal number of families (to maintain the prevalence π) from NAS. Since the families in AS are dual-assessed, we utilize the assessments from TA (*model*-AS-TA[6]) and the assessors (*model*-AS-A[7]) to study the effect of expertise on classifier training. All families in the NAS are annotated by only assessors.

Table 2: Training Families

Train-Set Case			
Case ID	Privileged	Not-Privileged	π- Prevalence
$AS-TA$	166	113	0.59
$AS-A$	169	110	0.60
NAS	166	113	0.59

Thus, in the results section, we use the classifiers' performance to (1) analyze the effect of expertise on training classifiers by comparing *model*-AS-TA and *model*-AS-A; (2) analyze the effect of selection bias on training classifiers by comparing *model*-AS-A and *model*-NAS.

5. CLASSIFIER DESIGN

Traditionally text classification applications have achieved successful results by using the bag-of-words representation. A number of approaches have sought to replace or improve the bag-of-words representation by adding complex features, however the results have been mixed at best. Although privilege classification can be viewed as a classic text classification problem, the parameters that determine attorney-client privilege depend strongly on (1) the people and (2) the content of the email communication. Since both people and content are equally important in finding privilege, we use both the network and content information of the families to define features. We do this by separating the information in each family into two disjoint components (henceforth called *views*). as shown in Table 3.

The first view $view_1$ exploits the metadata[8] information to obtain the importance score of each actor. We removed a small handful of labeled families (29 families) that are missing sender or/and recipient information during our experiments. In this view, a family is represented as a directed multi-graph (a graph in which multiple edges are permitted

[6] This notation denotes that the *model* is trained on families in AS with expert (TA) judgments.

[7] This notation denotes that the *model* is trained on families in AS with non-expert (Assessor) judgments.

[8] Data in From, To, Cc and Bcc fields

Table 3: Separation of email data

Actor-Centric Features or $view_1$	Content-Centric Features or $view_2$
Sender information - *From* field data	Content - *Subject* field data
Recipient information *To* , *Cc* and *Bcc* field data	Content - data in email body and attachments

between the same nodes) in which each node is an actor and each edge is an email communication between actors. We define $view_1$ as a Graph Model (GM). Our intuition is that, an email message sent/received by an actor "a" has a high probability of being privileged if actor "a" frequently communicates with other actors who have a higher probability of being involved in privileged communications. The second view $view_2$ utilizes the content information in each family. $view_2$ is defined as a Content Model (CM). In CM, we use only the words occurring in the subject field and the content field of the family to derive term features. For model performance comparison, we build a joint model called Mixed Model (MM). The MM uses the features from both the GM and CM. In our experiments, we used three types of classification algorithms: Linear Kernel Support Vector Machines (SVM), Logistic Regression and NaiveBayes, all using the implementations in the Python Scikit-Learn Framework. We report only linear kernel SVM classifier results since we did not observe any significant change in the model performance while using the other two classification algorithms. We compare the classifier results by deriving point estimates for recall and precision with two-tailed 95% approximate confidence intervals. In the next section, we describe the models in detail. Section 5.2 details the estimation and interval calculation.

5.1 Models

5.1.1 *Graph Model*

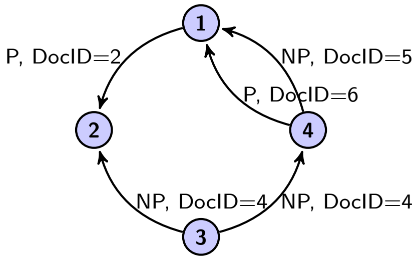


Figure 3: Sample Graph

One common way of representing the information extracted from $view_1$ is by a directed graph structure. Let $G = (V, E)$ denote a directed multi-graph with node set V and edge set E. For a single directed edge (u, v), u is called the sender and v the recipient of the email communication. In the model built using $view_1$ data, each node would represent an individual person and the edge linking the two nodes would represent a family. Consider an example graph sample space G as shown in Figure 3. Here, each edge connecting the nodes is a labeled family. Each labeled training family is represented by the nodes as its features. However our feature extraction technique faces challenges in identifying unique nodes in emails due to the absence of a named-entity linked knowledge base. Hence as a first step, we extract unique actors from emails using string pattern matching approach.

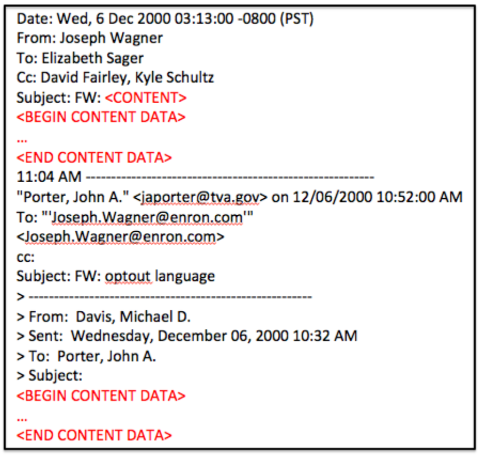
The task is defined as follows: an email is composed of multiple actors with a variety of name mentions as shown in Figure 5. The objective is to identify a set of unique actors across all email communications. To obtain a unique set of actors, we extract the $(sender, [recipient])$ from each family. Once this is done, we compute the similarity using a pattern recognition algorithm between every pair of nodes [7]. The steps for computing similarity in name mention of nodes in emails are as follows: (1) Remove suffixes (like "jr", "sr") and remove generic terms like "admin","enron america", "support", "sales", etc.; turn all white-space into a single hyphen. Next, we merge the first name with the last name using a single hyphen to recognize the person's full name as a single entry. This step ensures that mike.mcconnell and mike.riedel are not similar. Thus, at the end of this step we obtain a list of actor nodes N; (2) For each node n in the set N we identify a set of similar nodes using an approach to match string patterns based on the Ratcliff-Obershelp algorithm. We used the implementation provided by the Python "difflib" module with the cutoff threshold set to 0.75. For the examples shown in figure 5, given the target node "mark.taylor@ees.com", the following close matches are obtained: "mark-taylor, mark.taylor@enron.com". Next, we obtain the correct match by comparing the target word with all its close matches and identifying the matching subsequences. The accuracy of identifying unique nodes using this technique is 0.83 with false positive errors at a higher rate (0.62) than false negatives. As future work, we propose to undertake a better approach in clustering nodes to reduce the false positive errors.

5.1.2 *Content Model*

In this model, an email family is typically stored as a sequence of terms where the terms represent a collection of text from the email message together with the text in all its attachments. Information retrieval techniques have developed a variety of techniques for transforming the terms representing the documents to vector space models to perform statistical classification. In content model, we simply use the words occurring in the subject field and the content field of the family to derive term features. We remove any metadata information (text in black in figure 4) included in the body of the email message. Figure 4 shows the boundaries of the content data extracted from the email message. Text in the attachment is also included in the Content Model. After extracting the text content, we represent the text as a vector space model where the terms are scored using the TF-IDF weighting algorithm.

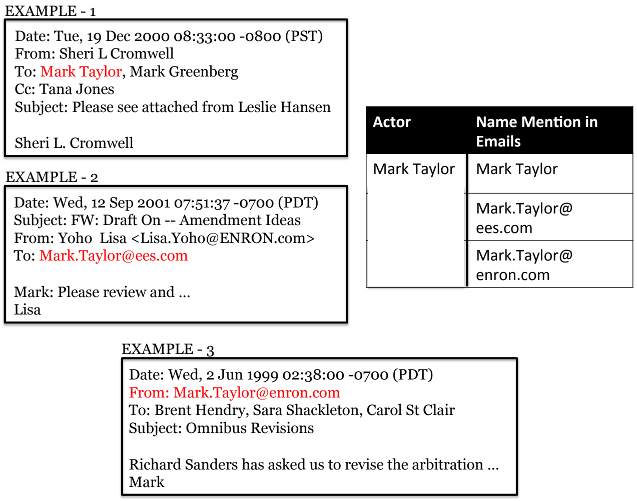
5.2 Evaluation Metric

The evaluation metrics are derived from two intersecting sets; the set of families in the collection that are privileged, and the set of families that a system retrieves (as shown in Table 4). Section 5.2.1 and section 5.2.2 explain the derivation of point estimates and confidence intervals respectively.



```
Date: Wed, 6 Dec 2000 03:13:00 -0800 (PST)
From: Joseph Wagner
To: Elizabeth Sager
Cc: David Fairley, Kyle Schultz
Subject: FW: <CONTENT>
<BEGIN CONTENT DATA>
...
<END CONTENT DATA>
11:04 AM ---------------------------------------------------------
"Porter, John A." <japorter@tva.gov> on 12/06/2000 10:52:00 AM
To: "'Joseph.Wagner@enron.com'"
<Joseph.Wagner@enron.com>
cc:
Subject: FW: optout language
> ---------------------------------------------------------
> From:  Davis, Michael D.
> Sent:  Wednesday, December 06, 2000 10:32 AM
> To:  Porter, John A.
> Subject:
<BEGIN CONTENT DATA>
...
<END CONTENT DATA>
```

Figure 4: Content-centric information in emails



EXAMPLE - 1

```
Date: Tue, 19 Dec 2000 08:33:00 -0800 (PST)
From: Sheri L Cromwell
To: Mark Taylor, Mark Greenberg
Cc: Tana Jones
Subject: Please see attached from Leslie Hansen

Sheri L. Cromwell
```

EXAMPLE - 2

```
Date: Wed, 12 Sep 2001 07:51:37 -0700 (PDT)
Subject: FW: Draft On -- Amendment Ideas
From: Yoho  Lisa <Lisa.Yoho@ENRON.com>
To: Mark.Taylor@ees.com

Mark: Please review and ...
Lisa
```

EXAMPLE - 3

```
Date: Wed, 2 Jun 1999 02:38:00 -0700 (PDT)
From: Mark.Taylor@enron.com
To: Brent Hendry, Sara Shackleton, Carol St Clair
Subject: Omnibus Revisions

Richard Sanders has asked us to revise the arbitration ...
Mark
```

Actor	Name Mention in Emails
Mark Taylor	Mark Taylor
	Mark.Taylor@ ees.com
	Mark.Taylor@ enron.com

Figure 5: Actor variants in emails

Table 4: Contingency Table

Prediction/Judgment	Privileged	Not Privileged	
Retrieved	N_{rp}	$N_{rp'}$	N_r
Not Retrieved	$N_{r'p}$	$N_{r'p'}$	$N_{r'}$
	N_p	$N_{p'}$	N

5.2.1 *Point Estimate*

This sections details the calculations used to estimate the recall and precision of the system. In order to estimate the precision for system T_i, we estimate N^i_{rp}, the number of privileged families returned by system T_i and the total number of families returned by that system N^i_r. Let N^h_{rp} be the number of privileged families in stratum h. The number of privileged families returned by System T_i is the sum of the number of privileged families in the strata returned by System T_i. Thus if $\hat{N}^h_{rp}$ is an unbiased estimator of N^h_{rp} then

$$\hat{N}^i_{rp} = \sum_{h:T_i \in T^h} \hat{N}^h_{rp} \tag{1}$$

is an unbiased estimator of N_{rp} for system T_i where T^h is the set of all systems that retrieved documents in the stratum h.

Now, let the number of documents in stratum h be N_h. A sample of size n_h is drawn from the stratum by simple random sampling without replacement, and n_{hp} of the families in the sample are observed to be privilege. Then, an unbiased estimator of N^h_{rp} is

$$\hat{N}^h_{rp} = N_h * \frac{n^h_r}{n_h} \tag{2}$$

Finally,the estimator of System T_i's precision can be obtained using

$$\hat{Precision}^i = \frac{\hat{N^i_{rp}}}{N^i_r} \tag{3}$$

In order to estimate recall, an estimate of N_p , the total number of privilege documents or yield of the collection, is also required. An unbiased estimate of N_p is obtained by summing the yield estimates for each stratum as shown below:

$$\hat{N}_p = \sum_{h:T_i \in T^h} \hat{N}^h_p \tag{4}$$

The recall estimate of the system T_i is then calculated using the expression

$$\hat{Recall}^i = \frac{\hat{N^i_{rp}}}{\hat{N}_p} \tag{5}$$

5.2.2 *Confidence Intervals*

The recall and precision values derived in section 5.2.1 are point estimates, and are subject to random variation due to sampling and measurement error. Here, we focus on providing an indication of the expected range of variability around a point estimate, and to account for it when comparing two scores. A two-tailed (1-α) confidence interval, [θ_l , θ_u], provides the range within which the population θ lies with confidence (1-α); in other words, if samples were to be repeatedly drawn from the population, and intervals calculated using the same method, then (1-α) of the time, that confidence interval would include θ, the parameter of interest. An exact confidence interval is calculated by finding the lowest upper and highest lower θ value that satisfy a one-tailed significance test. Exact confidence intervals, are often hard or impossible to calculate [3]. An approximate confidence interval is derived by other methods, and typically aims to achieve (1-α) coverage on average across values of the parameter θ, rather than guaranteeing it for every parameter. Throughout this paper, we calculate 95% approximate confidence intervals from beta-binomial posteriors on stratum yields [29].

5.3 Results

Here we analyze the influence of (1) annotator expertise; and (2) selection bias, on classifier training.

5.3.1 *Effect of Annotator Expertise*

We study the effect of annotator expertise on training by using the adjudicated families for training (families in set

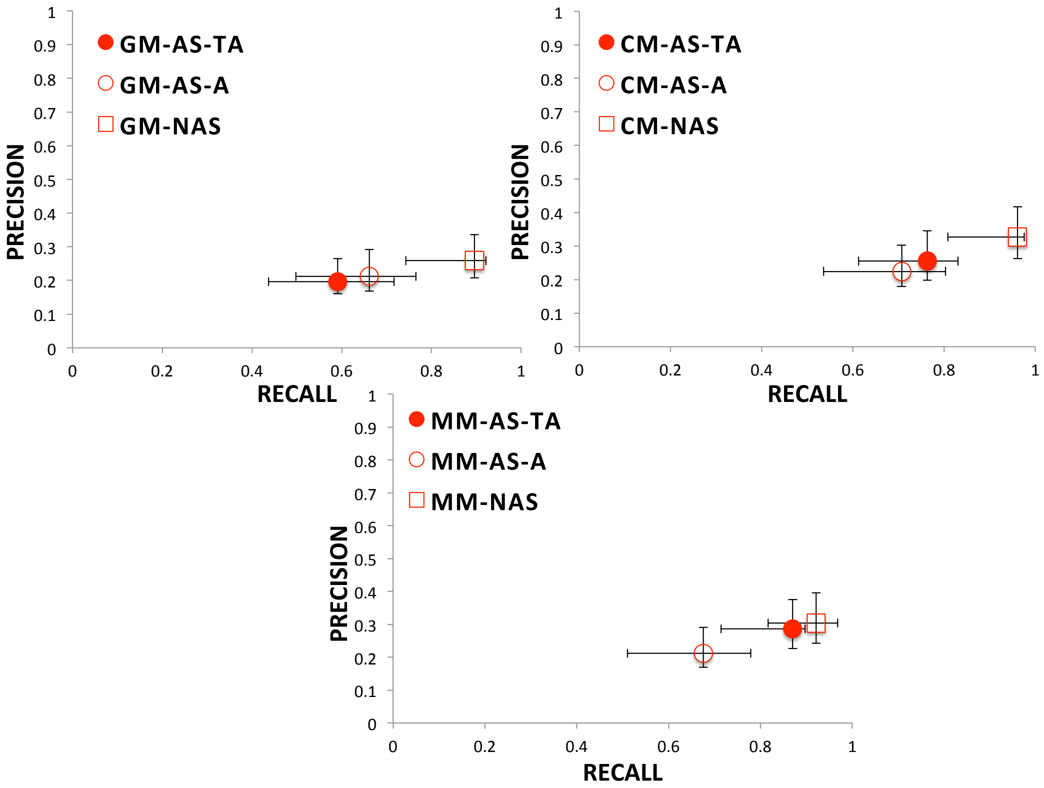


Figure 6: Effect of Annotator Expertise on Training

$AS-TA$ and $AS-A$), and the unbiased held-out set for testing. Although the sample drawn for adjudication in the test collection represents less than 8% of the total size of the official judgments, due to which the results yield fairly wide confidence intervals, the comparison discussed here does offer useful insights.

We compare the classifier performance using recall and precision values with 95% confidence intervals. Figure 6 shows the performance with (95%) confidence intervals on recall and precision for the three classifiers, each of which is trained on each of the three training cases (discussed in section 3). By comparing the performance of training the GM ($GM-AS-TA$ and $GM-AS-A$) and CM ($CM-AS-TA$ and $CM-AS-A$) classifiers on set AS, we observe that classifiers trained on neither expert nor non-expert annotations yield better results. However, by comparing the performance of the joint MM model, $MM-AS-TA$ and $MM-AS-A$, we observe a significant increase in the recall of the automated classifier trained on families in AS with the expert's (TA) annotations.

We explain this by collectively analyzing the classifiers' privilege predictions on the families in the test-set. Figure 7 shows the intersecting sets of all the classifiers' predictions on the privileged families in the test-set. By analyzing a pair of intersecting sets; (1) $CM-AS-TA$ and $MM-AS-TA$ (count of (22+0-1[9]) families), and the sets $CM-AS-A$ and $MM-AS-A$ (count of (15+4-0) families) (2) $GM-AS-TA$ and $MM-AS-TA$ (count of (7+7-1) families), and the sets $GM-AS-A$ and $MM-AS-A$ (count of (2+13-0) families), we deduce that the performance of $MM-AS-TA$ model gains a significant increase in recall over $MM-AS-A$.

5.3.2 Effect of Selection Bias

Comparing the performance of *model*-AS-A and *model*-NAS for each of the three classifiers (MM, GM and CM) in the figure 6 shows that, automated classifiers trained on the unbiased annotations from cheaper non-expert sources (Families in NAS) derive the best results. A significant increase in recall is noticed for all the classifier trained on NAS ($GM-NAS$, $CM-NAS$, $MM-NAS$) when compared to their corresponding classifiers trained on $AS-A$ (($GM-AS-A$, $CM-AS-A$, $MM-AS-A$)). A possible explanation to our finding is the presence of bias in choosing training families. Since families in AS have a selection bias due to the presence of (1) assessor disagreed families and (2) team appealed families, we argue that training classifiers on families in AS could affect the results due to the presence of families which are hardest to annotate (which explains the assessor disagreement) or which could strategically benefit the team's performance (which explains the team-appeals).

Nonetheless, we have shown some evidence that support our findings that: (1) Training classifiers on families chosen at random (annotated by non-expert reviewers) yields the best result and (2) Expert's annotations can also be useful in training automated privilege classifiers.

[9] Privileged family that is predicted as not-privileged by both $CM-AS-TA$ and $GM-AS-TA$

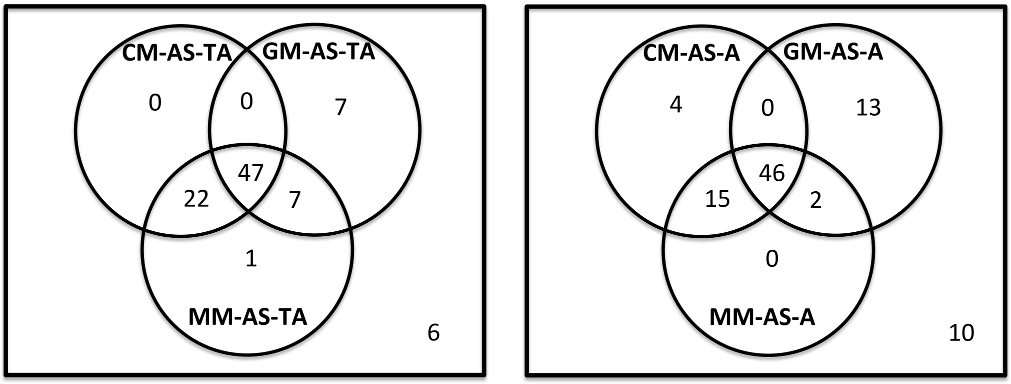


Figure 7: Analysis of Classifier Privilege Predictions

6. CONCLUSION

The growing concern related to the cost involved in privilege review process has forced several e-discovery professionals who are predominantly from the legal domain, to adopt technology-assisted review techniques. In this paper, we approach the issue by asking two simple questions about the effect of annotator expertise and seed-set selection while training a privilege classifier. To answer the questions, we utilize the privilege judgments from TREC Legal Track 2010. We conduct our analysis by training automated classifiers on privilege judgments from annotators with different levels of expertise. We studied the effect of selection bias in the annotated samples on training. Set-based evaluation technique using stratified sampling and approximate confidence intervals from beta-binomial posteriors on stratum yields is employed for comparing classifier results. We conclude that selection bias in training could hurt the classifier performance. Our results show that training privilege classifiers on randomly chosen, non-expert annotations yields the best results. We propose future work to study the effect of annotator expertise on training not only for privilege classifiers but also for responsiveness with the aim to arrive at a cost-effective training methodology.

7. ACKNOWLEDGMENTS

This work has been supported in part by NSF award 1065250. Opinions, findings, conclusions and recommendations are those of the authors and may not reflect NSF views.

8. REFERENCES

[1] Underwater Storage, Inc. v. United States Rubber Co. 314(Civ. A. No. 751-64):546, 1970.

[2] United States v. El Paso Co. 682(No. 81-2484):530, 1982.

[3] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 1998.

[4] P. Bailey et al. Relevance Assessment: Are Judges Exchangeable and does it matter. In *SIGIR*, 2008.

[5] J. R. Baron. The TREC Legal Track: Origins and reflections on the first year. In *Sedona Conf. J.*, 2007.

[6] J. R. Baron. Toward a New Jurisprudence of Information Retrieval: What Constitutes a Reasonable Search for Digital Evidence when Using Keywords. *Digital Evidence & Elec. Signature L. Rev.*, 2008.

[7] P. E. Black. Ratcliff/Obershelp pattern recognition. *Dictionary of Algorithms and Data Structures*, 2004.

[8] D. C. Blair and M. Maron. An Evaluation of Retrieval Effectiveness for a Full-text Document-Retrieval System. *Communications of the ACM*, 1985.

[9] L. Calkins. Enron fraud trial ends in 5 convictions. *The Washington Post*, 2004.

[10] J. Cheng, A. Jones, C. Privault, and J.-M. Renders. Soft Labeling for Multi-pass Document Review. In *ICAIL, DESI V Workshop*, 2013.

[11] G. V. Cormack et al. Overview of the TREC 2010 Legal Track. In *TREC*, 2010.

[12] Y. Diao, H. Lu, and D. Wu. A Comparative study of Classification based Personal E-mail Filtering. In *Knowledge Discovery and Data Mining.* 2000.

[13] E. S. Epstein. *The Attorney-Client Privilege and the Work-Product Doctrine.* ABA 2001.

[14] G. L. Fordham. Using Keyword Search Terms in E-Discovery and How They Relate to Issues of Responsiveness, Privilege, Evidence Standards and Rube Goldberg. *Rich. JL & Tech.*, 2009.

[15] M. Gabriel, C. Paskach, and D. Sharpe. The Challenge and Promise of Predictive Coding for Privilege. *ICAIL 2013 DESI V Workshop*, 2013.

[16] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, 2003.

[17] R. Laplanche, J. Delgado, and M. Turck. Concept search technology goes beyond keywords. *Information Outlook*, 2004.

[18] A. Linzy. Attorney-Client Privilege and Discovery of Electronically-Stored Information. *Duke L. & Tech. Rev.*, 2011.

[19] G. Manco et al. Towards an adaptive mail classifier. In *Italian Association for Artificial Intelligence Workshop*, 2002.

[20] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The Author-Recipient-Topic model for Topic and Role discovery in social networks: Experiments with Enron

and academic email. *Workshop on Link Analysis, Counterterrorism and Security*, 2005.

[21] D. W. Oard et al. Evaluation of Information Retrieval for E-discovery. *Artificial Intelligence and Law*, 2010.

[22] G. L. Paul and J. R. Baron. Information inflation: Can the legal system adapt? *Rich. JL & Tech.*, 2007.

[23] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal Electronic Discovery: Computer Classification vs. Manual Review. *Journal at ASIST*, 2010.

[24] J. Shetty and J. Adibi. Discovering important nodes through Graph Entropy The case of Enron Email Database. In *International workshop on Link discovery*. ACM, 2005.

[25] J. K. Vinjumur, D. W. Oard, and J. H. Paik. Assessing the Reliability and Reusability of an E-discovery Privilege Test Collection. In *SIGIR*, 2014.

[26] E. M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing & Management*, 2000.

[27] X. Wang, N. Mohanty, and A. McCallum. Group and Topic Discovery from Relations and Text. In *International Workshop on Link discovery*. ACM, 2005.

[28] W. Webber. Re-examining the Effectiveness of Manual Review. In *SIGIR Workshop*, 2011.

[29] W. Webber. Approximate Recall Confidence Intervals. *ACM TOIS*, 2013.

[30] W. Webber, D. W. Oard, F. Scholer, and B. Hedin. Assessor Error in Stratified Evaluation. In *CIKM*, 2010.

[31] W. Webber and J. Pickens. Assessor Disagreement and Text Classifier Accuracy. In *SIGIR*, 2013.

[32] D. Zhang et al. Modeling Interactions from Email Communication. IEEE, 2006.