ABSTRACT

Title of dissertation: PREDICTIVE CODING TECHNIQUES WITH MANUAL REVIEW TO IDENTIFY PRIVILEGED DOCUMENTS IN E-DISCOVERY

Jyothi K Vinjumur, Doctor of Philosophy 2018

Dissertation directed by: Professor Douglas W. Oard
College of Information Studies and UMIACS

In twenty-first century civil litigation, discovery focuses on the retrieval of electronically stored information. Lawsuits may be won or lost because of incorrect production of electronic evidence. Organizations may generate fewer paper documents, leading to an increase in the amount of electronic documents by many fold. Litigants face the task of searching millions of electronic records for the presence of responsive and not-privileged documents, making the e-discovery process burdensome and expensive. In order to ensure that the material that has to be withheld is not inadvertently revealed, the electronic evidence that is found to be responsive to a production request is typically subjected to an exhaustive manual review for privilege. Although the budgetary constraints on review for responsiveness can be met using automation to some degree, attorneys have been hesitant to adopt similar technology to support the privilege review process. This dissertation draws attention to the potential for adopting predictive coding technology for the privilege review phase during the discovery process.

Two main questions that are central to building a privilege classifier are addressed. The first question seeks to determine which set of annotations can serve as a reliable basis for evaluation. The second question seeks to determine which of the remaining

annotations, when used for training classifiers, produce the best results. As an answer, binary classifiers are trained on labeled annotations from both junior and senior reviewers. Issues related to training bias and sample variance due to the reviewer's expertise are thoroughly discussed. Results show that the annotations that were randomly drawn and annotated by senior reviewers are useful for evaluation. The remaining annotations can be used for classifier training.

A research prototype is built to perform a user study. Privilege judgments are gathered from multiple lawyers using two user interfaces. One of the two interfaces includes automatically generated features to aid the review process. The goal is to help lawyers make faster and more accurate privilege judgments. A significant improvement in recall was noted when comparing the users' review performance when using the automated annotations. Classifier features related to the people involved in privileged communications were found to be particularly important for the privilege review task. Results show that there was no measurable change in review time.

As cost is proportional to time during review, as the final step, this work introduces a semi-automated framework that aims to optimize the cost of the manual review process. The framework calls for litigants to make some rational choices about what to manually review. The documents are first automatically classified for responsiveness and privilege, and then some of the automatically classified documents are reviewed by human reviewers for responsiveness and for privilege with the overall goal of minimizing the expected cost of the entire process, including costs that arise from incorrect decisions. A risk-based ranking algorithm is used to determine which documents need to be manually reviewed. Multiple baselines are used to characterize the cost savings achieved by this approach.

Although the work in this dissertation is applied to e-discovery, similar approaches could be applied to any case in which retrieval systems have to withhold a set of confidential documents despite their relevance to the request.

PREDICTIVE CODING TECHNIQUES WITH MANUAL REVIEW TO IDENTIFY PRIVILEGED DOCUMENTS IN E-DISCOVERY

by

Jyothi K Vinjumur

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Spring, 2018

Advisory Committee:
Professor Douglas W Oard (Chair)
Associate Professor Hal Daumé III (Dean's Representative)
Assistant Professor Vanessa Frias-Martinez
Assistant Professor Beth St. Jean
Dr. Fabrizio Sebastiani, National Research Council of Italy

Dedication

To mom and dad, Arjun and Kiran.

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I want to thank my advisor Douglas W Oard. It has been a great journey working with him for the last few years. Among many lessons, he has taught me how good research is done. I would like to express my special appreciation to him for letting me make mistakes and grow as a researcher. His timely advice on research, teaching and career tips have been invaluable. I am grateful to have had a few fun filled opportunities to know him personally. Above all, I owe him my deepest gratitude for boosting my confidence by helping me realize my strengths and improve on my weakness during my PhD journey.

I would like to thank Beth St Jean, Fabrizio Sebastiani, Hal Daumè III and Vanessa Friaz-Martinaz for accepting the invitation to serve on my dissertation committee, and for the suggestions they provided during my thesis proposal and my dissertation work.

I would like to extend my gratitude to Fabrizio Sebastiani for his collaboration, co-authorship and research guidance on a few of the major contributions in my dissertation. I thank Jiaul Paik and Amittai Axelrod, for giving me an opportunity to learn and collaborate with them.

I am honored to have had the opportunity to meet multiple e-discovery experts along the way. My first thanks goes to Jason Baron whom I look up too with great admiration. He has always made me feel that both me and my work has great potential and has never missed an opportunity to introduce me to many of his colleges and acquaintances with great pride. I would like to thank David Lewis for letting me sit-in in one of his courses he offered in Georgetown University and providing me valuable inputs from time

to time. I would thank Maura Grossman for taking time to meet me at her office in NYC to discuss the privilege review interface design. I extend my gratitude to all the lawyers who participated in my study.

My work during my PhD journey was supported in part by NSF awards 1065250 and 1618695. I would like to thank NSF for supporting me and SIGIR travel grant for providing the financial assistance for conference travel.

I thank Maarten de Rijke for giving me the opportunity to work with him and his wonderful team of researchers at UvA. My stay at Amsterdam was an absolute delight because of the awesome people I met there. I owe my sincere thanks to David Graus, Zhaochun Ren, Marlies van der Wees, Manos Tsagkias, Tom Kenter, Fei Cai, Anne Schuth and Richard Berendsen. I would like to thank Hans Henseler for giving me an invitation to attend an e-discovery symposium in Amsterdam and learn about the difference of this domain in a different continent. I am grateful to Amanda Jones to have given me an opportunity to work with her and her awesome team at H5 during the summer of 2016.

The members of the Information Retrieval Research group in CLIP lab have contributed immensely to my personal and professional time at UMD. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful for William Webber, Ning Gao, Mossaab Bagdouri, Rashmi Sankepally, Jiaul Paik from CLIP Lab and Camli Badrya a graduate student in Aerospace Engineering Department. From William, I learned the importance of maintaining a detailed record of my experiments as a script that I can re-run in the future. I am grateful to Ning Gao, Mossaab Bagdouri, Rashmi Sankepally, Jiaul Paik and Petra Galuscakova for suggestions they provided on various occasions especially during my practice talks for conferences, dissertation proposal and dissertation defense.

I would like to thank June Ahn for encouraging me to talk to multiple professors in the iSchool and explore the area of research that interests me the most. Without his encouragement I would not have met my advisor Doug Oard.

I would like to thank Ben Shneiderman and Catherine Plaisant for giving me an opportunity to work with them during summer 2012 and letting me know everything about the iSchool and about the PhD program offered in the iSchool.

I extend my gratitude to Reinhard Radermacher and Vikrant Aute who supported and encouraged me to start my PhD program in part-time while I work full-time as a Faculty Research Assistant in the Mechanical Engineering Department.

I would like to thank all the friends I made and their families for making the duration of my life in Maryland to be one of the best in this country so far.

A special thanks to my husband, Arjun for supporting me throughout this experience. To my little darling daughter Kiran, I would like to express my thanks for keeping me company during the final year of my PhD.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AID	Avoiding Inadvertent Disclosure
ALvRS	Active Learning via Relevance Sampling
ALvUS	Active Learning via Uncertainty Sampling
AS	Adjudicated Set
EDRM	Electronic Discovery Reference Model
ESI	Electronically Stored Information
FA	Fully Automatic
FCRP	Federal Rules of Civil Procedure
FM	Fully Manual
MINECORE	MINimizing the Expected COsts of REview
NAS	Non-Adjudicated Set
NIST	National Institute of Standards and Technology
QUIS	Questionnaire for User Interaction Satisfaction
RM	Risk Minimization
RR	Relevance Ranking
TA	Topic Authority
TAR	Technology Assisted Review
TREC	Text Retrieval Conference
UR	Uncertainty Ranking

Chapter 1: Introduction

Civil litigation in United States jurisdiction is a legal dispute between two or more parties where either hold the right to request relevant evidence from each other. The term *discovery* refers to a process in the dispute where one party can request to obtain evidence (documents, tapes,etc,.) from the other party or parties. The party requesting documents during the discovery process is called the requesting party and the party who is responsible for producing the documents as per the request is called as the producing party. Although the question of what qualifies to be relevant is up for debate in the court of law, it is required by the producing party to perform some kind of review to provide documents that are relevant or responsive (in legalese) to the litigation requests and that are not subject to a claim of privilege (e.g., attorney-client privilege). During the *discovery* phase, the resulting transfer of documents from the producing party to the requesting party is referred to as *production*.

In the year 1989, a temporary restraining order to preserve a collection of Electronically Stored Information (ESI) was granted in a court in Washington, DC. The ESI had been shared between members of the National Security Council in the Executive Office of the President of the United States [3]. The basis for this order, was a claim that electronic messages could constitute as evidence of activity in an organization. As a consequence of this event, on December 6th, 2006, the Federal Rules of Civil Procedure (FRCP) amended the traditional discovery process to address the discovery of all ESI. This amendment to the FRCP resulted in the term "*electronic*" to precede the word "*discovery*" giving birth to 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 common practical process during civil

litigation.

The different stages of e-discovery process are illustrated in figure 1.1. E-discovery begins when a producing party is required to produce ESI for the requesting party from sources that it identifies as reasonably accessible. The production request is followed by the collection identification process, pre-processing and filtering before the manual review and analysis phase. In practice, producing parties conduct manual review linearly on all responsive document to assert privilege on some set of responsive documents to withhold confidential content before the final phase of document production.[1] The process concludes when the producing party produces all the responsive and not-privileged documents to the requesting party.

During this process, the cost of e-discovery is incurred at every stage: (1) Cost is incurred while locating the potential sources of ESI that collectively make up a searchable collection of electronic documents, (2) Pre-processing the collection of ESI and classifying the documents that are potentially responsive during the filtering stage and (3) Costs due to the manual review process of identifying responsive documents to be produced and privileged or confidential information to be withheld. Prior studies have shown that the majority of the cost in e-discovery is due to the manual review of documents for responsiveness and privilege (typically about 73 percent) [59].

Thus, it is the manual review phase of e-discovery process that this dissertation discusses. We aim to introduce predictive coding techniques to identify privileged documents. We develop algorithms, evaluation measure and conduct a user study. We conclude this dissertation by focusing on techniques to reduce review cost. The upcoming section details the dissertation design with research questions to explain how we think of handling the problems of privilege review in e-discovery.

[1] There are several grounds on which documents might properly be withheld from production, some of which are referred to as privilege and the others go by other names (e.g. Attorney Work-Product Doctrine, etc,.). For convenience, we group them together and refer to them collectively as privilege.

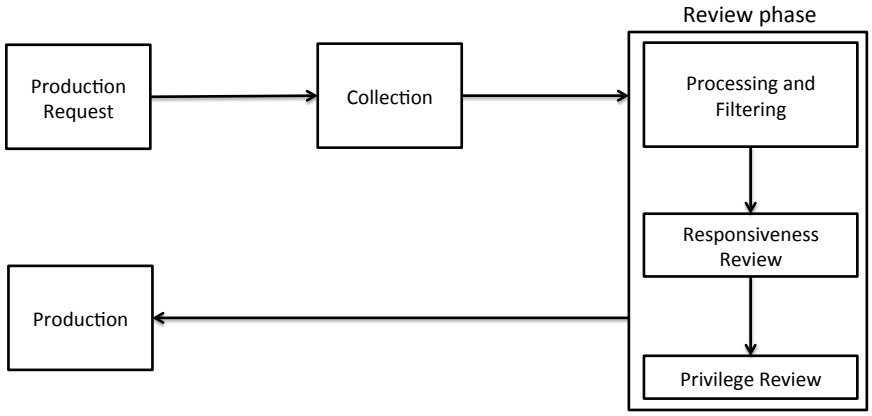

Figure 1.1: E-discovery process

1.1 Motivation

The document processing and filtering stage in e-discovery (refer figure 1.1) concentrates on balancing the document count which affects the manual review cost and review time. Due to the exponential growth in digital content, exhaustive manual review can almost become impossible. This has led to the introduction of a number of techniques for Technology-Assisted Review (TAR), which can be defined as a set of automated techniques that support legal professionals who need to perform an e-discovery review. These automated techniques are also called as *predictive coding* techniques.

One of the earliest articles to describe anything akin to predictive coding techniques was by Anne Kershaw [42]. She described a study that compared a human review team against a document assessment system. While the humans identified only 51% of relevant documents, the system identified more than 95%. The technology her article explained was not what we think of today as predictive coding because it lacks the sophistication of the statistical techniques to determine which documents were relevant. However, her analysis was an initial effort of TAR's eventual refinement.

The next turning point came in the year 2006. That year, the Text Retrieval Conference, an organization started in 1992 by the National Institute of Standards and Technology (NIST) to study information retrieval techniques, launched something called the TREC Legal Track devoted to the use of search and information retrieval in e-discovery.

Its annual research projects provided critical evidence of the efficacy of these techniques in e-discovery. Two e-discovery researchers, Maura R. Grossman and Gordon V. Cormack analyzed data from the 2009 TREC Legal Track involving the use of predictive coding processes. They concluded that predictive coding was not only more effective than human review at finding relevant documents, but also much cheaper [36]. This study found that the use of predictive coding techniques produced almost a 50-fold savings in cost over manual review. Thus, it is becoming increasingly common to perform predictive coding during the e-discovery process. The use of predictive coding techniques have thus revolutionized the filtering process to identify responsive documents. However, the use of predictive coding techniques for the privilege review stage, is still less common.

There are at least two factors causing this difference between the review for responsiveness and the review for privilege. The first one is due to an observed practice of relevance review being performed before privilege review. This is done because it reduces the number of documents that must be reviewed for privilege, thus rendering a linear manual review for privilege more affordable. Second, the failure to detect and properly withhold a privileged document might incur more serious consequences for the party performing the review than would the failure to detect a relevant document determined not to be privileged. Hence legal professionals are less inclined to adopt any fully automated techniques for conducting privilege review.

In this work we try to take an initial step to assure them that adopting predictive coding techniques to perform privilege review is a rational choice. Although we agree that there are cases in which fully manual review is the best choice, we argue that there exist cases where reliance on some degree of automation is a good choice. This work researches multiple ways to assist e-discovery practitioners to make these choices and to explain those choices once they have been made.

Thus the main question here is not just about what the technology is able to do or how legal professionals use what the technologists build, but also about how the legal professionals could use the technology to ensure ESI production at a proportionate cost.

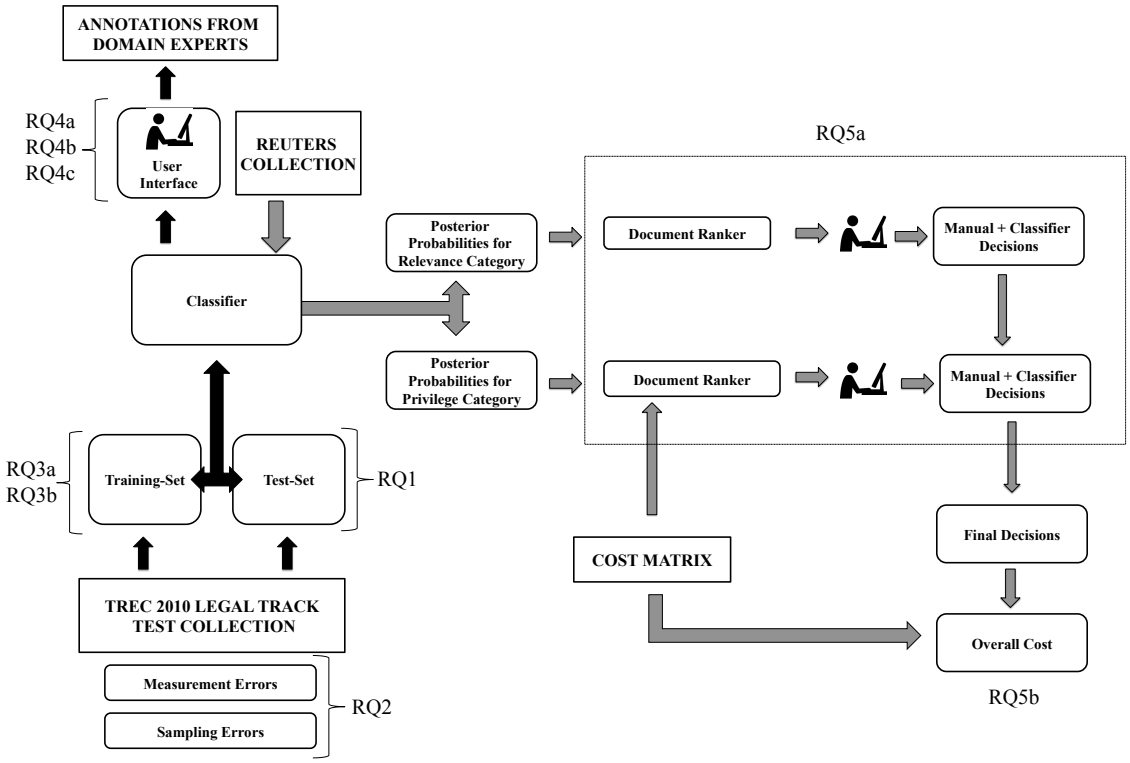

Figure 1.2: Dissertation Overview

1.2 Research Questions

The overall design of this dissertation can be divided into three main components: (1) Predictive Coding component (Chapter 3) (2) Manual Review component (Chapter 4) and (3) Predictive Coding with Manual Review component (Chapter 5). Figure 1.2 graphically illustrates the overall design of this dissertation with clear pointers showing where our research questions fit in.

The *Predictive Coding* component concentrates on building a probabilistic classifier to identify privileged documents. The *Manual Review* component aims to aid the manual privilege review process by utilizing the features from the probabilistic classifier. The *Predictive Coding with Manual Review* component empirically demonstrates the efficacy that can be achieved by our semi-automated system. We next discuss all the the research questions answered in each of the three main components listed above.

1.2.1 Predictive Coding

Most e-discovery vendors today are adopting automation to classify documents during the responsiveness review phase based on input from reviewers. This automation is employed as an effort to expedite the process of filtering the documents in the collection. However, the filtering process during privilege review phase is mostly done manually by domain experts.

Our main goal in this component of our dissertation is to build a binary classifier to identify privileged documents. To build a binary classifier, we need a test collection with privilege judgments that can be used both for training and for evaluation. Thus, before building a binary classifier, we start to think about the evaluation plan. In the year 2010, the initiative taken by the TREC Legal Track, released a Test Collection with privilege judgments for email communications. The document collection was derived from the Enron Email collection. Since that collection is the only public test collection available for conducting e-discovery research, the first research question we ask is:

> ***RQ1: Is it possible to create a labeled test set to enable unbiased classifier evaluation?***

Evaluation of predictive coding systems depends on test collections and the document judgments. During the TREC 2010 Legal Track, multiple teams submitted a total of five system. The results from those five systems were grouped to create a total of 32 categories. There are multiple ways to choose the documents from those categories that need to be manually judged. In TREC 2010, results from submitting teams were pooled to gather samples. Samples were drawn from those categories using a procedure called stratified sampling to obtain manual judgments. A procedure called adjudication was used to expedite the judgments on the sample to a senior assessor (who is an expert) for arbitration. Hence the judgments from this resulting collection were of two types; (1) A small number of documents had judgments from senior and junior (non-expert) assessors and (2) Most of the documents had judgments only from the junior assessor.

For evaluating our privilege classifiers fairly, we need to build an unbiased test

set utilizing the judgments from the TREC 2010 Legal Track collection. To create this unbiased set with senior assessors' judgments[2] for evaluation purpose, we need to eliminate the selection bias (introduced by the appeal process during TREC 2010). We eliminate the bias by re-sampling from the stratified document categories. We maintained the sampling probabilities for each stratum. The procedure resulted in creating a total of 252 document families[3] as gold standard for evaluation.

Although we were able to create a held-out test set for evaluation purposes, the issues of using the stratified approach for document selection during the TREC Legal Track raised two more concerns. The next research question we ask is:

RQ2: Are the privilege judgments obtained from the TREC 2010 Legal Track collection reliable and reusable?

The first issue, which we refer as reliability, is that different manual assessors may reach different judgments for the same document. A second concern about reusability, is that new systems could find some documents that did not contribute to the selection process. Assuming these new documents not to be relevant might adversely affect system comparisons.

In TREC 2010 Legal Track collection, multiple manual assessors with different levels of expertise were involved. Hence for reliability, the key question is to determine the extent to which privilege judgments correctly reflect the opinion of the senior assessor whose judgment is authoritative. For reusability, the key question is to determine the degree to which systems whose results contributed to the creation of the test collection can be usefully compared with other systems that use those privilege judgments in the future. These correspond to measurement error and sampling error, respectively. We performed set-based evaluation using a held-out set of families as test set for privilege classification using stratified sampling, with each strata defined by the overlapping classification results from different participating systems. We examine the impact of unmodeled assessor errors on evaluation results and show recall-precision graph with confidence intervals on the the held-out test set. Our results indicate that measurement errors by junior assessors are

[2]Senior assessor's judgments were always considered as Gold Standard.

[3]In this context, a "document family" (a legal term) refers to an email messages plus all its attachments.

sufficiently large to require their exclusion from the test set if reliable system comparisons are to be made.

Findings from *RQ2* revealed inconsistencies in estimating absolute measures particularly for recall while using junior assessor judgments. This means that, if uncorrected junior assessor judgments were a small fraction of the total judgments, this would be a smaller problem. But in TREC Legal Track 2010 collection, the judgments from the uncorrected junior assessors are being used for about 92% of the sampled documents. Hence the next questions we ask are:

RQ3a: Are the judgments from junior assessors useful for classifier training?

RQ3b: How does the process of selecting training documents (judged by senior assessors or junior assessors or both) affect the classifier performance?

During the creation of TREC Legal Track 2010 collection, relevance judgments gathering process followed a two-stage assessment procedure, whereby an initial relevance assessment was made by junior reviewers for each document in the evaluation sample. A portion of the initially assessed documents were escalated (based on some criteria) to the senior assessor to obtain final judgment. This two-stage assessment procedure created a selection bias both during training and testing our classifiers.

Traditionally, the data used to build a classifier usually comes from multiple datasets. The classifier is first trained on a set of labeled documents called the training set, validation sets can be used for parameter tuning and finally the test set is a set used to provide an unbiased evaluation of a final classifier fit on the training set. To answer the two research questions stated above, we build and evaluate our binary classifiers using multiple training sets and a single held-out test set.

In *RQ3a* we study the effect of utilizing the large amount of junior assessor judgments for training our classifier. Although some documents in this training set may also have the senior assessor's judgment, we consider only the junior assessor's judgment to answer our question.

To answer *RQ3b* we build multiple binary classifiers. We build the classifier using both the content and metadata features. We study the effect of classifier training on (1) multiple annotator types (expert annotators and non-expert annotators) for the same sample and (2) multiple training sets (with and without selection bias). We evaluate our binary classifiers using a held-out test set with senior assessor judgments. The findings show that, larger unbiased training set labeled by a number of junior annotators is about as useful as a smaller biased training set created by a senior annotator. We thus conclude that the use of labeled set from both junior and senior annotators together can be justified for training (although not for testing) the classifier. By building predictive coding models to identify privileged documents, we evaluate the efficacy of adopting predictive coding techniques. Our classifier has better recall measure than precision. Since recall is the more important measure during privilege review in e-discovery, this is a promising result.

We next concentrate on building an interface to determine whether automation can aid privilege review process, especially to avoid inadvertent disclosure of a privileged document.

1.2.2 Manual Review

As manual annotation during privilege review in e-discovery is inevitable, we now seek to build a positive synergy between automation and manual review. It is important to know who uses our system and more so to understand what is it that they are looking for. The motivating factor for designing a user study by building a research prototype was to determine whether the use of automation can aid lawyers perform the privilege review task. We study what type of visible clues via predictive coding assistance could help manual reviewers during the review phase. The objective here is to investigate the extent to which the use of automation (in the form of highlighting potentially useful features and patterns utilizing the metadata and content information) can benefit the manual privilege review process. To this end, we build a research prototype by designing an interactive system to support privilege review in which the objective is to improve the speed and accuracy of the manual privilege review process. At the end of the user study we conducted a semi-structured interview to understand which specific feature was more

beneficial to perform the task. This idea led to the following three sub-questions:

RQ4a: Do the accuracy of the manual reviewer's privilege review judgments improve when system-generated features are presented during privilege review?

Attorney-client privilege exist when the attorney and the client communicate in confidence about an active litigation. The first task we consider, is to identify the actors; who the client is and who the attorney is. We study the relationship of the actors in the email communication. We identify the organization information when available. We use email content to understand the context of the communication. And finally we utilize the time of the communication. Our system generates useful metrics (Discussed in detail in Chapter 4) using the information provided in the email family to provide visual cues to the manual reviewer during review. We then ask lawyers to label the email family as *Privileged* or *Not-privileged* or *Unsure*. We perform a hypothesis test by providing the lawyers two interfaces; one without any automation as a baseline condition and the other with automation as a treatment condition. We evaluate the accuracy of each of the reviewer's judgment by considering one of the senior attorney's judgments as gold standard.

RQ4b: Does the manual reviewer's review speed improve when system-generated features are presented during privilege review?

A substantial amount of cost in e-discovery results from the process of manual review for privilege. Due to the high-stakes in the privilege review process, review for privilege is usually performed by senior attorneys. As a result privilege review costs more money when compared to the review for responsiveness (which is usually performed by junior lawyers). Our motivation to measure the review speed was designed to indirectly measure the review cost as attorneys are usually billed by the hour. In e-discovery, manual review time is proportional to manual review cost.

To answer RQ4b, we study if the users perform the review faster in the presence of our system-generated features. We do this by recording the time-stamp of each event the

user performs. We record the duration spent by each user to review each email message. We run statistical tests to determine the difference in the average review speed. We carried out a paired t-test across the baseline interface and our treatment interface to compare the average speed of the privilege review task over the two sets of observations. Our findings reveal no significant difference in review speed.

RQ4c: Which system-generated features do the manual reviewers believe are most helpful?

To answer the question *RQ4c*, we performed a subjective evaluation using the Questionnaire for User Interaction Satisfaction (QUIS). Our questionnaire aim to measure the system satisfaction along multiple interface factors (screen factors, learning factors and system capabilities) on a 9-point scale. During a semi-structured interview, participants reported that the actor role and identity features exposed by the system were most useful to them, and that the present implementation of features based on content or date added no discernible additional value. Quantitative results indicate that substantial and statistically significant improvements in recall were achieved.

The scope of the research questions thus far were limited to the privilege review phase. However, review time is a factor that applies to both responsiveness and privilege review phase. In our next two research questions, we model the use of predictive coding system for the entire e-discovery review process. We consider both the review for responsiveness and the review for privilege. The main objective for the next two research questions is to understand how predictive coding can aid in making the e-discovery process more effective at the lowest possible incremental cost. We run our experiments on a different test collection to avoid the sampling challenges encountered during the creation of the TREC 2010 Legal Track collection.

1.2.3 Predictive Coding with Manual Review

All the above research questions are specifically aimed to tackle the privilege review phase in e-discovery. Our initial questions *RQ1, RQ2, RQ3a and RQ3b* aim at building a predictive coding system to identify privileged documents while *RQ4a, RQ4b and RQ4c*

concentrate on manual privilege review.

Findings from *RQ4b* motivate our last couple of research questions. Findings from *RQ4b*, reveal that having a lawyer look at the documents with features generated by our algorithm yields no improvement in review time. As time is directly proportional to money during privilege review, this process can be quite expensive. Consequently, we can infer that a fully manual review is not sustainable. Thus we aim to develop a semi-automated system where we utilize the manual reviewer's time only when it is cost-effective. The analysis addresses a ternary classification problem. We propose a semi-automated system whose goal is to identify, within a set of documents D, the documents that are at the same time (a) responsive to a certain topic, and (b) non-privileged. Documents that are both responsive and non-privileged should be produced by the producing party to the requesting party; documents that are responsive and privileged should be declared in a privilege log; non-responsive documents should be withheld.

We aim to make the review process of e-discovery more efficient. Using a fully manual system incurs huge amounts of manual annotation cost during review. Hence our goal is to involve the human only when the document review cost is smaller than the expected cost of accepting the decision of the automatic classifier (we call this as risk). We aim to achieve a reduction in cost by using a semi-automated system where-in we make use of a predictive coding system to automatically classify all the documents in the test set and use the reviewers to manually check the label for a document only when the risk involved in accepting the decision of the automatic classifier exceeds the review cost.

To understand how our semi-automated system can improve the efficacy of the review process, we first develop a ranking algorithm to determine which documents in the test set need a manual review. We then compute the overall expected cost of the review process. In our semi-automated system, the documents that are not manually labeled by the reviewers, use the classifier predictions as labels. Hence we have two types of review costs; (1) Cost incurred due to manual review and (2) Cost incurred due to classifier misclassifications. To model the cost of misclassification error, we quantify the different e-discovery outcomes in terms of liability cost. Our input cost structure is formed on the basis that some mistakes are more severe than others. Besides, if the probability of making

that type of mistake is small, the expected cost for making any one decision will also be small. To compare our system performance with other effective baselines, we develop a linear evaluation function where the total expected cost for the review is simply the sum of the expected costs of each of the outcome. One of the research questions we ask in this component is:

RQ5a: Which documents need to be manually reviewed?

We ask this question because we know that there are some cases where adopting some degree of automation is the best solution. By answering this question, we aim to help e-discovery practitioners decide when and to what extent adopt automation during privilege review. The classifier model we aim to build, balances for the cost of review and the risk of compromising a privileged document.

Our ranking algorithm utilizes the posterior probabilities and the cost of making a mistake. Unlike the traditional classification processes, the outcomes of our classifier vary significantly in terms of prediction errors; i.e., some type of classification errors are considered to be more acceptable than others. We first quantify the type of prediction errors as a representation of liability costs. We map the misclassification errors to a cost value. We develop a risk based ranking algorithm to determine which document needs to be reviewed by a human depending on the expected cost associated with each document. If the expected cost of accepting the decision of the automatic classifier is higher than the cost of manually reviewing that document, then it would be rational to manually review that document. And conversely, if the cost of reviewing a document exceeds the expected cost, then it would be rational not to manually review that document. The approach we take is to run the classifier on every document, sort the documents in decreasing order of the expected cost and then manually review documents from the top of the list as we go, until we reach the first document for which the expected cost is less than the cost of reviewing that document. We next ask our final research question;

RQ5b: Does our semi-automated system yield lower overall expected cost when compared to other baseline models?

To answer this question, we develop a suitable evaluation measure that is optimized for review cost. We define multiple effective baselines. Our baseline methods are of different types; completely automated, completely manual solutions and human-in-the-loop systems. Their classification decisions are obtained via some combination of manual annotation and automatic classification. We compute the overall expected cost for each of the baselines and our semi-automated system. Using the cost structures exemplified in 5 we can evaluate each system by computing the overall expected cost for all the seven models.

1.3 Thesis Statement

For the task of identifying privileged documents intermixed with responsive material during discovery, automation can be used to improve efficacy, accuracy, or both.

1.4 Dissertation Outline

The remainder of this dissertation is structured as follows: Chapter 2 discusses the background with related work. In Chapter 3, we discuss the research question related to building a classifier, evaluation of the collection and highlight with detailed experiments the drawbacks of the TREC Legal Track 2010 test collection. We provide a fix for the drawbacks and describe the use of the test-collection. Then in Chapter 4, we attempt to seek users' (lawyers) help to determine how our work could help them perform the task of privilege review better. We design and conduct a user study and a semi-structured interview with 6 legal professionals. Next in Chapter 5, we introduce our risk-minimization framework to show when and which document in the collection needs human input. We define and develop six baseline models and compare all the models with our model. We conclude in Chapter 6, with experimental limitations, looking to future directions, and articulating some of the broader impacts of our work.

Chapter 2: Background

The work reported in this dissertation is related to multiple lines of research. Section 2.1 introduces the research domain and its background. This dissertation uses the TREC Legal Track 2010 test collection. In section 2.2, we explain the related work done in the area of evaluating test collections along with the necessary background about the TREC test collection utilized for our experiments. We discuss prior work about manual review in section 2.3, interactive review that supports our study (discussed in chapter 4) in section 2.4 and the predictive coding techniques in section 2.5

2.1 E-Discovery and Privilege Review

In the United States, civil lawsuits generally proceed through distinct steps: pleadings, discovery, trial and possibly an appeal. E-discovery is a process in which a producing party involved in the lawsuit is responsible to produce all the relevant electronically stored documents to the requesting party. The legal professionals are increasingly confronting a new reality: massive and growing amounts of electronically stored information (ESI) required to be retained by law, in anticipation of litigation. Spotlight has now formed on how lawyers decide to meet their obligations in various e-discovery contexts. One major aspect involves the study about how researchers adopt technology to identify relevant electronic evidence in response to a discovery requests or due to some other external demand for information coming from a requesting party. Research on the process is increasingly important, given the legal costs. The e-discovery cost grows exponentially as a portion of relevant documents need to be protected due to the existence of *Privilege* or *Attorney work-product*. In litigation, there are many types of Privilege namely:

- Legal Professional Privilege or Attorney-Client Privilege

- Public Interest Privilege
- Without Prejudice Privilege
- Privilege Against Self-Incrimination
- Others

Attorney work-product is a doctrine that protects from discovery, the materials prepared by the attorney or attorney's representative [33]. In this dissertation, we conduct experiments that concentrates on this type of privilege.

In legal context, *Attorney-Client privilege* is a right given to the parties in a lawsuit to provide protection against the involuntary disclosure of information. Attorney-client privilege in particular exists to protect the information exchange between *privileged persons* for the purpose of obtaining legal advice. Privileged persons include [33]:

- 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

Since the 2006 amendments to the FRCP, the task of withholding documents on the basis of attorney-client privilege alone has faced multiple challenges in litigation [34, 48]. The attorney-client privilege is aimed to foster trust and promote at-will communication between the parties and their attorneys. However, privilege does not arise simply because privileged persons communicate; it can only be claimed when the content of the communication merits the claim. For example, an email from Jeff Skilling (Enron's president) sent only to James Derick (Enron's general counsel) about pending litigation would be privileged; an email with the same content sent to both James Derrick and a personal friend of Skilling's who was not involved in Enron's business operations would not be, and an email from James Derrick to Skilling that indicated (only) his intent to resign in order to spend more time with his family also would not be privileged.

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 privileged entities has been made in confidence for the purpose of obtaining legal advice, the existence of privilege can be waived due to the involvement of a third party [2] or sometimes even due to inadvertent disclosures. In practice, inadvertent disclosures appear at greater frequency [1,4,33]. 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 lawsuits 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. 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 [62]. Such numbers would cause the cost of the review process to grow quickly with the increase in collection size, making linear review impractical [60].

As linear review is becoming impractical, this dissertation attempts to determine if adopting automation to some extent can help in reducing the cost of review.

2.2 Test Collection Evaluation

The modern literature on the effectiveness and reliability of retrieval experiments is largely confined to the problem of constructing test collections for IR evaluation. The Text Retrieval Conference (TREC) was created to address the problem of IR evaluation for large datasets. TREC typically follows the Cranfield paradigm [76], which evaluates the results of participating systems against a gold standard that identifies every relevant document.

A test collection consists of documents and assessments of which documents are relevant to. These relevance assessments are made by human assessors. Depending on

the collection, some documents have multiple human judgments. Effectiveness measures are then calculated based on the return of relevant documents by systems under evaluation [77]. Gathering human relevance assessments is one of the most expensive and problematic aspects of test collection formation. Human judgment is subject to various cognitive, perceptual and motivational biases [61]. Researchers identify multiple factors that influence evaluation of test collection: documents; judgment conditions; judgment scales; and factors like human expertise [51]. Saracevic [64] surveys experimental work on these factors. Analysis by Voorhees [75] shows that while absolute effectiveness scores are sensitive to variations in relevance judgments, relative scores remain broadly stable. In e-discovery, evaluating the absolute effectiveness matters at least as much as than systems' relative scores.

The traditional test collection methodology assumes that all documents in a collection are judged in response to every query in the test set. As collection sizes have grown, exhaustive assessment has become infeasible. Evaluation campaigns such as TREC therefore make use of a pooling approach, where documents for assessment are taken from the answer lists of participating systems. Zobel [52,89] finds pooling robust in determining relative system rankings, but incomplete in identifying all relevant documents. Subsequent work has suggested that for very large collections, pooling may be unreliable even for relative comparisons [20, 21, 81]. Yilmaz and Aslam propose the simple random stratified sampling method [85] [87]. Pooling of results introduces bias against unpooled systems because distinctive documents returned by these systems are assumed to be irrelevant. A possible fix is to ignore unassessed documents in calculated metric scores. This was proposed by Buckley and Voorhees [21]. There has been considerable recent interest in techniques for the efficient estimation of effectiveness metrics. Yilmaz and Aslam [86] introduce infAP, a method for estimating average precision using uniform sampling from the set of complete relevance judgments. A refinement is statAP, which uses stratified sampling requiring smaller sample sizes than infAP for the same accuracy [23]. Stratified sampling was also used in the TREC Filtering Track [45].

Chapter 3 of this dissertation explores the reusability of the TREC Legal Track 2010 test collection. We utilize the collection and address multiple issues related to (1)

Use of the judgments from different assessors for building and evaluating classifiers and (2) adjudication conditions. In the next section we discuss the details about how that test collection was created.

2.2.1 TREC Legal Track Collection

The first effort at creating a platform for e-discovery domain research and evaluation was initiated by the TREC Legal Track after the 2006 amendments to the Federal Rules of Civil Procedure (FRCP). The principal goal of the TREC Legal Track was to develop multiple ways of evaluating search technology for e-discovery [13]. Keyword search approach was one of the initial attempts taken to help the lawyers manage the enormous amounts of documents [14]. 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. Some extensions to keyword search approach called concept search are employed to extend the search terms to include context information [44]. However, as corporate collections have continued to grow, filtering by keywords have left huge document sets to be linearly reviewed [19] making linear review procedure insupportable [60].

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 [55]. Thus use of automated classifiers with a higher degree of technological assistance using machine learning techniques is currently being studied [35]. Although many types of electronically stored 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 received by its employees. Since email collection one avenue to search for communications that could be withheld on the grounds of attorney-client privilege, we utilize the relevance and privilege judgments obtained from TREC 2010 Email Test Collection in our experiments reported in chapter 3 and chapter 4.

TREC 2010 Legal Track focuses on evaluation of search technology for discovery of ESI in litigation and regulatory settings. The TREC 2010 Legal Track consisted of

two distinct tasks: the Learning task, in which participants were required to estimate the probability of relevance for each document in a large collection, given a seed set of documents, each coded as responsive or non-responsive; and the Interactive task, in which participants were required to identify all relevant documents using a human-in-the-loop process. We used Interactive Task topics for our experiments.

2.2.2 Topics & Assessment

In the 2010 TREC Legal Track's "Interactive task",[1] one off the three relevance topics (Topic 303) required finding *"all documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto."* [29]. And the privilege topic in the 2010 TREC Legal Track[2] requested *" all documents or communications that are subject to a claim of attorney-client privilege, work-product, or any other applicable privilege or protection"*. Although privilege classification is normally performed as a second pass after classification for relevance, nothing in the definition of privilege is specific to any litigated matter. The collection to be searched was version 2 of the EDRM Enron Email Collection, which includes both messages and attachments. The items to be retrieved were "document families," where (following typical practice in e-discovery) a family[3] was defined as an email message together with all of its attachments.

Once the submissions from the participants were received during TREC 2010, the collection was stratified for each topic and evaluation samples were drawn. Stratification followed the pooling-based design whereby one stratum was defined for email families all participants found relevant (the *All-R* stratum), another for email families no participant found relevant (the *All-N* stratum), and others for the various possible cases of conflicting assessment among participating teams. The operative unit for stratification

[1] A task in which participants design both a system and an interactive process for using that system

[2] For bookkeeping purposes, the (non-topical!) privilege task was Topic 304.

[3] Use of families is referred to as "message" evaluation in [29].

was the document family, and families were assigned intact (parent email together with all attachments) to strata. Samples were composed following the allocation plan whereby strata are represented in the sample largely in accordance with their full-collection proportions. An exception to proportionate representation was made in the case of the very large All-N stratum, which is under-represented in the sample relative to its full-collection proportions. To manually gather relevance and privilege assessments, selection within each stratum was made using simple random selection without replacement. The process of gathering assessment followed a two-stage procedure, whereby an initial relevance assessment is made of each document in each evaluation sample and then a selection of those first-pass assessments are escalated to a subject matter expert or Topic Authority (TA) for final adjudication.

Once the evaluation samples were drawn, they were made available to review teams for first-pass assessment. The review teams, were all staffed by commercial providers of document-review services. At the outset of each review team's work, an orientation call was held with the Topic Authority for the team's topic; on the call, the Topic Authority outlined his or her approach to the topic, and the review team had the opportunity to ask any initial questions it had regarding the relevance criteria to be applied in assessing documents. Finally, once the review got under way, an email channel was opened, whereby the review team could ask the Topic Authority any questions that arose, whether regarding specific documents or regarding the relevance criteria in general, in the course of their assessment of the evaluation sample.

Dual assessments were gathered on a subset of the families. The dual-assessment subset was chosen by random selection from families already included in the sample. Both assessments were supplied by the same review team; indeed, it is not impossible that, in some cases, the same individual supplied both assessments. What we can say about the dual assessments is that they represent distinct assessments of the same message on two different occasions.

A set of first-pass assessments are escalated to the Topic Authority for adjudication. The families that are escalated are derived from multiple sources: (1) First-pass assessments can be appealed by one or more of the participants (2) Assessments with dis-

agreements (only those that are dual assessed) and (3) A sample of non-appealed families with first-pass assessments. Once selected, the families were made available to the Topic Authority for final assessment. In making their assessments, the Topic Authorities had access to the assessment guidelines they had prepared for the first-pass assessors, as well as any other materials they had compiled in the course of their interactions with the participants. Once the Topic Authorities had completed their reviews of their adjudication sets and the sample assessments had been finalized, the relevance judgments gathering process were deemed complete.

2.3 Manual Review

In e-discovery, documents that are initially marked as responsive to a production request (i.e., a specific request for evidence by the counterparty) are typically subjected to a linear manual review for privilege in order to be sure that content that could properly be withheld is not inadvertently revealed. Failure to identify a privileged document could jeopardize the interests of the party performing the review, so it is common practice to have highly qualified (and thus expensive) lawyers perform the privilege review. However, it is well known that human assessors frequently disagree on the relevance of a document to a topic. Experienced TREC assessors working from only sentence length topic descriptions, had an average overlap (size of intersection divided by size of union) of between 40% and 50% on the documents they judged to be relevant [75]. Voorhees concludes that 65% recall at 65% precision is the best retrieval effectiveness achievable, given the inherent uncertainty in human judgments of relevance. Bailey et al. [12] survey other studies giving similar levels of inter-assessor agreement. One way of characterizing accuracy is by measuring inter-assessor agreement, which has consistently proven to be lower than one might expect [75,79]. When searches are done by different users, disagreement might reflect different notions of relevance or, in our application, different ways of reaching decisions regarding privilege. Reasons for disagreement between different relevance assessors, such as the instructions given to judges or the different topics have also been analyzed [79, 83]. In e-discovery, however, there is a single senior attorney who ultimately certifies the

correctness and completeness of the review process, and their interpretation of privilege is thus taken to be authoritative [58].[4] The Interactive Task of the Legal Track of TREC includes such a topic authority, and provides a process of appeal to this authority for uncovering assessor errors. The appeal results for TREC 2009 found that, on an assessment set in which 90% of documents were actually irrelevant, 33% of relevant assessments were in error, as were 3% of irrelevant assessments [37]. This is likely a lower bound to the error rate, since some errors may not have been appealed (although conversely some appeals may have been erroneously upheld).

Carterette and Soboroff have found that when judgments from one person are used to predict system preferences that would be obtained by computing evaluation measures using the judgments of another person, the quality of the prediction can be enhanced by selecting a relatively conservative assessor (i.e., one that has a lower tendency to make a false positive error) as the source of judgments that are the basis for the prediction [24]. This is an intriguing result for our privilege review task because in privilege review it is the risk of false negative errors that would generate the greatest concern on the part of the party performing the review.

The Legal Track of TREC provides an objective environment in which to validate and compare different retrieval methods for e-discovery [15]. Two other known studies have compared the quality of automated retrieval and manual review, one by a re-review of an earlier manual production [62], the other through an analysis of data from the TREC 2009 Legal Track [36]. The former study finds automated retrieval to be at least as consistent as manual review, while the latter concludes that automation gives superior reliability. While there has also been some work on the design and evaluation of automated classifiers to actually perform the privilege review task [29, 35, 74], there is a widely held belief among attorneys that (absent compelling reasons to the contrary such as a need for privilege review at a scale that would otherwise be impractical), reliance on a fully automated classifier for privilege review would incur an undesirable level of as-yet uncharacterized risk.

[4] This certification can itself be litigated; in such cases the court would make the authoritative determination.

Thus automated classifiers are more often used for consistency checking on the results of a manual privilege review process than as the principal basis for that review. A part of the work in this dissertation (Chapter 5) explores a second possible use of the technology. That is, use of automated annotations to (hopefully) improve the accuracy and reduce the cost of a manual review process.

2.4 Interactive Review

Chapter 4 focuses on building tools that can help lawyers to make faster and more accurate privilege judgments. We do that by scoring the importance of specific email addresses to determine each actor's propensity to engage in privileged communication. We choose email messages along with the attachments as our document collection because much of the activity of most organizations is ultimately reflected in the emails sent and/or received by its employees. For this reason, email provides an excellent environment to initially develop techniques to improve the productivity and accuracy of privilege review when it is rational to conduct it manually [57].

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") [31, 49]. 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 [66]. McCallum et al took an initial step towards building a model that captures actor roles and email relationships using dependencies between topics of conversation [50]. Since then, several other generative models have been proposed [78, 88]. 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 [41], litigation [22], 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 using metadata information and the email content information to build features for our classification system. To evaluate classifiers, availability of reliable annotated data is required.

The process of gathering reliable annotations are fraught with multiple problems. In e-discovery, one such problem is the requirement for skilled legal annotators for review who make the review process more expensive. Thus, 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 reliability of the retrieval effectiveness while evaluated on expert judgments [82] and empirical findings have shown that annotations from experts would lead to better classifier accuracy [12]. However, Cheng et al. describes the benefits of utilizing noisy annotations to enhance classifier performance in a multiple annotation type environment [25]. Thus it is reasonable to accept that many factors like sampling, annotator expertise, etc., affect the process and quality of gathering relevance assessments. Although it is not realistic for human annotators to be infallible [75, 79, 80], we make the assumption that human annotators to be infallible in chapter 5.

2.5 Predictive coding with cost-sensitive learning

Automating the process of search, analysis, and review are different tasks with different objectives. The objective of search is to find enough documents to satisfy an information need, such as a request for documents that are relevant to a topic. However, one of the results from the TREC legal track is that many relevant documents are missed by the best present search methods [15, 69]. Thus, automated methods for retrieving relevant documents could take advantage of predictive coding techniques and ranking algorithms. The state-of-the-art in the application of predictive coding technique in e-discovery is reviewed in [58], and has been the subject of many recent studies [27, 28, 36, 62, 63, 74]. Predictive coding for privilege classification has been recently addressed [35, 71, 73]. Four recent cases has brought predictive coding techniques to the forefront [5–8]. These cases have attracted considerable attention in law and technology blogs.

Attorneys, typically senior attorneys, work to train or calibrate the predictive coding system. Most of the prior studies on predictive coding technique like [27, 28] begins either with attorneys selecting a seed set of responsive and nonresponsive documents, or

reviewing and coding a random sample of documents. These initial documents are then analyzed by the predictive coding system. The system begins to make judgments on probable relevance of other documents. The attorneys review further samples produced by the system, again applying their own judgment as to relevance, responsiveness, and privilege. The process continues until the attorneys are satisfied that the software is properly calibrated. At that point, the results are said to be optimized.

In Chapter 5 of this dissertation we use predictive coding system to optimize for the overall cost of the e-discovery process. We do this by limiting the total number of documents that need to be manually reviewed for relevance and for privilege. We quantify the different types of classifier errors in terms of costs. Depending on type of the classifier error, the cost value varies. In other words, the cost value of each mistake is sensitive to the type of errors produced by our predictive coding system. It is thus important to take the cost of every type of error into account so as to avoid the costliest of errors.[5] Some of the principles applied in our work are described in [16].

We utilize the idea of gain presented in [16] for ranking automatically classified documents in order to optimize the work of human reviewers who annotate some of them. One major difference is that [16] is more theoretical, while this dissertation work can be seen as an application to an e-discovery context. The cost matrix emerges from the evaluation function (e.g., F_1), which is given as an input to the problem [16], while in our model it is the evaluation function which emerges from the cost structure, which is given as an input to the problem.

The framework we discuss in chapter 5 employs cost-sensitive active learning. The work related to ours are [17, 40, 70], where the cost of manually annotating a document is an explicit variable in a model that ranks items for presentation to a human reviewer. However, the goal of [17, 40] is not prioritizing the documents whose annotation would bring about the highest reduction in overall cost, but annotating the documents that would prove most valuable when used as training examples for retraining the classifier. In other words, the task we deal with is not retraining the best possible classifier, but

[5]The work discussed here is currently under review and was done in collaboration with Douglas W. Oard and Fabrizio Sebastiani; Minimizing the Expected Costs of Review for Responsiveness and Privilege in E-Discovery [54].

reviewing a set of documents at the minimum possible overall cost; this difference in goals shapes the difference between that technique and our model. Other work in cost-sensitive active learning (e.g., [32, 65, 68]) are even more different from ours since they focus on modelling the fact that different types of items may involve different annotation costs, and an issue that we do not address in our model.

The next chapter explores the first component of our dissertation: Designing and building a predictive coding system to identify privileged documents.

Chapter 3: Predictive Coding

In e-discovery, the task of withholding documents on the basis of privilege (attorney-client privilege or attorney work-product doctrine) has surfaced many challenges in litigation cases [34] [48]. As more and more litigators today are familiar with the use of automated classifiers, the effectiveness and evaluation of such classifiers has gained the interest of not only commercial e-discovery vendors but also the courts [55]. This chapter details the contribution of building and evaluating privilege classifiers using the only existing test collection[1] . As a first step, we develop a test set to enable fair classifier evaluation. We next evaluate how reliable and reusable the existing test collection is. We finally build binary privilege classifiers that utilize the judgments from the test collection [74].

Evaluation of information retrieval systems relies on test collections in which relevance judgments can be created for only a small portion of the collection [67]. One approach for evaluating our systems is using a test collection with relevance and privilege assessments. Since collection of a realistic size are too large to exhaustively evaluate for relevance and then for privilege, the approach taken instead is, to assess documents that are highly ranked or retrieved by at least one retrieval system. By selecting these documents the focus is on those documents found by the systems that are to be compared. This approach, known as pooling, has been widely used in the Text Retrieval Conference (TREC) and elsewhere.

The procedure of pooling the system results during the TREC 2010 Legal Track collection created two major concerns. We study the first one by asking a question about reliability since different assessors may reach different judgments for the same document.

[1]The work discussed in this chapter is published in SIGIR and ICAIL conferences and was done in collaboration with Douglas W. Oard and Jiaul Paik; Assessing the reliability and reusability of an E-discovery privilege test collection [74] and Evaluating expertise and sample bias effects for privilege classification in e-discovery [71].

Voorhees has shown that absolute measures of effectiveness are sensitive to this effect but that relative comparisons between systems are relatively insensitive to inter-assessor disagreement [75]. A second concern, reusability, is that new systems will generally find some documents that did not contribute to the pool, and assuming such documents not to be relevant might adversely affect even relative comparisons. Reusability is important because reusable test collections allow the cost of relevance judgments to be amortized over future uses of a test collection. Reusability of pooled judgments was examined by Zobel [89], who found that TREC pooling had likely found no more than half of the relevant documents, but that relative comparisons remained reliable. Buckley et al. [21] later highlighted a key limitation of that conclusion, finding that when distinctive systems had contributed to the pool, removing one such system could yield a substantial adverse effect on measurements of mean average precision. One way to partially address this concern, introduced by Yilmaz and Aslam, is to sample the documents to be judged from the full collection and then to estimate the evaluation measure from the sampled judgments [85, 87].

Random samples drawn from very large collections yield confidence intervals that are so large as to be uninformative, so in this chapter we explain and focus on the sampling design used in the interactive task of the TREC Legal Track, in which set intersections were used as a basis for stratification [56]. Between 2006 and 2011, the TREC Legal Track created relevance judgments for more than 100 topics (which in e-discovery are called "production requests"). In 2010, this was augmented by the world's first (and to date only) shared-task evaluation of privilege classification [29].

We study the reliability and reusability of the resulting test collection. When working with Legal Track test collections we need to think a bit differently about reliability and reusability. For reliability, we are interested not just in relative comparisons, but also in the reliability of absolute measures of effectiveness, and most particularly in estimates of recall. Point estimates from samples are (in expectation) insensitive to sample size, so characterizing the reusability of stratified samples requires comparing confidence intervals for systems that did and didn't contribute to the stratification. What we call reliability thus corresponds to the statistical concept of measurement error, reusability to the statis-

tical concept of sampling error. We then study the effect of building classifier by training privilege classifiers on two sets of families using the TREC Legal Track test collection.

3.1 Test Collection

This section introduces the required background to answer our research question *RQ2*. In the 2010 TREC Legal Track, the document collection used for all Interactive tasks (including the privilege detection task) was derived from EDRM Enron Collection, version 2, which is a collection of Enron email messages. The privilege task during TREC Legal Track 2010[2] was to retrieve to withhold "all documents or communications that are subject to a claim of attorney-client privilege, work-product, or any other applicable privilege or protection". Although privilege classification is normally performed as a second pass after classification for relevance, nothing in the definition of privilege is specific to any litigated matter. The collection to be searched was the EDRM Enron collection, version 2, which is a collection of Enron email messages for which text extracted from attachments is provided with the collection. Following the practice for privilege review in e-discovery, the items to be classified were "document families," in this case a family was defined as an email message together with all of its attachments.[3]

3.1.1 Stratified Sampling

Two teams (A and H)[4] submitted system results (runs) for the TREC 2010 privilege classification task: Team A submitted four runs ($a1$, $a2$, $a3$, $a4$); Team H submitted one ($h1$). Each run was a binary assignment of families to one of two classes: privileged or not privileged. Following TREC convention, we refer to these five runs as participating systems; each run was produced by people and machines working together (TREC refers to this as interactive task). The collection was partitioned into 32 strata, each defined by a unique 5-bit vector (e.g., 01010 for the stratum containing families runs $a1$, $a3$, and $h1$ classified as not privileged and runs $a2$ and $a4$ classified as privileged) [29]. The 00000

[2]For bookkeeping purposes, the (non-topical!) privilege task was "topic 304."
[3]Use of families is referred to as "message" evaluation in [29].
[4]In [29] Team A was called CB, team H was called IN.

stratum included 398,233 of the 455,249 families (87% of the collection), but only 3,275 of the 6,766 samples (48%) were allocated to that stratum. The resulting sampling rate for the 00000 stratum (0.8%) was far sparser than for any other stratum (which averaged 6.1%). The allocation of samples was a bit denser for smaller strata since a 6.1% sampling rate might otherwise result in very few samples being drawn. Few samples were allocated to these very small strata in aggregate, so the sampling rate remained above 6% for every stratum other than the 00000 stratum.

3.1.2 Privilege Assessments

First-tier junior level privilege assessors (henceforth, assessors), who were lawyers employed by a firm whose business included provision of document review services for e-discovery, were provided with detailed guidelines written by a senior privilege review attorney (the Topic Authority (TA)). Assessors recorded ternary judgments: privileged, not privileged, or unassessable (e.g., for display problems, foreign-language content, or length). As expected, assessors sometimes made judgments that disagreed with the TA's conception of privilege. For other tasks, differing judgments might be treated as equally valid, but in e-discovery the TA's judgments are authoritative (because the TA models the senior attorney who will certify that the review was performed correctly). Judgments that disagree with those of the TA are therefore considered incorrect. In TREC 2010, an assessor's judgment regarding whether a family should be classified as privileged could be escalated to the TA for adjudication in three ways. First, a team might appeal the decision of an assessor to the TA. A total of 237 such appeals were received. Of course, teams might not as easily notice, nor would they rationally appeal, assessor errors that tended to decrease their estimated classification accuracy. In particular no team would rationally appeal an erroneous assessor judgment of privileged in the 11111 stratum, nor an assessment of not privileged in the 00000 stratum. The set of appealed judgments is thus biased [80]. To create an unbiased sample, 223 assessor judgments were thus independently drawn using simple random sampling. Since this is a random sample of a stratified sample, it results in a smaller stratified sample of the full collection. To facilitate symmetric comparisons among assessors, a second simple random sample containing 730

Table 3.1: TA adjudication rates.

Category	Assessed	Adjudicated	Rate
Random sample	6,766	223	3.3%
Team appeal	6,766	237	3.5%
Assessor disagreement	730	76	10.8%

Table 3.2: Training Families

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

families was drawn and each family in that sample was duplicated in the set of families to be assessed. This was done in a manner that had been expected to result in the duplicated families being assigned to different assessors.[5] When conflicting assessments for a family were received, the judgment was adjudicated. Table 3.1 summarizes the selection process. Families chosen to be adjudicated by the TA will henceforth be called as Adjudicated Set or AS; Families that are not adjudicated as Non-Adjudicated Set or NAS.

3.2 Evaluation Plan

In this section, we answer our research question *RQ1* and introduce the framework for the questions *RQ3a and RQ3b*. We study the effect of training privilege classifiers on two sets of families. We utilize the 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 3.1 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[6], we randomly draw 2 families from appeal (A) category and one from assessor

[5]Some pairs may have been judged by the same assessor.
[6]This is the sampling probability of the random category.

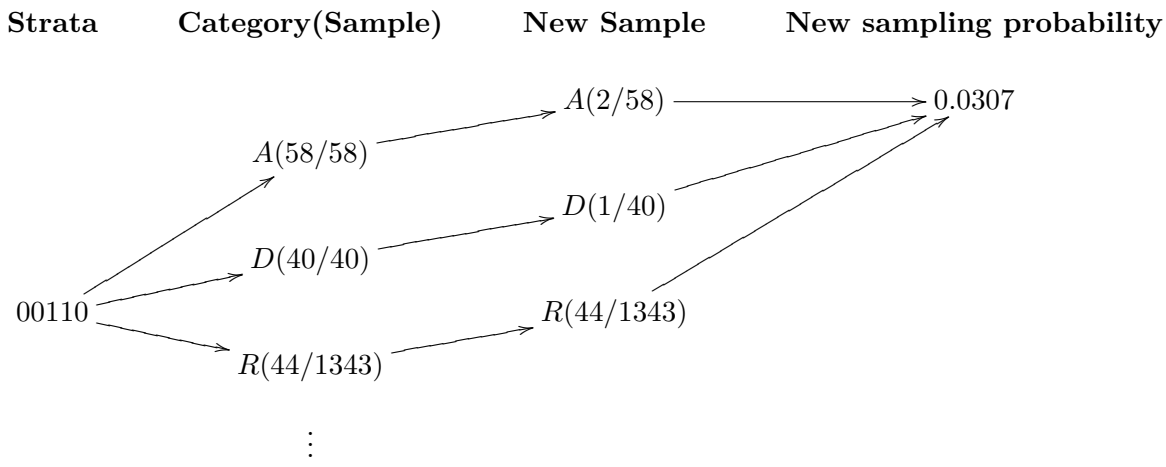

Figure 3.1: Re-sampling Procedure

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 [74]. The remaining families in the AS and NAS are used for training our classifiers.

Figure 3.2 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 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 [74]. 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 3.2 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 3.3. 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 assess-

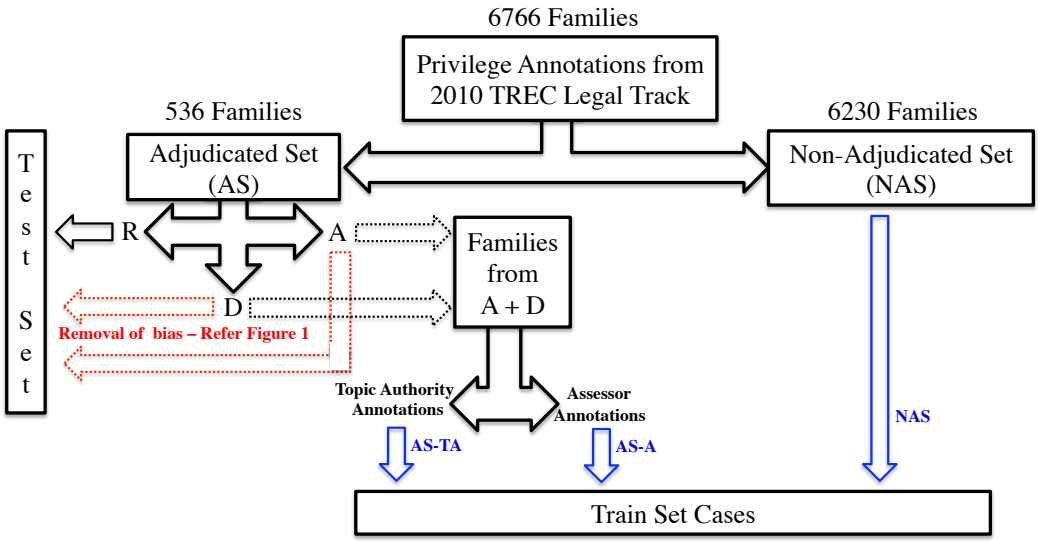

Figure 3.2: Train-Set and Test-Set Split Procedure

ments from TA (*model*-AS-TA[7]) and the assessors (*model*-AS-A[8]) to study the effect of expertise on classifier training. All families in the NAS are annotated by only assessors.

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.

3.3 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 important in finding privilege, we use both the network and

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

Table 3.3: Separation of email data

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

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 3.3.

The first view $view_1$ exploits the metadata[9] 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 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.[10] 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.

[9]Data in From, To, Cc and Bcc fields
[10]http://scikit-learn.org/stable/

3.3.1 Models

In this section, we describe the models in detail. We explain the estimation and interval calculation.

3.3.1.1 Graph Model

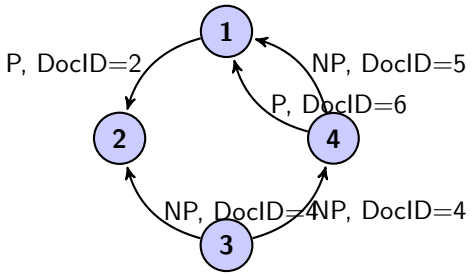

Figure 3.3: Sample Graph

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 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 3.4. 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 [18]. 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",

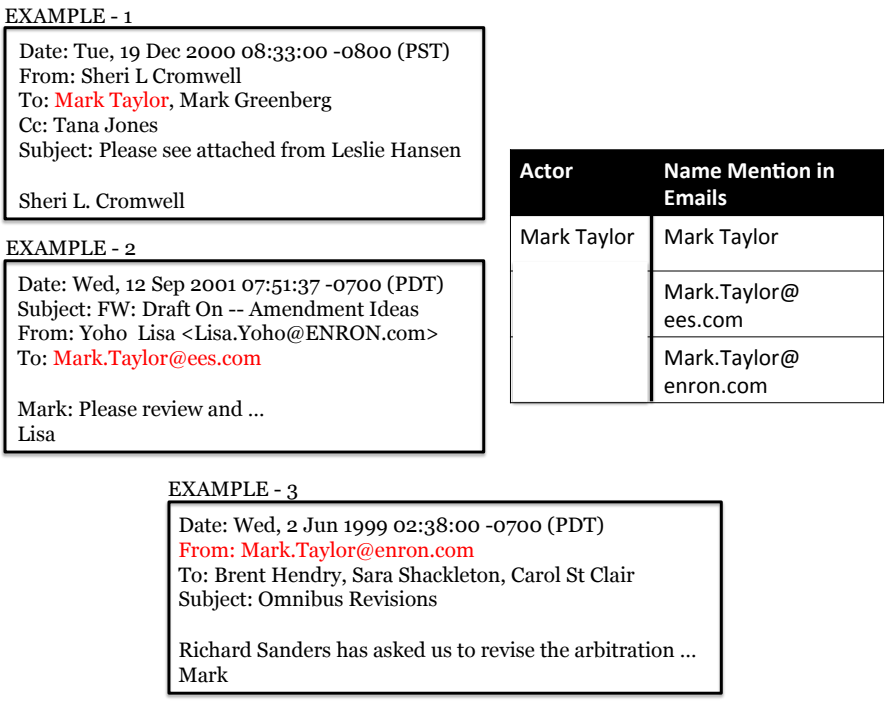

Figure 3.4: Actor variants in emails

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 [18]. We used the implementation provided by the Python "difflib" module with the cutoff threshold set to 0.75. For the examples shown in figure 3.4, 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 sub-sequences. 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.

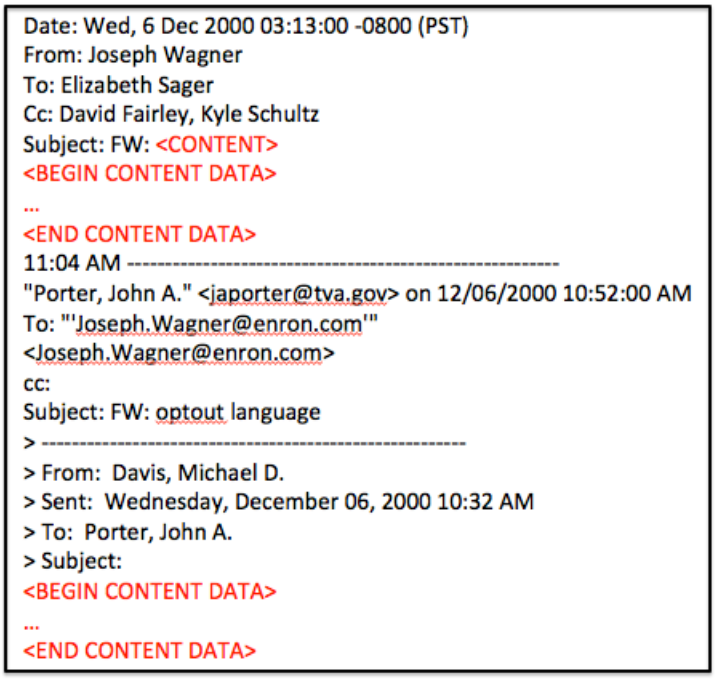

Figure 3.5: Content-centric information in emails

3.3.1.2 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 researchers 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 3.5) included in the body of the email message. Figure 3.5 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 a TF-IDF weighting algorithm.

Table 3.4: 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

3.3.2 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 3.4). Section 3.3.2.1 and section 3.3.2.2 explain the derivation of point estimates and confidence intervals respectively.

3.3.2.1 Point Estimate

This section 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

$$\hat{N}^i_{rp} = \sum_{h:T_i \in T^h} \hat{N}^h_{rp} \tag{3.1}$$

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 privilege. Then, an unbiased estimator of N^h_{rp} is

$$\hat{N}^h_{rp} = N_h * \frac{n^h_r}{n_h} \tag{3.2}$$

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

$$\hat{Precision}^i = \frac{\hat{N}^i_{rp}}{N^i_r} \tag{3.3}$$

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:

$$\hat{N}_p = \sum_{h:T_i \in T^h} \hat{N}^h_p \tag{3.4}$$

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

$$\hat{Recall}^i = \frac{\hat{N}^i_{rp}}{\hat{N}_p} \tag{3.5}$$

3.3.2.2 Confidence Intervals

The recall and precision values derived in section 3.3.2.1 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 [9]. 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. In the experiments reported in this chapter, we calculate 95% approximate confidence intervals from beta-binomial posteriors on stratum yields [**?**].

3.4 Results

In this section we report the results of RQ2 (section 3.4.1), RQ3a and RQ3b (section 3.4.2).

3.4.1 Test Collection Bias

Here we analyze the reliability and reusability of the TREC 2010 Legal Track privilege task test collection.

Analysis of Measurement Error

The use of assessor judgments for families that the TA had not adjudicated would be reasonable if the appeal process had identified most of the assessor errors. This is a testable hypothesis. Although the TA might also make errors, we ignore that factor because we believe its effect to be small. We therefore treat the TA's judgments as a gold standard. As a further simplification, we treat the small handful of unassessable documents (13 families) as not privileged in our analysis. One way of visualizing the effect of assessor errors is to use only some or all of the families that were selected for adjudication, plotting confidence intervals using TA judgments in one case and using assessor judgments in the other. The adjudicated sample is less than 8% of the size of the full set of official judgments, so this yields fairly large confidence intervals, but the comparison does offer useful insights.

Figure 3.6 compares the (95%) confidence intervals on recall for each participating system using only the families that were selected for adjudication by the simple random sample; Figure 3.7 shows a similar comparison using all of the adjudicated families. From Figure 3.6 we can observe that judgments from assessors yield somewhat higher recall estimates than the judgments from TA, but Figure 3.7 shows the opposite effect. The difference results from some combination of sampling error, appeals that disproportionately benefit participating systems, or systematic biases in the families on which assessors disagree. As the size of the error bars illustrates, we cannot reject sampling error as an explanation. Nonetheless, there is some evidence to support the hypothesis that appeals disproportionately benefit participating systems.

Table 3.5 shows how the overturn rate varies with the reason for adjudication and

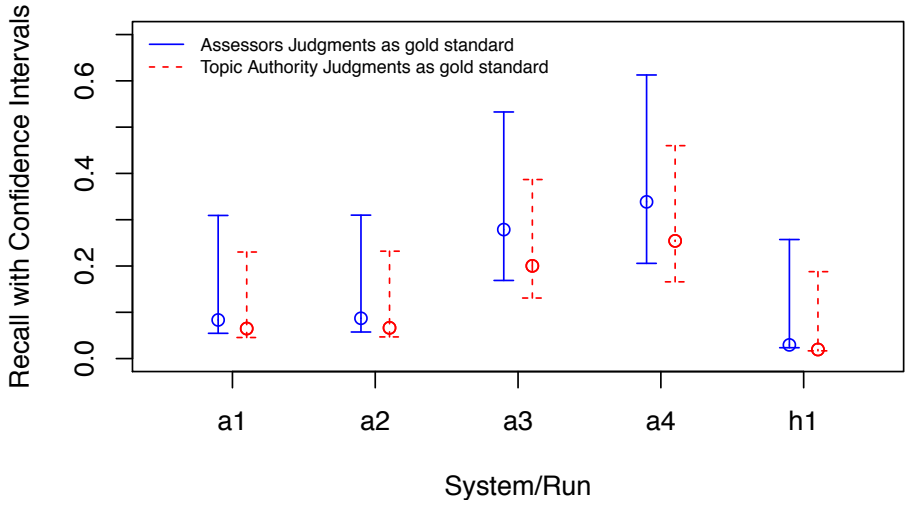

Figure 3.6: Recall, $a4$ ablated, random adjudication

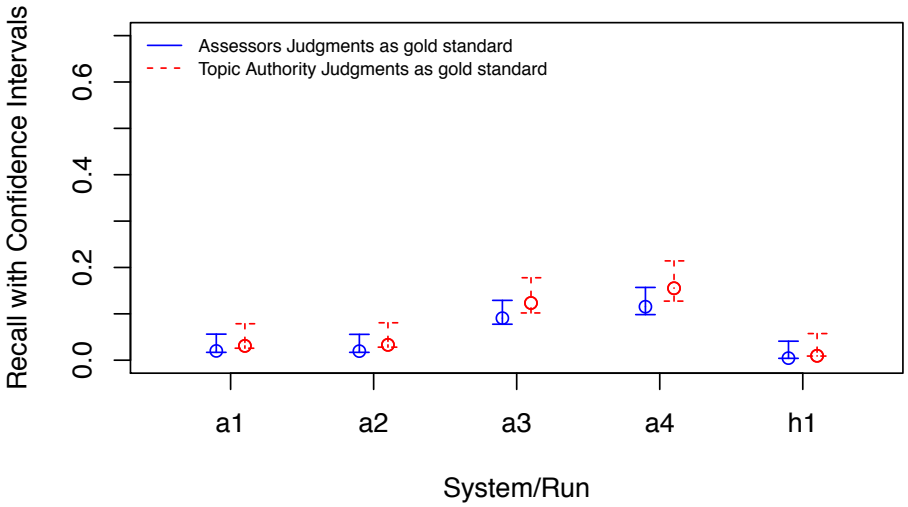

Figure 3.7: Recall, $a4$ ablated, all adjudication

with the original judgment. As the random sampling results show, assessors are more likely to mistakenly judge a family as privileged than as not privileged. Specifically, a z-ratio test for independent proportions finds the $1 \rightarrow 0$ overturn to be significantly more likely than a $0 \rightarrow 1$ overturn ($p < 0.05$). The same is not true for documents appealed by participating teams, however, where the overturn rates in each direction are statistically indistinguishable. Said another way, the increase in total overturn rate from 23% to 36% between randomly sampled adjudications and appealed adjudications (a 58% relative increase) can be largely explained by participating teams being no better than chance at recognizing an assessor's false positive judgments, but by being much better than chance

at recognizing an assessor's false negative judgments.

The implications of this for the reliability of the test collection are clear: estimating absolute measures, and particularly absolute estimates of recall, using assessor judgments that exhibit systematic errors results in estimates that are open to question. If uncorrected assessor judgments were a small fraction of the total judgments, this would be a relatively minor concern, but uncorrected judgments are being used for about 92% of the sampled families. On the positive side, the availability of adjudicated random samples offers the potential for modeling differential error rates conditioned on the first-tier assessor's judgment. On the negative side, the inability to associate judgments with individual assessors in TREC 2010 means that such corrections can only be applied on an aggregate basis. We note, however, that relative comparisons between participating systems can still be informative, so long as assessor errors penalize all participating systems similarly.

Analysis of Sampling Error

To assess reusability, we need to assess the comparability of evaluation results for systems that did and did not contribute to the development of the test collection. A standard way of performing such analyses is through system ablation [89]: removing a system that in fact did participate in the stratification and then rescoring all systems, including the ablated system, and observing the effect on system comparisons. With pooling, ablation results in removing judgments for documents that were uniquely found by one system. With stratified sampling, by contrast, ablation results in re-stratification. For example, when system $a4$ (the participating system with the highest recall) is ablated, the 00000 stratum and the 00010 stratum become merged into a 000?0 stratum (where ? indicates a don't-care condition), the 11001 stratum gets merged with the 11011 stratum to form a 110?1 stratum, and similarly for each other stratum pair that is differentiated only by the ablated system. If we then reapply the process for deciding on the number of families to sample from each merged stratum, we will see little effect on the sampling rate for most strata. The one important exception is the 000?0 stratum (continuing with our example of ablating system $a4$), where we are merging large strata with quite different sampling rates (very small strata can also see substantial changes in their sampling rate, but their effect on the overall estimate will be small). We therefore model the effect of

Table 3.5: Overturn rates

	Assessor → Topic Authority	
Adjudication Basis	0 → 1	1 → 0
Random sample	31 of 161 (19%)	20 of 62 (32%)
Team appeal	32 of 77 (42%)	54 of 160 (34%)
Disagreement	28 of 49 (57%)	9 of 27 (33%)

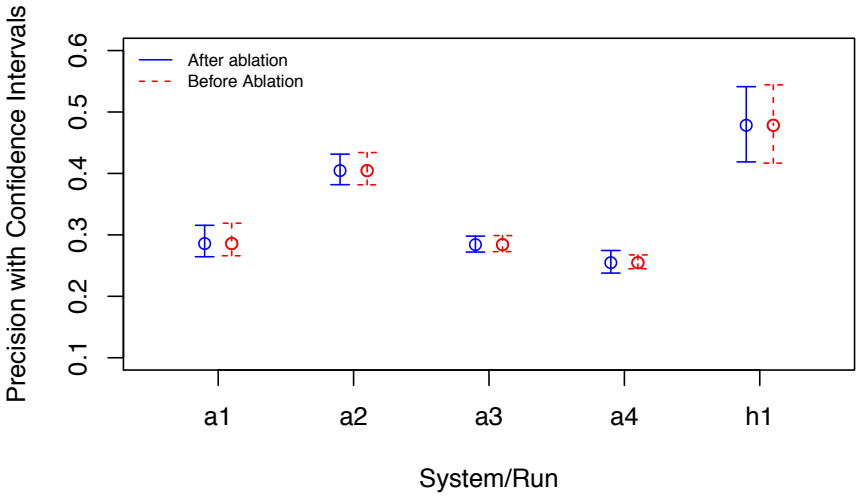

Figure 3.8: Precision, $a4$ ablated, all adjudication

ablation by allocating all of the samples in each pair of strata to the corresponding merged stratum, adjusting the contributions of each sample to the estimate of the yield for the merged stratum to be equal.

To generalize, let a refer to the stratum in the pair including families classified as privileged by the ablated run, b to the corresponding stratum containing families classified as not privileged by the ablated run, and c to the merged stratum. We assume that the merged stratum would include the same number of samples that the two original strata contained separately; that is $n_c = n_a + n_b$ and the sampling rate for merged stratum c is $p_c = n_c/N_c$, where $N_c = N_a + N_b$.

We performed three ablation experiments, in each case ablating one system with high, medium or low recall and then recalculating point estimates and confidence intervals for every system. Comparing post-ablation to pre-ablation results, we see that point estimates are unchanged, as expected, but as Figure 3.8 shows confidence intervals for precision increase for the ablated system (system $a4$ in this figure). We attribute this to

the reductions in the sampling rate for the 00010 stratum (from merging with the 00000 stratum, which results in documents in the former 00010 stratum being sampled at a far lower rate), since we expect families classified uniquely by any reasonable system as privileged to more often actually be privileged than families that no system classified as privileged. The same pattern is evident in our other two ablation experiments (ablating systems $a2$ or $h1$; not shown). No similar effect was observed for confidence intervals on recall, however, perhaps because the estimates for the retrieved set contribute to both the numerator and the denominator of the recall computation.

3.4.2 Expertise and Sample Bias in Classifier Results

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

Effect of Annotator Expertise

We study the effect of annotator expertise on training by using the adjudicated families for training (families in set $AS - TA$ 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 3.9 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. 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 3.10 shows the intersecting sets of all the classifiers' predictions on the privileged families in the test-set. By analyzing a pair of intersecting

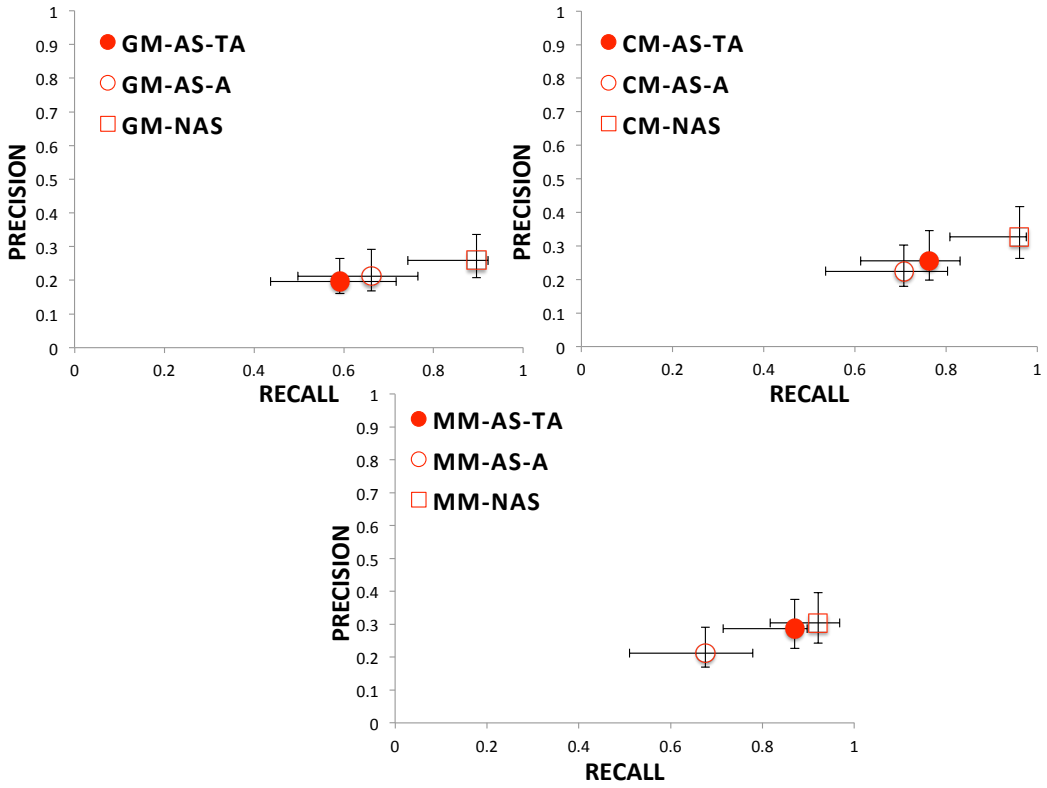

Figure 3.9: Effect of Annotator Expertise on Training

sets; (1) CM-AS-TA and MM-AS-TA (total of (22+0-1[11]) families), and the sets CM-AS-A and MM-AS-A (total of (15+4-0) families) (2) GM-AS-TA and MM-AS-TA (total of (7+7-1) families), and the sets GM-AS-A and MM-AS-A (total of (2+13-0) families), we conclude that the performance of MM-AS-TA model gains a significant increase in recall over MM-AS-A.

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 3.9 shows that, automated classifiers trained on the unbiased annotations from cheaper non-expert sources (Families in NAS) derive the best results. An 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

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

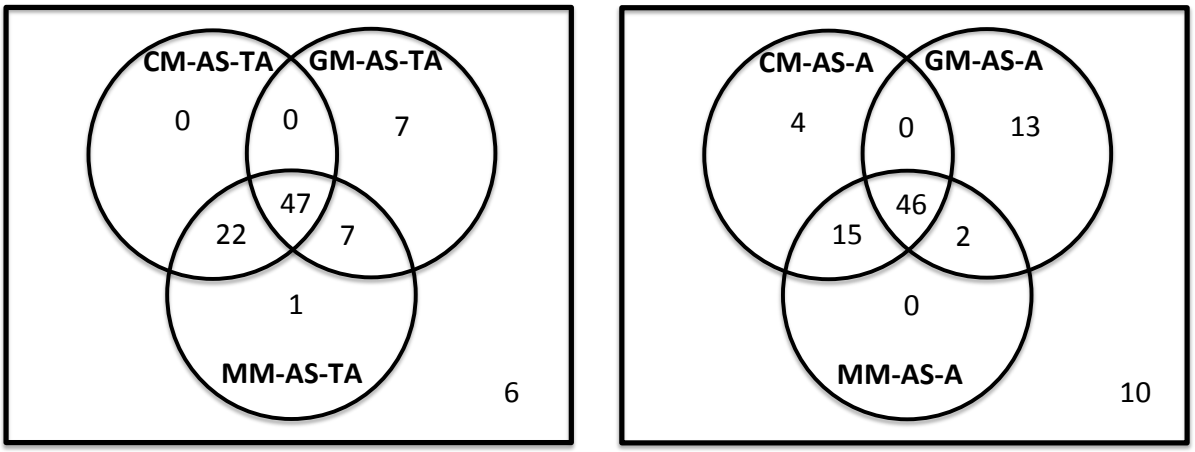

Figure 3.10: Analysis of Classifier Privilege Predictions

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.

3.5 Summary

In this chapter, we have explored set-based evaluation for privilege classification using stratified sampling, with strata defined by the overlapping classification results from different participating systems. We have studied collection reliability by examining the impact of unmodeled assessor errors on evaluation results, and collection reusability by showing that confidence intervals are affected when we reconstruct the test collection in a way that does not rely on the contributions of one participating system. We show that assessor errors do adversely affect absolute estimates of recall.

To study effect of training data and classifier accuracy, we utilize the privilege judg-

ments 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. 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 generally yields the best results. As future work, we motivate 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.

Chapter 4: Manual Review

Manual review denotes a process in which every document that is marked for production is reviewed for relevance (responsiveness and/or privilege) by at least one human reviewer. Exhaustive manual review involves having a human reviewer examine every document in a collection and code each document as relevant or non-relevant, and perhaps apply additional labels such as "privileged" or not, "hot document" or not, and sometimes, specific issue tags. It is not uncommon to have human reviewers exhaustively annotate documents during privilege review phase. Manual review is often accompanied by some sort of quality control process in which a portion of the documents is re-reviewed and, where indicated, re-coded by a second, more authoritative reviewer or a senior attorney. When the coding decisions disagree, action may be taken to diagnose and mitigate the cause. However, the vast majority of documents in the collection are reviewed only once, and the original reviewer's coding is the sole determinant of the disposition of the document.

Automated review denotes a situation in which the decision to produce or not to produce some proportion of the documents is made algorithmically, without a linear manual review. The term "technology-assisted review" is often used instead. In this chapter we introduce what we call technology-assisted manual review to utilize automation during the manual review process.[1]

Lawyers have shown interest to adopt predictive coding technique for finding relevant evidence. As the stakes involved in inadvertent disclosure of privileged content are high, it is natural to doubt any fully computerized technique to accurately recognize content

[1]The work discussed in this chapter is published in CHIIR and ASIST conferences and was done in collaboration with Douglas W. Oard and Amittai Axelrod; An AID for Avoiding Inadvertent Disclosure: Supporting Interactive Review for Privilege in E-Discovery [73] and Finding the Privileged Few: Supporting Privilege Review for E-Discovery [72].

that can properly be withheld. Hence, attorneys are reluctant to trust fully automated techniques for privilege review.[2] This chapter describes the design of an interactive system to support privilege review in which the goals are to improve the speed and accuracy of privilege review.

4.1 Problem Design

Our work in this chapter is focused on providing useful tips to human reviewers during the privilege review process in e-discovery. Several types of privilege might be asserted, but in this chapter we focus principally on attorney-client privilege.[3] Our basic approach to supporting privilege review is to train features or model annotators[4] to label specific components of a message with information that we expect might help a reviewer to make a correct decision. We use a total of five annotators to enrich three types of components: people (or, more specifically, the email addresses for senders and recipients of a message), terms (words found in the message or in attachments to the message), and the date (on which the message was sent). In each case, we compute a numerical score for which higher values indicate a greater likelihood of privilege [72]; for people we also annotate job responsibilities (when known) or organization type (when known, if the job responsibilities are not known).

We study the usefulness of different types of features to human reviewers using a within-subjects user study in which six lawyers each reviewed two sets of documents (email messages, together with their attachments), one set using a baseline system with no annotations, and the second set using our AID system (named for our goal of Avoiding Inadvertent Disclosures) in which annotations were shown for people, terms, and dates. Quantitative measures of review accuracy (e.g., precision and recall) and of review speed are augmented with analysis of self-reported response to questionnaires and interviews. We seek to answer the three research questions (*RQ4a, RQ4b and RQ4c*):

[2] So long as the scale of the privilege review (i.e., the number of relevant documents) is not so great as to preclude manual review.

[3] The rationale behind attorney-client privilege is that justice will be best served when attorneys can communicate freely with their clients (e.g., on matters of fact, intent, or legal strategy), and open communication can be fostered by prospectively protecting such communication from disclosure.

[4] We use the word "annotator" here to refer to an automated system that generates the features.

- Do the accuracy of the user's privilege review judgments improve when system-generated annotations are presented during privilege review?
- Does the user's review speed improve when system-generated annotations are presented during privilege review?
- Which system-generated annotations do users believe are most helpful?

Our results indicate that recall can be enhanced by displaying annotations. Although the improvements in recall come at some cost in precision, given the nature of this application, that cost may be acceptable. Participants in the study principally attribute the beneficial effects to annotations of people (rather than of terms or of dates). These formative evaluation results have implications for annotator and interface design.

4.1.1 Privilege Features

Privilege in legal context is a right given to the parties in a lawsuit to provide protection against the involuntary disclosure of information. Attorney-client privilege in particular exists to protect the information exchange between "privileged persons" for the purpose of obtaining legal advice. Privileged persons include [33]: (1) the client (an individual or an organization), (2) the client's attorney, (3) communicating representatives of either the client or the attorney, and (4) other representatives of the attorney who may assist the attorney in providing legal advice to the client. However, privilege does not arise simply because privileged persons communicate; it can only be claimed when the content of the communication merits the claim.

Our intuition is that, an email message sent or received by a person (e.g. Person3) has a higher probability of being involved in privileged communication if that person frequently communicates with other people (Person5, Person6, etc.) who themselves have a higher probability of being involved in a privileged communication. Figure 4.1 illustrates this idea. As shown in the figure, the node Person3 in the example email network has multiple privileged (P) email exchanges with the node Person5 which in-turn has privileged email exchange with Person6. The privilege propensity of node Person3 depends not only on the emails sent/received by Person3, but also on the email traffic of all the nodes

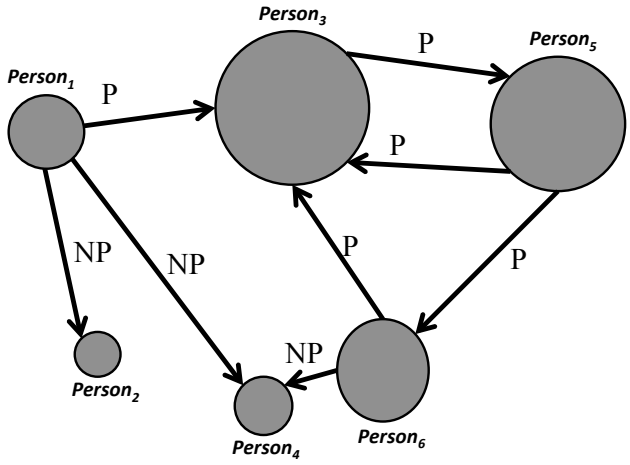

Figure 4.1: Our depiction of Privileged Communication Network
†P ⇒ Privileged, NP ⇒ Not-Privileged

Person3 communicates with. Thus we define "propensity" as a measure of the degree to which we expect a person to engage in privileged communication. It is a number between 0 (low propensity) and 1 (high propensity).

While there has also been some work on the design and evaluation of automated classifiers to actually perform the privilege review task [29, 35, 74], there is a widely held belief among attorneys that reliance on a fully automated classifier for privilege review would incur an undesirable level of uncharacterized risk. Thus automated classifiers are more often used for consistency checking on the results of a manual privilege review process than as the principal basis for that review. In this paper, we explore a second possible use of the technology. That is, use of automated annotations to (hopefully) improve the accuracy or the cost of a manual review process.

4.1.2 Document Collection

For our study, we need a set of documents that we know to be relevant to some request that we might typically see in e-discovery. To train our annotators, we also need a set of similar documents that we know to be privileged. We thus need a test collection that contains some relevance and some privilege judgments. One such collection, which we used in this chapter, was produced during the TREC Legal Track in 2010.

In the 2010 TREC Legal Track's "Interactive task",[5] one task (Topic 303) was to find *"all documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto."* [29] The collection to be searched was version 2 of the EDRM Enron Email Collection, which includes both messages and attachments. The items to be retrieved were "document families," where (following typical practice in e-discovery) a family was defined as an email message together with all of its attachments. Five teams contributed a total of six interactive runs for Topic 303, with each run being a binary assignment of all families as relevant or not relevant. A stratified sample of families was drawn from submitted runs, and 1,090 of those families were judged to be relevant [29]. We have drawn a random sample of 200 of those relevant families for use in our study. Our automated annotation pipeline failed on 12 of those 200 families which lacked a critical field (From, To, or Date), so we removed those 12 families from consideration and randomly split the remaining families into two disjoint sets of 94 families each, which we refer to as D_1 and D_2. We consistently use set D_2 with our Baseline system and set D_1 with the treatment[6] system.

In the 2010 TREC Legal Track's Interactive task, a second task (called "Topic 304") was to find *"all documents or communications that are subject to a claim of attorney-client privilege, work-product, or any other applicable privilege or protection"* [29]. Two teams submitted a total of five runs, with each run being a binary assignment of every family as Privileged or Not Privileged. A stratified sample of 6,736 families were marked as privileged or not privileged by experienced reviewers,[7] and prior work has shown that these annotations can be used to train a privilege classifier with reasonable levels of accuracy [74]. A total of seven families from this random sample were, by chance, also present in either

[5] A task in which participants design both a system and an interactive process for using that system

[6] The treatment system uses an interface that have system-generated features highlighted during review.

[7] 13 of these 6,736 had actually been marked as Unjudged, but during our experiments those 13 were treated as Not Privileged. The effect of this is negligible.

Table 4.1: TREC 2010 privilege judgments (For training and review)

	Training	D_1	D_2
Privileged	932	2	3
Not Privileged	5,799	1	1

D_1 or D_2, and we removed the five that had been judged as Privileged from the set that we used for training our numerical annotators.[8] As Table 4.1 indicates, this resulted in a total of 932 families annotated as Privileged and 5,799 families annotated as Not Privileged that could be used for training our automated annotators. Most of the judgments are from junior annotators. We refer to this set as NAS as described in Chapter 3. The smaller set of training documents are from AS-TA. We first study which of the two training set of judgments give the best coverage.

4.2 The AID System

Our web interface system which we name "AID" (which starts for Avoiding Inadvertent Disclosure) system is a research prototype that is designed to help explore the design space for providing automated assistance to users during privilege review. In this section, we first describe the design of the five types of automated annotators that we have built. We next explain the user interface and interaction design of our AID system.

4.2.1 Propensity Annotation

We define propensity of a person to engage in privileged communication as a number between 0 (indicating low propensity) and 1 (indicating high propensity). We utilize the expert or non-expert labels to indicate whether the family (represented as an edge) connecting the persons in the multi-graph is privileged or not-privileged. Given the labels of the edges, the task is to assign a score to the nodes that depends on the edge labels.

We start by computing a privilege weight value that is associated with each edge in the graph as a prior using the network information from labeled families in the train-set. We then use the idea behind the weighted PageRank technique to score the propensity for

[8] Because of presentation order neither of those Not Privileged documents was seen by any participant in the user study that we describe in this paper.

Algorithm 1 Missing Person Score Algorithm

Input:
$Graph_{test}$
$PRdictionary_{score}(A_{train})$
$uniqueNodes_{test}$

procedure $GetMissingNodeScores_{test}$
  $rankDictionary \leftarrow sort(PRdictionary_{score}(A_{train}))$
  $uniqueNodeScoreDict \leftarrow NULL$
  for <each edge(s, r) in $Graph_{test}$ > **do**
    if s in $uniqueNodes_{test}$ **then**
      $sum \leftarrow zero$
      if **then** r in rankDictionary.keys()
        $sum \leftarrow sum + d[r]$
      else
        $sum \leftarrow sum + [(min(d.values()) + max(d.values())) \div length(d)]$
      end if
    end if
    $score_n \leftarrow sum \div num(recipients)$
    $unqNodeScoreDict[n] \leftarrow score_n$
  end for
  return $unqNodeScoreDict[n]$
end procedure

Figure 4.2: Missing Person Score Algorithm

each person [84]. We define $w[Edge(x, y)]$ as the edge weight between x and y given the label of each communication edge as:

$$w[Edge(x,y)] = \sum_{e \in E_{train}} \frac{n(x,y)_{e_p}}{(n(x,y)_{e_p} + n(x,y)_{e_{np}})} \tag{4.1}$$

where E_{train} is a set of labeled edges used in training set; $n(x, y)_{e_p}$ and $n(x, y)_{e_{np}}$ is the number of edges labeled as privileged and not-privileged respectively, with x as sender and y as the recipient. The weight of $Edge(x, y)$ indicates the privilege probability between the two people. To score the individual nodes, we use these weighted edges in the graph as an input to a power iteration algorithm to obtain the "propensity score" or PR_{score} for each person using:

$$PR_{score}(x) = (1-d) + d \sum_{v \in E_x} \frac{PR_{score}(v)}{N_v} \tag{4.2}$$

where $d = 0.85$[9] is the dampening factor; E_x is the set of edges where x is the recipient; and N_v is the total number of edges where v is the sender.

Given the PR_{score} of each person seen in the labeled training set families, the final step of our person scoring algorithm is to calculate the PR_{score} of each person seen in the test set. Only 32% of the senders or recipients of unannotated emails have a PR_{score} greater than zero when trained on labeled training set NAS and 30% when trained on the labeled training set AS-TA. The other 56% are not present even once in either training sets. To estimate propensity for people who are not present in the training set, we leverage each unknown persons' egocentric communication network, ultimately increasing the number of people to whom we can assign a propensity score to 94% of senders and recipients in the test set when trained on documents from NAS (93% when trained on documents from AS-TA). Figure 4.3 shows an example family where none of the 6 persons are seen in the training set. However, our missing person algorithm scores 3 of the 6 (shown in bold font). To calculate the propensity score for each person in the test set, our algorithm follows two steps:

Common Person Scoring: We obtain a set of common persons (persons seen in both train and test set) $Common_a$. For each person i in the test set, if $i \in Common_a$ then we use the $PR_{score}(i)$.

Missing Person Scoring: For each person i in the test set, if $i \notin Common_a$ we take the approach described in Algorithm 4.2. For each person in the test graph who is not seen in the train graph, we exploit the the person's network information. If the missing sender is connected to one or more recipients who are seen in the train graph, we assign the average of recipient's node scores as the missing sender score. However, if the sending person is connected to only missing recipients, we assign the sender the average of all PR_{score} values in the train graph. We take this conservative approach to scoring missing persons as we do not want to mislead the annotator by providing a zero propensity score when we are actually simply unsure about the propensity of a person.

[9] We fix the value d to 0.85 to assign 15% privilege likelihood for persons with no prior labeled privileged communication.

Figure 4.3: Privileged Email

4.2.2 Person Role Annotation

Propensity annotation is intended to help call a user's attention to a specific person, but actually knowing how to interpret the importance of that person requires additional information. Professional reviewers would typically have information about the roles of specific people (e.g., they might know who the attorneys and the senior executives are), and in complex cases such lists could be quite extensive. The speed, and perhaps the accuracy, of the review process might be enhanced if we could embed that information in the review system. For this purpose, we need a role annotator that can associate each email address with some (generic or specific) version of their job tile. For our experiments we therefore built a simple role annotator using table lookup. We manually populated this table for 160 of the 1,611 unique email addresses that appear in at least one of the 188 families in either of our two test-sets. We obtained these roles from the MySQL database released by Shetty and Abidi [66], from ground truth produced for evaluating the Author-Recipient-Topic model of McCallum et al [50], from other lists found on the Web,[10] from manual examination of automatically inserted signature blocks in email messages throughout the collection, from public profiles such as LinkedIn, and through manual Web searches. The

[10] http://cis.jhu.edu/~parky/Enron/employees, http://www.desdemonadespair.net/2010/09/bushenron-chronology.html

roles were manually edited for consistency and conciseness.

4.2.3 Organization Type Annotation

When the role of a specific person is not known, reviewers might benefit from knowing the type of the organization for which that person works. We therefore used the same lookup table to annotate the organization in such cases. We did this by manually examining the domain name of an email address and then using a current domain name registry, a Web search, or our personal knowledge to label the organization's type, when possible. For example, some messages in the Enron collection are from addresses with the domain 'brobeck.com', and Wikipedia indicates that (at the time) Brobeck, Phleger & Harrison was a law firm.

4.2.4 Content Analysis

Term unigrams have been reported to be a useful feature set for privilege classification [71], so it is natural to also consider annotating terms. The families in our collection contain many more terms than email addresses. Hence some approach to feature selection is needed if we are to avoid the display clutter that would result from annotating every term. We perform this feature selection by obtaining the entropy difference for each term. The entropy difference score identifies words that are like words in the Privileged set and also unlike words in the Not Privileged set [53]. To do this, we first tokenize the email message subject field, email message body and extracted text from each attachment for each family in the training set and in the test-set. We then build two unigram language models on these terms (i.e., the unstemmed tokens), one for the 932 families in the training set that were labeled as Privileged, and the other for the 5,799 families in the training set that were labeled as Not Privileged. We then rank each term w present in either of the test-set families using the entropy difference:

$$score(w) = H_p(w) - H_{np}(w)$$

Figure 4.4: Indicative terms

where $H_p(w)$ and $H_{np}(w)$ respectively represent the entropy of the token w in the Privileged and the Not Privileged language models [11]. Negative Entropy difference scores indicate terms that are indicative of privilege. Figure 4.4 shows the Indicative terms where larger the font size; higher the negative Entropy Difference. We used the top 10% unique terms with a high negative entropy difference value. Out the the top 350 terms, we annotate 117 terms with the highest negative entropy difference as strongly indicative of privilege, the middle set of 117 terms as moderately indicative of privilege, and the remaining 116 terms as somewhat indicative of privilege.

4.2.5 Temporal Likelihood

Email communications that focus on the lawsuits often occur during specific time intervals, so it seems reasonable to expect that privileged communication regarding those events might exhibit some predictable temporal variation. We therefore also built an annotator for dates that estimates the likelihood of privileged communication on (or near)

that date. To do that, we parse the date field of the email that heads each family in the training set. We then use maximum likelihood estimation with Laplace smoothing to estimate the probability that a family sampled from the set of training families sent on a specific date would be privileged. We calculate that probability estimate as:

$$P(d_i|n_d^x) = \frac{n_{d_i}^p + 1}{n_{d_i}^p + n_{d_i}^{np} + 2} \tag{4.3}$$

where d_i is the date of the message, $n_{d_i}^p$ and $n_{d_i}^{np}$ are the total number of Privileged and Not-Privileged families sent on d_i respectively. Because TREC performed stratified sampling, designed to oversample potentially privileged families, we expect this to be a substantial overestimate of the actual probability. Nonetheless, we would expect relative values of the estimate to be informative.

4.2.6 User Interface

Our research prototype is designed to help explore the design space for providing automated assistance to users during privilege review. We use the design of the five types of automated annotators that we have built. We then explain the interface and interaction design of our review system. We characterize the coverage of each of our automated annotators as the fraction of the unique items (people, terms or dates) in the 61 families for which annotations are available.

Figure 4.5 shows a screenshot for our AID system. Documents are presented to every user in the same order, and the user must record a judgment (Privileged, Not Privileged, or No Decision) before being shown the next document. They could return to any previously judged document to change their judgment if they chose to do so. Annotations are provided as visual scaffolds during the privilege review process. Whenever a person role or organization type annotation is available, the associated email address is displayed with a red background, and the role or type annotation can be displayed in a manner similar to a "tool tip" (using a graphical control element that is activated when the user hovers the mouse over the shaded area). We shade the background with variations of the color red to indicate the propensity category (darker red for strong propensity,

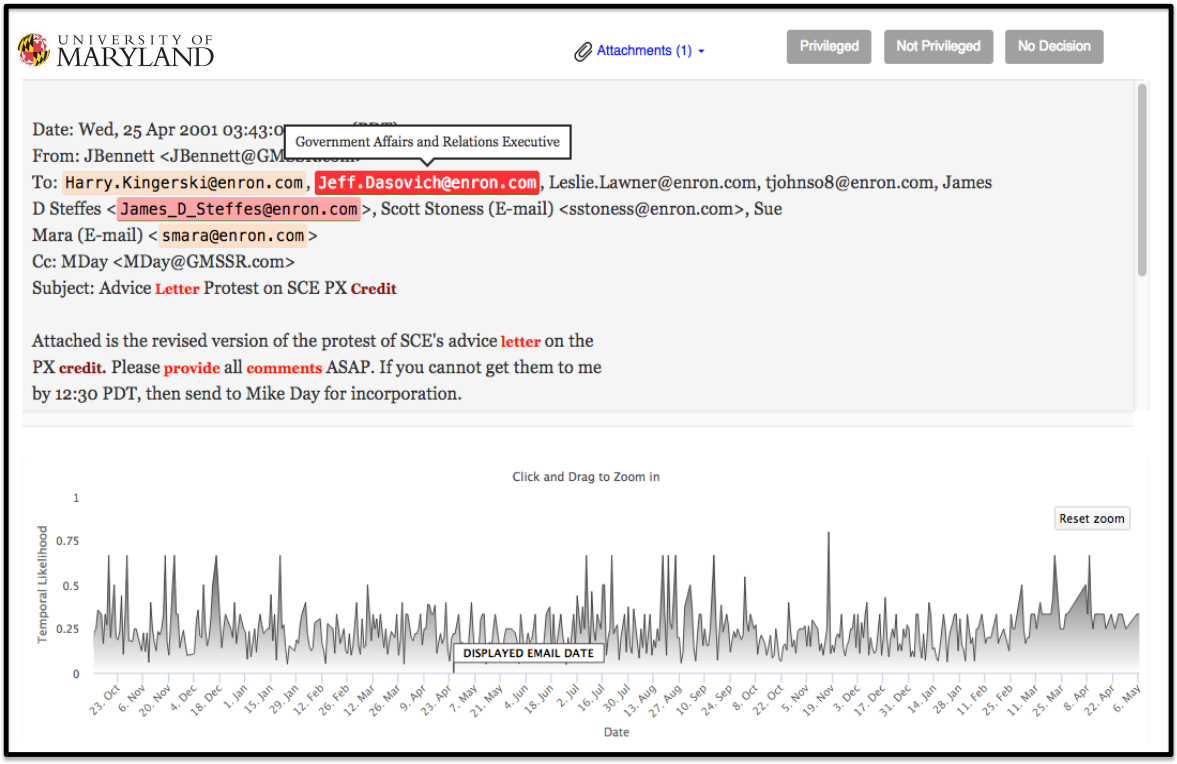

Figure 4.5: The AID system.

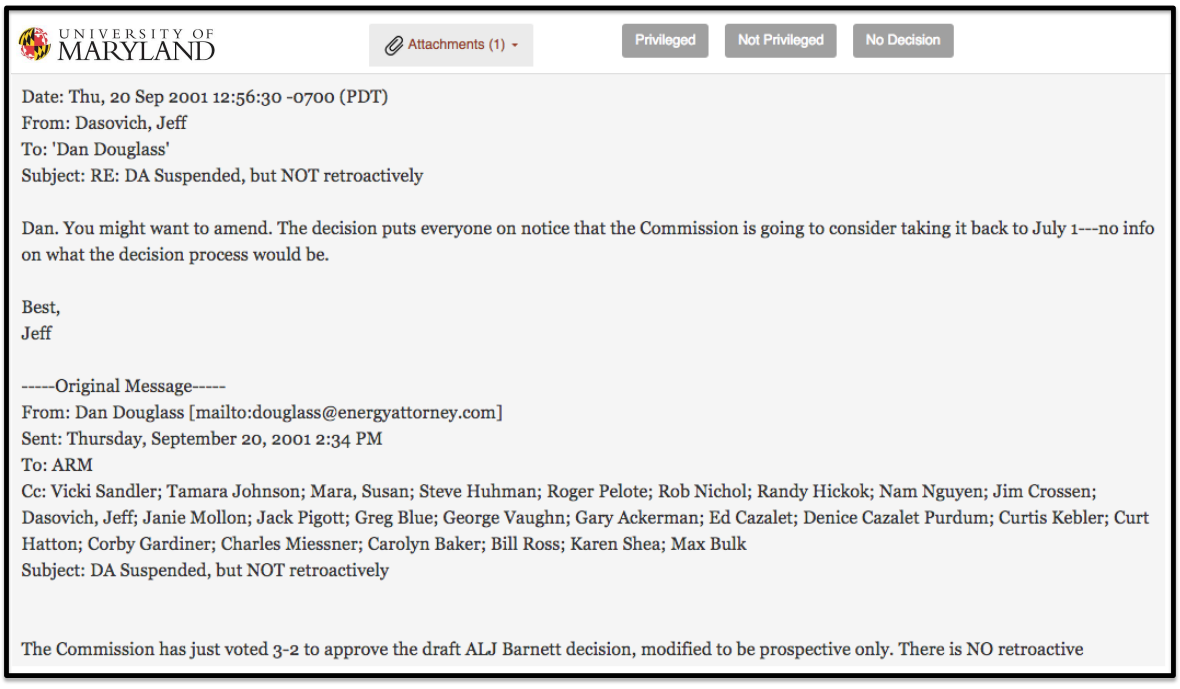

Figure 4.6: The Baseline system.

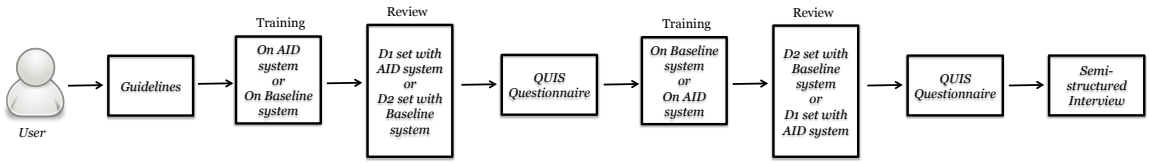

Figure 4.7: User study procedure.

lighter red for moderate propensity, very light red for all other cases in which role or type information is available).[11] On average (across the 61 viewed families), 58% of the email addresses appearing as senders or recipients had a role or a type annotation available (55% for person role, 3% for organization type). About two-thirds of those cases in which role or type annotation were available, were displayed with shading indicating strong or moderate propensity.

The display of terms that are indicative of privilege in the subject line, email message body, or attachments follows a similar pattern, but by altering the color of the typeface rather than the background. For example, the term "credit" is rendered in the darkest shade of red[12] in Figure 4.5, thus indicating it was strongly indicative of privilege. On average (across the 61 viewed families), 2% of all term occurrences are highlighted.

Temporal likelihood is plotted as a connected line plot, with date as the horizontal axis and temporal likelihood as the vertical axis. This has the effect of visually performing linear interpolation of temporal likelihood for dates on which that likelihood can not be computed directly. The displayed date range can be reduced (by a click and drag zoom-in functionality) by the user for finer-grained display.

Figure 4.6 shows the user interface of our Baseline system. As can be seen, the only differences from the AID system are that none of the annotations are present, and the omission of the temporal likelihood plot permits more of the content to be displayed. Both the systems log the time, family ID, user ID and judgment (Privileged, Not Privileged, or No Decision) for each reviewed family.

The principal goal of our user study was to determine whether any of our system-

[11] Low propensity addresses for which no role or organization type information is available have no background shading.

[12] We chose to use the same color gradations for terms and email addresses to simplify training, but the question of optimal color choices merits further investigation.

generated annotators could help the users to perform the review task more quickly, more accurately, or both. A secondary goal was to determine whether there were usability issues with our current interface design that might adversely affect our ability to determine the effects of specific annotators. A third goal was to use our current AID system design as an artifact around which we could discuss specific as-yet unimplemented capabilities that experts might believe would provide useful support for the task.

4.2.7 Study Participants and Procedure

We were able to recruit a total of six participants from the first two groups, which we judged to be adequate for the comparisons we wished to make, so we limited our study to those six participants. Two of the six were senior attorneys employed by law firms with a current e-discovery practice. These senior attorneys are experienced litigators who have extensive experience conducting both relevance and privilege review for email using commercial Technology Assisted Review (TAR) tools.[13] We refer to these senior attorneys as S_1 and S_2.

The remaining four participants were law school graduates. Two of the four had prior experience conducting relevance and privilege review using commercial TAR tools, but neither was currently working in an e-discovery practice; one of the two is a graduate student in another discipline, the other is an intellectual property attorney. We refer to this pair of experienced reviewers as E_1 and E_2. By coincidence, E_2 had experience working as a reviewer during the original Enron litigation.

The remaining two participants had experience conducting e-discovery reviews some time ago, principally on paper, but neither had experience using current TAR tools. One was a retired attorney, the other was currently a faculty member in another discipline. We refer to these (TAR) inexperienced reviewers as I_1 and I_2. I_2 had little direct experience using computers.

Figure 4.7 summarizes the study procedure for one of the six single-participant sessions.[14] Each participant completed the study in about two hours, with a 10 minute

[13] Tools like Recommind, Nuix, kCura, etc.

[14] Refer to Appendix A for details about the IRB approval.

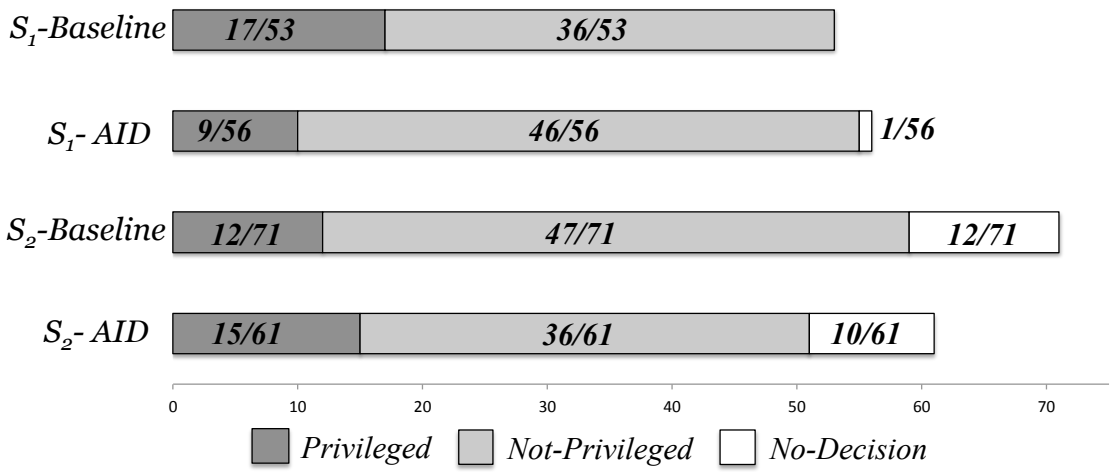

Figure 4.8: S_1 and S_2 Judgments by type

Table 4.2: Contingency table; for review of same families by S_1 & S_2)

	S_1: Privileged	S_1Not Privileged	S_1: No Decision	S_1: Not Seen
S_2: Privileged	15	7	0	5
S_2: Not Privileged	5	62	0	16
S_2: No Decision	6	12	1	3
S_2: Not Seen	0	1	0	75

†There was one family that was skipped in sequence by chance by S_2; but was not skipped by S_1.

break at the end of the first hour. Participants were given an overview of the review task and were asked to read a written description of the study that we provided before signing a consent form. Each participant then received a 5 minute tutorial on the first system they would use, presented by the investigator, in which the different parts of the system were demonstrated using a few example families.

4.3 Results

In this section we first focus on quantitative results for accuracy and speed. Following that we contextualize these results from qualitative results from our interview and from our usability questionnaire. We then draw insights from each of these analyses to discuss what we see as the most important conclusions that can be drawn from this study.

4.3.1 Selecting a Benchmark for Evaluation

If we are to make any useful statements about the accuracy of a privilege review, we must first select an informative set of judgments as benchmark against which accuracy can be measured. This benchmark judgments need not be perfect for the resulting measures to be informative, but we will have the greatest confidence in our results if we select the best available benchmark judgments. Thus it is natural to begin by characterizing the results from the two senior attorneys, since we would expect their judgments to be natural candidates as a benchmark.

Figure 4.8 shows the number of judgments of each type made by S_1 and S_2 for each of the two conditions. As can be seen, S_2 is somewhat faster than S_1 (making 33% more judgments in the same 30 minutes in the Baseline condition, and 9% more in the AID condition). S_2 records many more No Decision judgments (22 for S_2 vs. 1 for S_1).[15] As Table 4.2 shows, senior attorney S_1 marked a total 15+5+6=26 families as Privileged while S_2 marked a total of 15+7+5=27 families as Privileged. Among the families seen by both senior attorneys (using either system), 15 families were marked as Privileged by both. Computing chance corrected inter-annotator agreement between S_1 and S_2 using Cohen's Kappa (κ) yields 0.68, a value that Landis and Koch [43] characterize as "substantial." Indeed, given the class prevalence in our test sets, chance agreement would be 0.57, making very high levels of κ difficult to achieve [10].

TREC 2010 Interactive Task Topic 304 privilege judgments are available for seven of the families in our test set. Of those seven, 5 were Privileged and 2 were Not Privileged. Of the 5, three families were adjudicated by the Topic Authority (a senior attorney whose judgments were authoritative) who was responsible for providing guidance and adjudicating disputes. Out of the three Privileged families adjudicated by the TREC Topic Authority, two were reviewed by both S_1 and S_2. S_1 agreed with the Topic Authority on one of the two families by marking one of the two families as Privileged while the other as Not-Privileged. S_2 never agreed with the Topic Authority. S_2 marked one of the two families as Not Privileged (the same family marked as Not Privileged by S_1) and the other

[15] Participants mark a family as No Decision when a clear distinction between Privileged and Not Privileged could not be made on the email message or any of its attachments.

was marked as No Decision. Comparisons on two judgments is not sufficient to determine whether the two senior attorneys in our user study are (1) generally more inclined to judge documents as Not Privileged than the TREC Topic Authority would have been (2) generally inclined to agree with each other, but we can say that there is no evidence to refute such a claim.

From this analysis, either senior attorney could reasonably be chosen as a benchmark against which the other participant's judgments could be measured for accuracy. However, because S_2 left 19 families unjudged and skipped reviewing 1 family throughout the review sequence and all 24 of the families that were not seen by S_1 were late in the review sequence, a larger number of useful judgments are available from S_1. We therefore use judgments from S_1 as a benchmark for evaluation. We evaluate participants on the basis of precision and recall estimates that we report in Figure 4.9.

4.3.2 Accuracy

Figure 4.9 shows the privilege review effectiveness of S_2, E_1, E_2, and I_1 for the Baseline and AID conditions, evaluated as if the judgments by S_1 were the ground truth. We calculate point estimates for precision and recall using only the cases judged as Privileged or Not Privileged by both S_1 and by the participant whose decisions are being evaluated (i.e., we omit No Decision and Not Seen cases from both). Because we are comparing estimates for different sets of documents, we also show the 95% confidence intervals for recall and for precision, computed using the standard approximation method described by Agresti et al. [9]. Results for I_2 are not shown because after removal of the 21 No Decision judgments recorded by I_2 there were only 7 families judged by I_2 (3 in the AID condition, 4 in the Baseline condition), a number insufficient for useful estimation of intervals.[16]

From Figure 4.9 we can conclude that there is a consistent and statistically significant improvement in recall when the review task is performed using our AID system for all four participants (S_2, E_1, E_2, I_1).[17] This improvement is, however, accompanied by

[16] All 7 were judged as Privileged, suggesting that participant I_2 may have intended to record judgments of Not Privileged and instead incorrectly selected No Decision. It was participant I_2 who had only limited personal experience using computers.

[17] We consider a difference to be statistically significant if each point estimate lies outside the 95% confidence interval for the other condition.

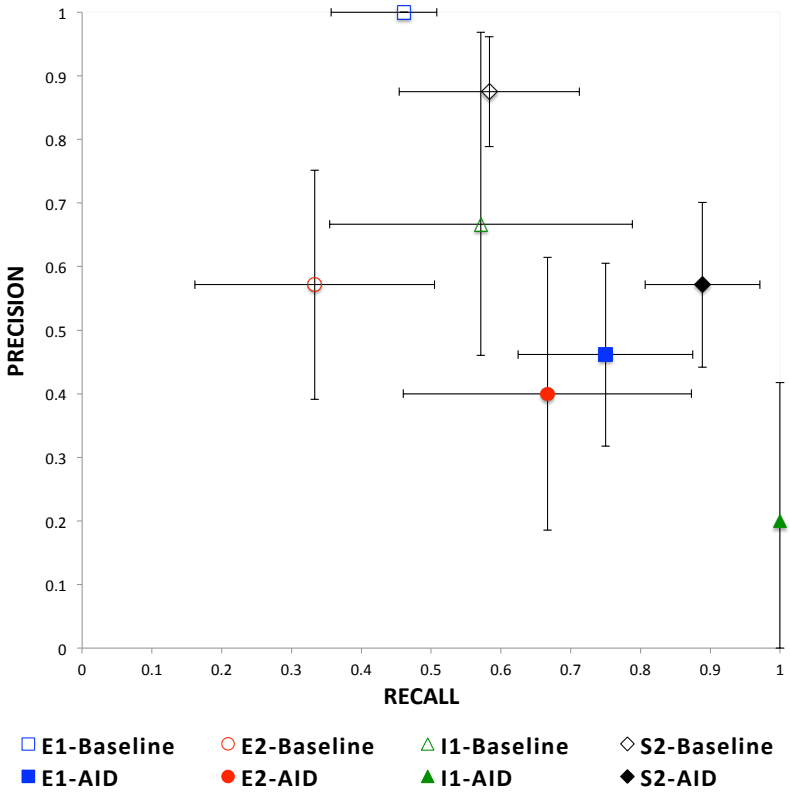

Figure 4.9: Evaluation - S_1 judgments as Benchmark.

a statistically significant reduction in precision for three of the four participants. Instead using S_2 as a reference to evaluate S_1, E_1, E_2, and I_1 (not shown) yields similar results, with statistically significant improvements in recall in 1 of 4 cases and statistically significant decreases in precision in 2 of 4 cases. Since the principal goal of our AID system is to avoid inadvertent disclosures, this consistent bias in favor of recall (i.e., in avoiding false negatives), regardless of which senior attorney we select as a reference, is well in line with that goal.

4.3.3 Speed

To characterize the effect of the choice of system on review speed, we computed the number of families reviewed by each participant in 30 minutes using the Baseline and the AID systems, observing little difference in the means (averaging 40.1 families for the AID condition and 43.6 families for the Baseline condition).[18] A paired t-test found no

[18]Data from participant I_2 is omitted from this analysis.

Table 4.3: QUIS Summary

	S_1 BL	S_1 AID	S_2 BL	S_2 AID	E_1 BL	E_1 AID	E_2 BL	E_2 AID	I_1 BL	I_1 AID	I_2 BL	I_2 AID
Review experience	NF	Good	Bad	Good	Good	Great	Good	Good	Bad	Good	Fair	Good
Info was adequate	D	A	SD	MD	A	SA	A	SA	SD	MD	SD	A
People info was useful		SA		A		A		SA		SA		NF
Term info was useful		SA		NF		MA		NF		NF		SD
Date graph was useful		NF		D		D		MA		D		A
Logical use of colors		A		A		SA		NF		A		A

†SA=Strongly Agree, A=Agree, MA=Moderately Agree, SD=Strongly Disagree, D=Disagree, MD=Moderately Disagree; NF=Neutral Feedback, blank indicates not applicable; BL=Baseline.

detectable difference in average review speed across the two conditions ($p > 0.38$). From these results we conclude that there is no indication that our AID system results in faster review, and indeed it is possible that our AID system might result in marginally slower review.

4.3.4 Usability

Table 4.3 summarizes participant responses to six of the seven QUIS questions (a seventh question, about layout, evoked no useful differences in the responses). Five of the six participants assigned a higher rating to the overall experience with the AID system than with the Baseline system (and the sixth participant noted no difference). All six participants gave more positive scores to the AID system than to the Baseline system in response to the question about adequacy of the displayed information. Person highlighting was reported to be useful (to at least some degree) by five of the six participants, whereas term highlighting and the date graph were each reported to be useful to some degree by only two of the six participants.

4.3.5 Usefulness

During the semi-structured interview session, we asked each participant which type of system-generated annotation they found to be most useful; five of the six named person annotation. The following excerpts are representative of responses that participants gave to our open-ended questions.

"I think having the role or type information in-line on the user interface was very helpful. All I had to do was to hover over the name instead of looking it up on a sheet of paper as we normally do." — S_1

"I would honestly like the people highlighting concept much more if it would give me more information about the meta-data. Having information about the domain addresses of people who are not Enron employees is one such information." — S_2

"The presence of highlighted people made me look into the documents more carefully in non obvious cases for the presence of potentially privilege content. It help me to make a filtering decision about which document need more attention. The highlighting helped me to be quicker." — E_1

"I think the trickiest part was to review the document when the information about a sub-set of the people was missing. For example, if there were 6 people and we have information about 3 of them but not the other 3, it is hard to predict who the other players are." — E_2

"I think the highlighting of the people was useful to do the review; the highlighting of the terms were less useful because almost all emails contain the same boilerplate language and the term highlights did not provide much information; and about the dates, I did not feel the need to use the date information displayed on the graph." — I_1

"The ideas presented in the AID system are good, however the information provided was sometimes confusing to me. The role and type information provided was useful but the term highlighting was distracting; mainly because the highlighted terms did not make sense to determine privilege and I lost my faith on the terms." — I_2

4.4 Summary

Our quantitative results clearly indicate that our AID system resulted in a greater ability to detect privileged documents. The QUIS and our semi-structured interviews provide consistent support to our belief that our annotation of people (or, more specifically, of email addresses) is principally responsible for this improvement. Of the three ways we annotate people (for propensity, for person role, or for organization type) we have the strongest evidence for a claim that role and organization type annotation was believed by our participants to be useful; we do not have sufficient evidence to separately identify the effect of propensity annotation. Neither our present implementations of term highlighting

nor the date graph were often commented on favorably by the participants. From these observations, we conclude that our current AID system achieves its principal objective of helping to avoid inadvertant disclosure, that further study is needed to separately analyze the value of propensity annotation, that the value of term annotation has not yet been shown, and that further refinement of our approach to date annotation will not be among our highest near-term priorities. We base this last conclusion in part on the following comment by S_1, who said *"Date information could be helpful during responsiveness review. But for privilege review, it is less likely to be useful"*.

We were somewhat surprised by the magnitude and consistency of the drop in precision that accompanied the increases in recall that we observed from the use of our AID system. In privilege review, low precision could result in incorrectly withholding some families that should properly have been turned over to the requesting party. Perhaps such cases might be discovered and corrected in a second stage of privilege review, but a two-stage review process would naturally lead to higher costs. Future work aimed at understanding the reason for the reduction in precision will thus be a high priority. Moreover, trade-offs between recall and precision are natural, so it may be that similar results might be obtained in other ways (e.g., by providing financial incentives based on the number of privileged documents found). In future work it will therefore be important to develop task-tuned utility measures that account for the relative importance of recall and precision for the privilege review task and to develop study designs in which recall at comparable levels of precision can be studied.

Our participants made some suggestions for improvements that might be made to our AID system. One useful suggestion was to consider highlighting multi-word expressions that are indicative of privilege, rather than only single-terms as our present system does. Another useful suggestion was to consider augmenting our role annotations with an opportunity to drill down to learn more (e.g., date assigned to that role, previous roles, or supervisory relationships). In future work we are interested in exploring the potential for viewing privilege review as a structured collaboration task, and when we asked about this several of our participants (three of the five who we asked) indicated that system support for collaboration might be of interest for privilege review.

Chapter 5: Predictive Coding With Manual Review

Adoption of predictive coding technique to categorize each document in a collection as privileged or not, and to prioritize the documents based on expected risk before the manual review is the key approach we discuss in this chapter. The party performing the review before production may incur costs of two types, namely, annotation costs (deriving from the fact that human reviewers need to be paid for their work) and misclassification costs (deriving from the fact that failing to correctly determine the responsiveness or privilege of a document may adversely affect the interests of the parties in various ways). Relying exclusively on results from the predictive coding model would minimize manual annotation costs but could result in substantial misclassification costs, while relying exclusively on manual review could generate the opposite consequences. The principal focus of the work presented in this chapter is therefore on developing a semi-automated process. The goal of the semi-automated system is to develop an efficient way of automatically ranking documents based on classifier decisions[1] and partially reviewing those ranked documents manually to minimize the overall cost of the e-discovery process.

Our approach is based on a realistic intuition that automation is imperfect. Thus attorneys will often perform partial or complete manual review depending on the classifier's results. If the manual review of a sample of the classifier's output reveals an unacceptably high error rate, then additional manual review would be needed. Additional training data might yield improved classifier accuracy, but ultimately some limit will be reached beyond which an alternative strategy is needed. If the best error rate that the automatic classifier can achieve remains worse than what human reviewers can achieve, then additional manual

[1]The work discussed in this chapter is currently under review and was done in collaboration with Douglas W. Oard and Fabrizio Sebastiani; Minimizing the Expected Costs of Review for Responsiveness and Privilege in E-Discovery [54].

review can further decrease the overall error rate. This approach works because in e-discovery we are ultimately classifying some finite population of documents and it is thus the accuracy of the classification decisions, and not of the classifier itself, that we care about.

The main contributions of this piece of our work are (1) to be able to quantify a classifier error to a cost value, (2) to derive a cost function as the basis of our evaluation and (3) to determine when and to what extent is it rational to adopt automation in this human-in-the-loop application domain. This chapter answers research questions *RQ5a and RQ5b* introduced in Chapter 1, section 1.2.

5.1 Problem Design

We model our algorithm based on an assumption that all relevant costs can be quantified. These costs are of two types, namely, annotation costs; resulting from the wages paid to human reviewers for their time and work, and misclassification costs. Misclassification costs result from the fact that failing to correctly determine the responsiveness or privilege of a document results in incorrect decisions, which would have consequences that we model as costs. The notion of risk arises naturally in a cost-sensitive classification context, due to the existence of multiple outcomes in probability theory. Depending on the outcome, each outcome has its own cost (e.g., incurring a sanction for having entered on the privilege log a document that should instead have been produced). Minimizing this risk requires avoiding certain outcomes for which a combination of probability of occurrence and cost is high. Here, the notion of "risk" $R(d)$ is the converse of the notion of *utility*; one usually speaks of "risk" when each of the possible events has an associated cost (i.e., the amount at risk due to an undesired consequence), whereas one usually speaks of "utility" when each possible event has an associated gain (i.e., a desired consequence). Anyway, the two notions are interchangeable; we prefer speaking of "risk" here since the entire process involves costs, and not gains, for the producing party, and it is the expectation over these costs that we want to minimize.

Thus, we formalize the problem of the e-discovery process as a risk minimization

framework (called MINECORE, for "MINimizing the Expected COsts of REview") that seeks to strike an optimal balance between the annotation and the misclassification costs.

MINECORE is defined as a semi-automated system whose goal is to identify documents that need to be produced (responsive and nonprivileged documents) to an e-discovery request; documents that are responsive and privileged should be put on a privilege log; nonresponsive documents should be withheld. In other words we model the problem as a classifier generating $h : \mathcal{D} \rightarrow \mathcal{C}$ such that $\mathcal{C} = \{c_P, c_L, c_W\}$ three target classes, where

- c_P is the class of the responsive nonprivileged documents, that should be Produced to the requesting party;
- c_L is the class of the responsive privileged documents, that should be entered on the privilege Log;
- c_W is the class of the nonresponsive documents, that should be Withheld by the producing party.

Since different classification errors bring about different costs, our problem defined above is quite sensitive to the value of the misclassification costs. For instance, producing a document that should have been on the privilege log typically brings about a higher cost than producing a document that should instead have been withheld. Hence we assume the existence of a cost matrix $\Lambda^m = \{\lambda^m_{ij}\}$ (for $i, j \in \{P, L, W\}$) as an input to our algorithm. The structure of the cost matrix is illustrated in Table 5.1(b) above, where each entry λ^m_{ij}, in *unit cost* is a nonnegative value representing the cost incurred when misclassifying an element of c_j into c_i (the m superscript stands for "misclassification").

In the next few sections, we explain the six baseline methods in detail and compare their performance against our MINECORE algorithm.

5.2 Fully Automated baseline model

In the fully automated baseline model, we train two automated classifiers h_r (binary classifier for responsiveness) and h_p (binary classifier for privilege), and we apply them to

Table 5.1: Contingency table D (a) and cost matrix Λ^m (b) for our problem.

		actual		
		c_P	c_L	c_W
pred	c_P	D_{PP}	D_{PL}	D_{PW}
	c_L	D_{LP}	D_{LL}	D_{LW}
	c_W	D_{WP}	D_{WL}	D_{WW}

(a)Contingency table D

		actual		
		c_P	c_L	c_W
pred	c_P	0	λ^m_{PL}	λ^m_{PW}
	c_L	λ^m_{LP}	0	λ^m_{LW}
	c_W	λ^m_{WP}	λ^m_{WL}	0

(b)Cost Matrix Λ^m

the collection $\mathcal{D}$. The classifiers are generated independently of each other. In this chapter we make the simplifying assumption that training and running automated classifiers has zero cost.

For each document $d \in \mathcal{D}$, h_r and h_p generate two posterior probabilities $\Pr(c_r|d)$ and $\Pr(c_p|d)$, which represent the classifiers' confidence in the fact that d is responsive and that d is privileged respectively. For $\Pr(c_r|d)$ a value of 1 represents total certainty that $d \in c_r$, a value of 0.5 represents total uncertainty, and a value of 0 represents total certainty that $d \in \overline{c}_r$; the same for $\Pr(c_p|d)$.

From $\Pr(c_r|d)$ and $\Pr(c_p|d)$, posterior probabilities $\Pr(c_P|d)$, $\Pr(c_L|d)$, $\Pr(c_W|d)$ are obtained as

$$\Pr(c_P|d) \equiv \Pr(c_r|d)\Pr(\overline{c}_p|d) \tag{5.1}$$

$$\Pr(c_L|d) \equiv \Pr(c_r|d)\Pr(c_p|d) \tag{5.2}$$

$$\Pr(c_W|d) \equiv \Pr(\overline{c}_r|d) \tag{5.3}$$

We next classify each document d in the class with the lowest expected cost using equation 5.4.

$$h(d) = \arg\min_{c_i} R(d, c_i) \tag{5.4}$$

where $R(d, c_i)$ (the risk associated with assigning d to class c_i) is defined as

$$R(d, c_i) = \sum_{j \in \{P,L,W\}} \lambda_{ij}^m \Pr(c_j|d) \tag{5.5}$$

As a result, the risk brought about by this classification is

$$R(\mathcal{D}) = \sum_{d \in \mathcal{D}} R(d, h(d)) \tag{5.6}$$

In other words, each document d is assigned a class (c_P or c_L or c_W) that brings about the minimum expected misclassification cost. Here the expected misclassification cost is computed as the sum of the misclassification costs of all possible events (i.e., classes to which d might truly belong), each multiplied by the probability of occurrence of the event (which is estimated by the classifier). For measuring misclassification cost, we use the equation below;

$$C^m(\mathcal{D}) = \sum_{i,j \in \{P,L,W\}} \lambda_{ij}^m D_{ij} \tag{5.7}$$

where the m superscript stands for "misclassification". Note that $C^m(\mathcal{D})$ is linear, i.e., it can alternatively be written as $C^m(\mathcal{D}) = \sum_{d \in \mathcal{D}} C^m(d)$, where $C^m(d) = \lambda_{h(d)y(d)}^m$ is the cost of predicting a document to be in class $h(d)$ when its true class is $y(d)$.

5.3 Fully Manual baseline model

In the Fully Manual baseline model a reviewer (typically: a junior lawyer) annotates all documents in $\mathcal{D}$ for responsiveness. All the documents in $\mathcal{D}$ that the reviewer deems responsive are forwarded to another reviewer (usually a senior lawyer) who annotates them for privilege, while all the others are withheld. All the documents that this latter reviewer deems nonprivileged are produced to the requesting party, while all the documents that the senior lawyer deems privileged are entered on the privilege log. The two reviewers usually work sequentially, rather than in parallel. This is justified by cost issues. It is a waste of resources to annotate by privilege a document that has already been ruled out

on counts of responsiveness, and that the reviewers who deal with responsiveness usually work at cheaper hourly rates than the reviewers who deal with privilege. This suggests to have a first pass carried out by the former before the latter intervene. We also assume, for ease of explanation, that there is only one reviewer for responsiveness and only one reviewer for privilege. In real applications there are often several reviewers of each type; however, what we describe straightforwardly applies to the case of more than one reviewer of each type. In this chapter we make the simplifying assumption that our reviewers are perfectly reliable (i.e., they do not make annotation errors); we defer the study of a model which relaxes this assumption to future work.

Let the pair $\Lambda^a = (\lambda_r^a, \lambda_p^a)$ denote the costs of annotating a single document for responsiveness (λ_r^a) and for privilege (λ_p^a), where the a superscript stands for "annotation". As a function for measuring annotation cost (which derives from the intervention of human reviewers) we use the equation below;

$$C^a(\mathcal{D}) = \lambda_r^a \tau_r + \lambda_p^a \tau_p \tag{5.8}$$

where τ_r and τ_p are the numbers of documents manually annotated for responsiveness and for privilege, respectively.

Note that for the fully manual solution, τ_r is the number of documents in $\mathcal{D}$, and τ_p is the number of responsive documents in $\mathcal{D}$. Similarly to the cost matrix Λ^m, we assume the unit costs in Λ^a to be input parameters, since they are not under the control of the experimenter.

5.4 Our MINECORE model

Both the baselines in section 5.2 and section 5.3 have drawbacks. The fully automatic model has the advantage of zero annotation cost, but bears the drawback of having non-negligible classifier error rate. As a consequence, this model is susceptible to the case of withholding documents that should have been produced, and (more dangerously) producing documents that should have been withheld. The costs generated by too many such misclassifications might be severe. On the other end, the fully manual model has

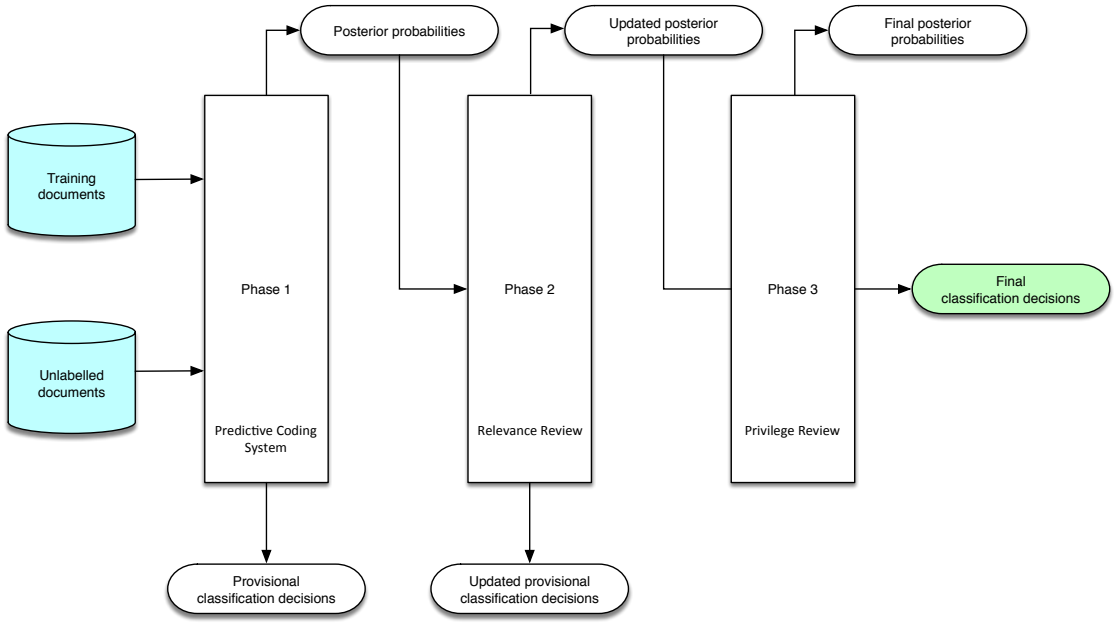

Figure 5.1: MINECORE Framework Overview

the advantage of perfect accuracy (assuming manual review is perfect) but is expensive, since the costs involved in manual annotation are high, and is sometimes infeasible, since it might be impossible to manually annotate each document given the time constraints imposed by the lawsuit.

Thus, we propose our MINECORE model where we try to strike a balance between the two. Figure 5.1 shows the overall architecture of our semi-automated MINECORE model. The execution of this model can be described in three phases;

1. All the documents in $\mathcal{D}$ are first assigned a class in $\{c_P, c_L, c_W\}$ by an automatic classifier that classifies according to Equation 5.4, following which

2. Junior annotators annotate a subset $\mathcal{D}'$ of the documents in $\mathcal{D}$ for responsiveness which may cause some of the documents in $\mathcal{D}'$ to be reassigned a class in $\{c_P, c_L, c_W\}$ different from the one assigned in Phase 1.

3. In the final phase, senior annotators annotate a subset $\mathcal{D}''$ of the documents in $\mathcal{D}$ for privilege, which may cause some of the documents in $\mathcal{D}''$ to be reassigned a class in $\{c_P, c_L, c_W\}$ different from the one assigned in Phase 1 and 2.

Of course, the right question here is how to strike an *optimal* balance, i.e., how to decide which documents should be annotated for responsiveness in Phase 2, and for privilege in Phase 3, and which others should instead be left unchecked. Our solution to arrive at such a balance makes use of

- the posterior probabilities $\Pr(c_r|d)$ and $\Pr(c_p|d)$ generated by the automated classifiers h_r and h_p;
- a cost matrix Λ^m and a pair Λ^a of unit annotation costs.

From now on, by the term cost structure we indicate a pair $\Lambda = (\Lambda^m, \Lambda^a)$, with Λ^m a cost matrix and Λ^a a pair $(\lambda^a_r, \lambda^a_p)$ of unit annotation costs. The only constraints we impose on Λ are that (i) all unit misclassification costs in Λ^m and both unit annotation costs in Λ^a must be nonnegative; (ii) all $\lambda^m_{ii} \in \Lambda^m$ must be 0; and (iii) it must hold that $\lambda^a_r \leq \lambda^a_p$.

Thus, the overall cost of the process can be quantified as

$$C^o(\mathcal{D}) = C^m(\mathcal{D}) + C^a(\mathcal{D}) \tag{5.9}$$

where the o superscript stands for "overall", and where $C^m(\mathcal{D})$ and $C^a(\mathcal{D})$ are the costs defined in Equations 5.7 and 5.8. $C^o(\mathcal{D})$ is the evaluation function we adopt in this work for all systems we experimentally compare, and not just for MINECORE. Note that for the fully automated solution $C^o(\mathcal{D})$ coincides with $C^m(\mathcal{D})$, since for this solution we have assumed the annotation cost to be zero, and for the fully manual solution $C^o(\mathcal{D})$ coincides with $C^a(\mathcal{D})$, since for this solution we have assumed the misclassification cost to be zero.

5.4.1 Document Ranking

MINECORE consists of an automatic classification phase (Phase 1), followed by two human annotation phases (Phase 2 and Phase 3) in which only the documents whose manual annotation is expected to reduce the overall cost are annotated. For each phase ϕ and for each document d, two posterior probabilities $\Pr_\phi(c_r|d)$ and $\Pr_\phi(c_p|d)$ are generated. Based on these probabilities, a class $h_\phi(d)$ is assigned in Phase ϕ to each document d as

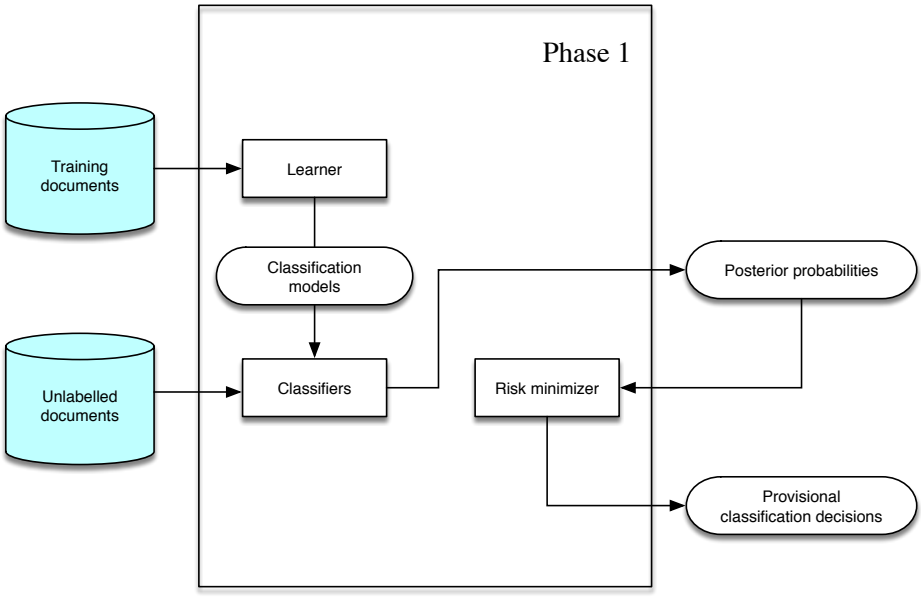

Figure 5.2: Phase 1 of the MINECORE Framework

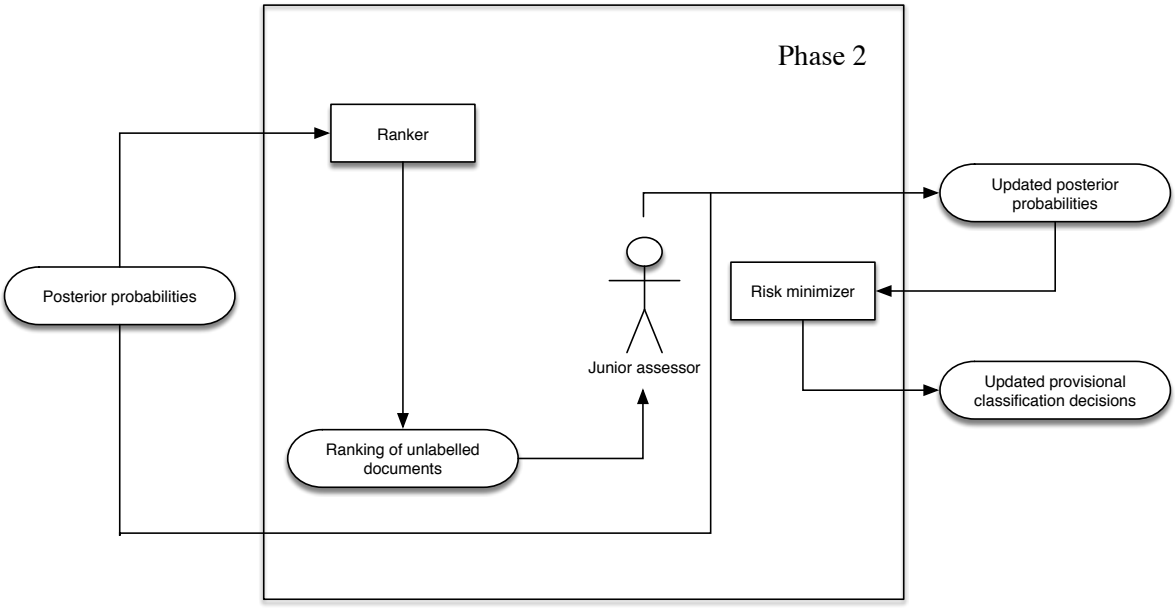

Figure 5.3: Phase 2 of the MINECORE Framework

$$h_\phi(d) = \arg\min_{c_i} R_\phi(d, c_i) = \arg\min_{c_i} \sum_{j \in \{P,L,W\}} \lambda_{ij}^m \Pr_\phi(c_j|d) \tag{5.10}$$

where c_i ranges on $\{c_P, c_L, c_W\}$. Equation 5.10 is just Equation 5.4 where the phase ϕ in which the probabilities are computed and the class is assigned is made explicit.

In Phase 1 of MINECORE, shown in figure 5.2, we train two automated classifiers, h_r (which classifies for responsiveness) and h_p (which classifies for privilege), from training data that we assume available, and we apply them to $\mathcal{D}$.

As in the fully automated solution described in Section 5.2, the two classifiers generate two posterior probabilities $\Pr_1(c_r|d)$ and $\Pr_1(c_p|d)$ for each document $d \in \mathcal{D}$. The two posterior probabilities represent the classifiers' confidence that d is responsive and that d is privileged, respectively. Using these posterior probabilities, we assign a class $h_1(d) \in \{c_P, c_L, c_W\}$ to each document $d \in \mathcal{D}$ using Equation 5.10.

In Phase 2 of MINECORE, shown in figure 5.3, the documents in $\mathcal{D}$ are ranked, and the reviewer (typically: a junior lawyer) annotates the top-ranked τ_r documents for responsiveness. Annotating d has the effect of eliminating the uncertainty on the responsiveness of d. As a consequence, if d is annotated as responsive we set $\Pr_2(c_r|d) = 1$, while if d is annotated as nonresponsive we set $\Pr_2(c_r|d) = 0$; no annotation for privilege is performed in this phase, so $\Pr_1(c_p|d) = \Pr_2(c_p|d)$. At this point, by using Equation 5.10, d is assigned a class $h_2(d) \in \{c_P, c_L, c_W\}$, which is possibly different from $h_1(d)$.

The documents d from the $(\tau_r + 1)$-th position onwards are not manually annotated; everything remains unchanged for these documents, i.e., $\Pr_2(c_r|d) = \Pr_1(c_r|d)$ and $\Pr_2(c_p|d) = \Pr_1(c_p|d)$, which implies that $h_2(d) = h_1(d)$.

In order to maximize the cost-effectiveness of this approach it is necessary to choose (i) an optimal ranking of the documents in $\mathcal{D}$ and (ii) an optimal threshold τ_r (which acts as the stopping condition for the annotation process).

Concerning point (i), similarly to the approach of [16] we adopt the principle that the documents in $\mathcal{D}$ are to be ranked in terms of the reduction in overall risk that annotating the document brings about; the documents whose manual annotation brings about the highest reduction are top-ranked. If by $C_\phi^m(d)$ we indicate the misclassification cost

brought about by attributing class $h_\phi(d)$ to d, the difference $\Delta^{or}(d)$ in overall cost that annotating d for responsiveness brings about can be written (using Equation 5.9) as

$$\begin{aligned}\Delta^{or}(d) &= C_2^o(d) - C_1^o(d) \\ &= C_2^m(d) + C_2^a(d) - C_1^m(d) - C_1^a(d) \\ &= C_2^m(d) + \lambda_r^a - C_1^m(d)\end{aligned} \tag{5.11}$$

However, at the time of ranking $\mathcal{D}$ the true class of d (noted as $y(d)$) is not known, so $C_1^m(d)$ and $C_2^m(d)$ are also unknown. Therefore, at the time of ranking $\mathcal{D}$ what we can actually compute, instead of $\Delta^{or}(d)$, is an *expectation* of $\Delta^{or}(d)$ over the $y(d)$ random variable, i.e.,

$$\begin{aligned}E_y[\Delta^{or}(d)] &= E_y[C_2^m(d) + \lambda_r^a - C_1^m(d)] \\ &= E_y[C_2^m(d)] + \lambda_r^a - E_y[C_1^m(d)] \\ &= R_2(d, h_2(d)) + \lambda_r^a - R_1(d, h_1(d))\end{aligned} \tag{5.12}$$

Actually, at the time of ranking $\mathcal{D}$ we also do not know the value of the $y_r(d)$ variable (a binary variable that indicates whether, if the reviewer had to annotate d, s/he would deem it responsive or not). This means that also the class $h_2(d)$ that would be assigned as a result of annotating d is not known. $R_2(d, h_2(d))$ is thus not known either, which means that Equation 5.12 cannot be used directly as a criterion for ranking $\mathcal{D}$.

At the time of ranking $\mathcal{D}$ we thus must compute an expectation of $E_y[\Delta^{or}(d)]$ over the $y_r(d)$ random variable, i.e.,

$$\begin{aligned}E_{y_r y}[\Delta^{or}(d)] &= E_{y_r}[R_2(d, h_2(d)) + \lambda_r^a - R_1(d, h_1(d))] \\ &= E_{y_r}[R_2(d, h_2(d))] + \lambda_r^a - R_1(d, h_1(d))\end{aligned} \tag{5.13}$$

where we have shortened $E_{y_r}[E_y[\cdot]]$ as $E_{y_r y}[\cdot]$, and where the last simplification is justified by the fact that $R_1(d, h_1(d))$ does not depend on $y_r(d)$.

$E_{y_r}[R_2(d, h_2(d))]$ is computed by assigning probabilities to the events c_r (i.e., "the reviewer annotates d as responsive") and $\overline{c}_r$ ("the reviewer annotates d as nonresponsive"). To do this, the best we can do is to "trust" our classifiers and assume that d will be

annotated as responsive with probability $\Pr_1(c_r|d)$ and nonresponsive with probability $\Pr_1(\overline{c}_r|d)$. Each of these probabilities is multiplied by the misclassification risk that the annotation would bring about, i.e.,

$$E_{y_r}[R_2(d, h_2(d))] = R_2(d, h_2(d)|c_r) \cdot \Pr_1(c_r|d) + R_2(d, h_2(d)|\overline{c}_r) \cdot \Pr_1(\overline{c}_r|d) \tag{5.14}$$

where by $R_2(d, h_2(d)|c_r)$ we indicate the misclassification risk that would result from assuming that $\Pr_2(c_r|d) = 1$ and $\Pr_2(c_p|d) = \Pr_1(c_p|d)$, and by $R_2(d, h_2(d)|\overline{c}_r)$ we indicate the misclassification risk that would result from assuming that $\Pr_2(c_r|d) = 0$ and $\Pr_2(c_p|d) = \Pr_1(c_p|d)$.

Equation 5.13 finally gives us a concrete method for ranking the automatically classified documents: for each $d \in \mathcal{D}$ compute $E_{y_r y}[\Delta^{or}(d)]$ (the expected increase in overall cost brought about by annotating d for responsiveness), and rank the documents in $\mathcal{D}$ according to their $E_{y_r y}[\Delta^{or}(d)]$ score, top-ranking those with the *lowest* scores. This guarantees that the reviewer will first annotate the documents characterized by the highest expected *reduction* in cost that manually annotating them would bring about. In turn this guarantees that, whatever the amount τ_r of documents that the reviewers annotate, the expected cost-effectiveness of the annotation work will be maximized.

Equation 5.13 gives us also a concrete method for addressing point (ii) above, i.e., for setting the τ_r threshold. The overall cost $C^o(d)$ is expected to decrease as a result of annotating d (i.e., $E_{y_r y}[\Delta^{or}(d)] < 0$) when the cost λ_r^a of annotating d is more than offset by the expected reduction $(R_1(d, h_1(d))) - E_{y_r}[R_2(d, h_2(d))]$ in misclassification cost that annotating d brings about; conversely, if $E_{y_r y}[\Delta^{or}(d)] \geq 0$, then the expected reduction in misclassification cost is not worth the additional annotation effort. Therefore, the criterion we adopt is in order to decide when to stop annotating is:

> **Stopping condition (responsiveness)**. Let d be the document at the k-th rank position. If $E_{y_r y}[\Delta^{or}(d)] < 0$, then annotate d by responsiveness and move on to the document in the $(k+1)$-th rank position, else stop annotating.

The rationale for this criterion is that a reviewer will annotate a document only if

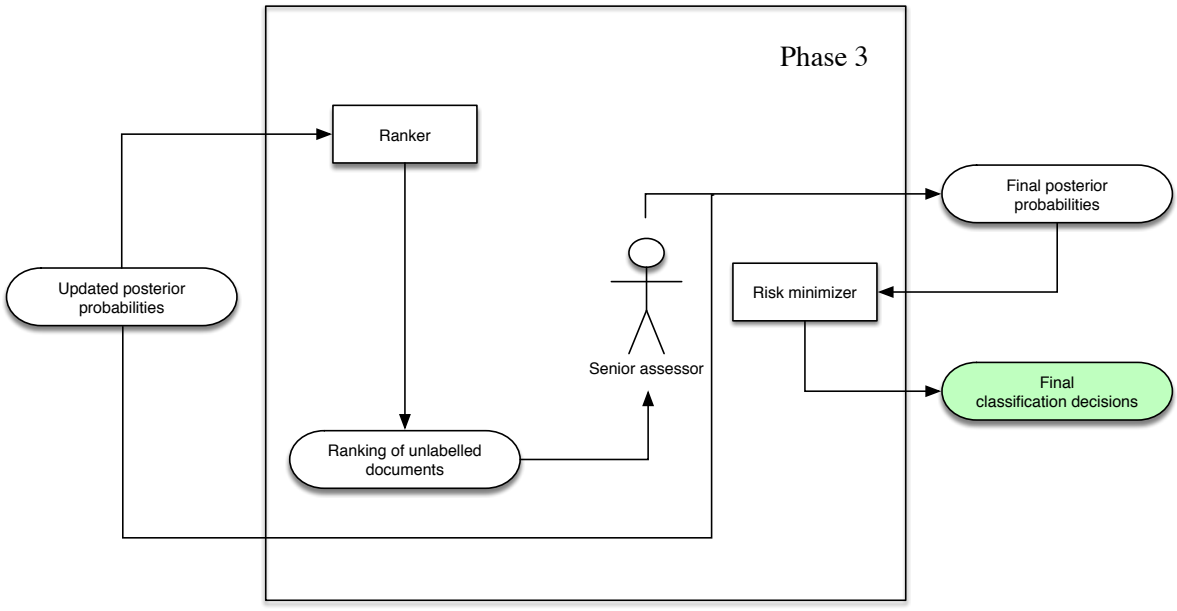

Figure 5.4: Phase 3 of the MINECORE Framework

this action is expected to diminish overall cost. Since the likelihood of diminishing overall cost decreases the more we go down the ranking, it follows that we should choose τ_r to be

$$\tau_r = |\{d | E_{y_r y}[\Delta^{or}(d)] < 0\}| \tag{5.15}$$

At this point, in Phase 2 the human reviewer has manually annotated the τ_r documents characterized by the lowest value of $E_{y_r y}[\Delta^{or}(d)]$.

Phase 3 of MINECORE, shown in figure 5.4, does for privilege essentially what Phase 2 did for responsiveness. In Phase 3 the documents in $\mathcal{D}$ are again ranked, and the reviewer (typically: a senior lawyer) annotates the top-ranked τ_p documents for privilege. If the reviewer annotates d as privileged we set $\Pr_3(c_p|d) = 1$, while if the reviewer annotates d as nonprivileged we set $\Pr_3(c_p|d) = 0$; no annotation for responsiveness is performed in this phase, so $\Pr_2(c_r|d) = \Pr_3(c_r|d)$. At this point, by using Equation 5.10, d is assigned a class $h_3(d) \in \{c_P, c_L, c_W\}$, which is possibly different from $h_2(d)$. The documents d from the $(\tau_p + 1)$-th position onwards are not manually annotated for privilege; for these documents, $\Pr_3(c_r|d) = \Pr_2(c_r|d)$ and $\Pr_3(c_p|d) = \Pr_2(c_p|d)$, which implies that $h_3(d) = h_2(d)$. Class $h_3(d) \in \{c_P, c_L, c_W\}$ is the final class assigned to d by MINECORE, and the class that determines whether the document is produced to the

requesting party ($h_3(d) = c_P$), entered on the privilege log ($h_3(d) = c_L$), or withheld ($h_3(d) = c_W$).

The difference $\Delta^{op}(d)$ in overall cost that annotating d for privilege brings about is

$$\begin{aligned}\Delta^{op}(d) &= C_3^o(d) - C_2^o(d) \\ &= C_3^m(d) + C_3^a(d) - C_2^m(d) - C_2^a(d) \\ &= C_3^m(d) + \lambda_p^a - C_2^m(d)\end{aligned} \tag{5.16}$$

Similarly to Equation 5.11, and for the same reasons, Equation 5.16 cannot be used directly as a criterion for ranking $\mathcal{D}$. At the time of ranking $\mathcal{D}$ we thus compute the expected difference in cost

$$\begin{aligned}E_y[\Delta^{op}(d)] &= E_y[C_3^m(d) + \lambda_p^a - C_2^m(d)] \\ &= E_y[C_3^m(d)] + \lambda_p^a - E_y[C_2^m(d)] \\ &= R_3(d, h_3(d)) + \lambda_p^a - R_2(d, h_2(d))\end{aligned} \tag{5.17}$$

Due to the fact that the value of $y_p(d)$ (a binary variable that indicates whether, if the reviewer had to annotate d, s/he would deem it privileged or not) is not known at the time of ranking, we must compute an expectation of $E_y[\Delta^{op}(d)]$ over the $y_p(d)$ random variable, i.e.,

$$\begin{aligned}E_{y_p y}[\Delta^{op}(d)] &= E_{y_p}[R_3(d, h_3(d)) + \lambda_p^a - R_2(d, h_2(d))] \\ &= E_{y_p}[R_3(d, h_3(d))] + \lambda_p^a - R_2(d, h_2(d))\end{aligned} \tag{5.18}$$

where we have shortened $E_{y_p}[E_y[\cdot]]$ as $E_{y_p y}[\cdot]$. To compute $E_{y_p}[R_3(d, h_3(d))]$, we assume that d will be annotated as privileged with probability $\Pr_1(c_p|d)$ and nonprivileged with probability $\Pr_1(\overline{c}_p|d)$, thus bringing about

$$E_{y_p}[R_3(d, h_3(d))] = R_3(d, h_3(d)|c_p) \cdot \Pr_1(c_p|d) + R_3(d, h_3(d)|\overline{c}_p) \cdot \Pr_1(\overline{c}_p|d) \tag{5.19}$$

Analogously to Equation 5.13, Equation 5.18 now gives us a concrete method for ranking the documents: rank the documents in $\mathcal{D}$ according to their $E_{y_p y}[\Delta^{op}(d)]$ score, top-ranking those with the lowest scores. The same equation also gives us a concrete

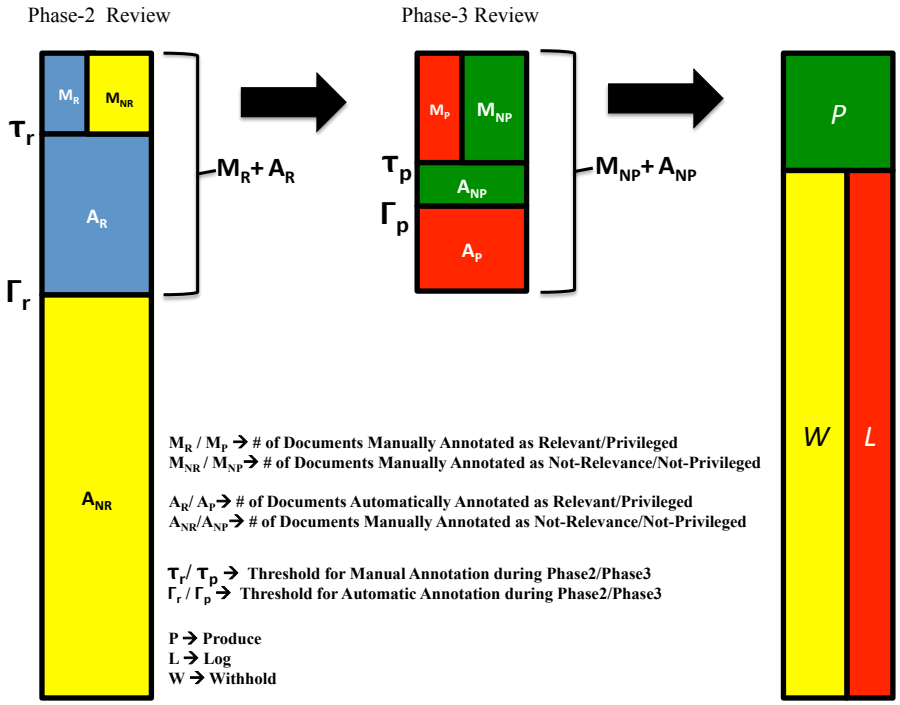

Figure 5.5: Model Parameters

method for setting the τ_p threshold: along the same lines discussed for Phase 2, the criterion we adopt in order to decide when to stop annotating is:

> **Stopping condition (privilege).** Let d be the document at the k-th rank position. If $E_{y_p y}[\Delta^{op}(d)] < 0$, then manually annotate d by privilege and move on to the document in the $(k+1)$-th rank position, else stop annotating.

and we should choose τ_p to be

$$\tau_p = |\{d | E_{y_p y}[\Delta^{op}(d)] < 0\}| \tag{5.20}$$

Thus to summarize, equations 5.13 and 5.18 gives us a concrete method for ranking the automatically classified documents. The rank order guarantees that the assessor will first annotate the documents characterized by the highest reduction in expected cost that manually annotating them would bring about. In turn this guarantees that, whatever the total number of documents that the assessors annotate, the expected cost-effectiveness of

the annotation work will be maximized in both Phase 2 and Phase 3. This also gives us a criterion for deciding when to stop the manual annotation. Let us assume that, d is the document at the $k-th$ rank position, and assume we are considering whether annotating d or stop annotating. The criterion we adopt is: If $\Delta_{or}(d) < 0$, then annotate d and move on to the document in the $(k+1)-th$ rank position, else stop annotating. The condition for review is that an assessor will annotate a document only if this action is expected to exceed the unit annotation cost for a document, i.e., if the cost of annotating the document is expected to be offset by a decrease in misclassification cost.

5.4.2 Algorithm & Evaluation Plan Overview

The overall algorithm of our MINECORE model can be depicted as shown in figure 5.1. In Phase 1 of our model we train two automated classifiers h_r (which classifies according to responsiveness) and h_p (which classifies according to privilege) from the training data (that we assume is available at a zero labeling cost) and we apply them to set D. The documents in D are ranked using equation 5.11 (section 5.4.1). In Phase 2, the assessor (typically: a junior lawyer) annotates the top-ranked (τ_r) documents for responsiveness. As shown in figure 5.5, at this point, we have M_R documents manually annotated as responsive and M_{NR} documents manually annotated as not-responsive. The documents that are not manually annotated in Phase 2 fall into two categories based on our automatic classifier results. Thus we have A_R documents automatically classified as responsive and A_{NR} documents automatically classified as not-responsive. At the end of Phase 2, we obtain a responsive document set $D^{'} = (M_R + A_R)$ which is then passed on to Phase 3 for manual privilege annotation.

In Phase 3, The documents in $D^{'}$ are ranked using equation 5.16 and the assessor (typically: a senior lawyer) annotates a total of τ_p documents for privilege. After the Phase 3 annotation step, (similar to Phase 2) we have M_P documents manually annotated as privilege and M_{NP} documents manually annotated as not-privileged. We divide the documents that are not manually annotated in Phase 3 into two categories based on our automatic classifier results. Thus we have A_P documents automatically classified as

Input : A training set Tr_r of documents labeled for responsiveness;
A training set Tr_p of documents labeled for privilege;
Documents $\mathcal{D}$ to be analysed for possible production to the requesting party;
Cost structure $\Lambda = (\Lambda_m, \Lambda_a)$.
Output: A partition of $\mathcal{D}$ into the following three sets:
– Set $\mathcal{D}_P$ of documents to be produced to the receiving party;
– Set $\mathcal{D}_L$ of documents to be put on the privilege log;
– Set $\mathcal{D}_W$ of documents to be withheld;
Annotation cost $C^a(\mathcal{D})$ incurred in the process.

`/* Phase 1 */`
Train classifiers h_r and h_p from Tr_r and Tr_p, respectively;
for $d \in \mathcal{D}$ **do**
  Compute $\Pr_1(c_r|d)$ by means of h_r and $\Pr_1(c_p|d)$ by means of h_p;
  Compute $h_1(d)$ via Equation 5.10;
end

`/* Phase 2 */`
for $d \in \mathcal{D}$ **do**
  $\Pr_2(c_r|d) \leftarrow \Pr_1(c_r|d)$; $\Pr_2(c_p|d) \leftarrow \Pr_1(c_p|d)$; Compute $E_{y_r y}[\Delta^{or}(d)]$ using Equation 5.13;
end
Generate a ranking $R^r_{\mathcal{D}}$ of $\mathcal{D}$ in increasing order of $E_{y_r y}[\Delta^{or}(d)]$;
`/*` $R^r_{\mathcal{D}}(k)$ `denotes the document at the` k`-th rank position in` $R^r_{\mathcal{D}}$ `*/`
$k \leftarrow 1$; $\tau_r \leftarrow 0$;
while $E_{y_r y}[\Delta^{or}(R^r_{\mathcal{D}}(k))] < 0$ **do**
  Have the reviewer annotate document $R^r_{\mathcal{D}}(k)$ for responsiveness;
  if $R^r_{\mathcal{D}}(k)$ *has been judged responsive by the reviewer* **then**
    $\Pr_2(c_r|R^r_{\mathcal{D}}(k)) \leftarrow 1$
  else
    $\Pr_2(c_r|R^r_{\mathcal{D}}(k)) \leftarrow 0$
  end
  $\tau_r \leftarrow \tau_r + 1$; $k \leftarrow k + 1$;
end
for $d \in \mathcal{D}$ **do**
  Compute $h_2(d)$ using Equation 5.10;
end

`/* Phase 3 */`
for $d \in \mathcal{D}$ **do**
  $\Pr_3(c_r|d) \leftarrow \Pr_2(c_r|d)$; $\Pr_3(c_p|d) \leftarrow \Pr_2(c_p|d)$; Compute $E_{y_p y}[\Delta^{op}(d)]$ using Equation 5.18;
end
Generate a ranking $R^p_{\mathcal{D}}$ of $\mathcal{D}$ in increasing order of $E_{y_p y}[\Delta^{op}(d)]$;
`/*` $R^p_{\mathcal{D}}(k)$ `denotes the document at the` k`-th rank position in` $R^p_{\mathcal{D}}$ `*/`
$k \leftarrow 1$; $\tau_p \leftarrow 0$;
while $E_{y_p y}[\Delta^{op}(R^p_{\mathcal{D}}(k))] < 0$ **do**
  Have the reviewer annotate document $R^p_{\mathcal{D}}(k)$ for privilege;
  if $R^p_{\mathcal{D}}(k)$ *has been judged privileged by the reviewer* **then**
    $\Pr_3(c_p|R^p_{\mathcal{D}}(k)) \leftarrow 1$
  else
    $\Pr_3(c_p|R^p_{\mathcal{D}}(k)) \leftarrow 0$
  end
  $\tau_p \leftarrow \tau_p + 1$; $k \leftarrow k + 1$;
end
for $d \in \mathcal{D}$ **do**
  Compute $h_3(d)$ using Equation 5.10;
end

$\mathcal{D}_P \leftarrow \{d|h_3(d) = c_P\}$; $\mathcal{D}_L \leftarrow \{d|h_3(d) = c_L\}$; $\mathcal{D}_W \leftarrow \{d|h_3(d) = c_W\}$;
Compute $C^a(\mathcal{D})$ using Equation 5.8.

Algorithm 1: MINECORE, a model for MINimizing the Expected COsts of REview for responsiveness and privilege.

privilege and A_{NP} documents automatically classified as not privilege.

Equation 5.9 is the primary evaluation function we will adopt. When running the experiment, (in which we indeed know the labels of the test documents) we will compute, at the end of the process, the overall cost $C_o(D)$ of the process for each of our 6 baseline models and MINECORE. We compute the cost of manual annotation $C_a(D)$ using equation 5.8, and the cost of misclassification $C_m(D)$ using estimates as in equation 5.7. Since, due to the involvement of an automatic classification component, we are in the presence of uncertainty, in developing our MINECORE method we use a risk minimization approach, where we try to minimize an expectation over the overall cost described in Equation 5.9; i.e., we want to minimize

$$E[C_o(\mathcal{D})] = E[C_m(\mathcal{D}) + C_a(\mathcal{D})] \tag{5.21}$$

where $E[\cdot]$ stands for "expected value". Note that $E[C_m(\mathcal{D}) + C_a(\mathcal{D})]$ does *not* break down as $E[C_m(\mathcal{D})] + E[C_a(\mathcal{D})]$, since $C_m(\mathcal{D})$ and $C_a(\mathcal{D})$ are not independent. That is, we can easily bring down $C_m(\mathcal{D})$ to zero by choosing to manually annotate all documents, which would however make $C_a(\mathcal{D})$ very high; and we can easily bring down $C_a(\mathcal{D})$ to zero by choosing to automatically annotate all documents, which would however make $C_m(\mathcal{D})$ very high. Thus our attempt is to *jointly* minimize $E[C_m(\mathcal{D})]$ and $E[C_a(\mathcal{D})]$.

The overall algorithm that implements MINECORE is summarized in Algorithm 1.

5.5 Other baselines

We are here proposing some baseline methods against which to compare MINECORE. Throughout this chapter we use the same vector representations for the documents, the same supervised learning algorithm, and the same classifier outputs, for all the methods being compared. Each method (be it MINECORE or a baseline method) assigns, for each test document d, a class in $\mathcal{C} = \{c_P, c_L, c_W\}$.

Our baseline methods are (aside from the fully automated and fully manual solutions) mixed-initiative, "human-in-the-loop" systems, i.e., their classification decisions are obtained via some combination of manual annotation work and automatic classifi-

cation. Using the cost structures exemplified in Table 5.2 we can evaluate each system using the evaluation measure described in Equation 5.21; that is, for each system we compute the misclassification cost $C^m(\mathcal{D})$, the annotation cost $C^a(\mathcal{D})$, and the overall cost $C^o(\mathcal{D}) = C^m(\mathcal{D}) + C^a(\mathcal{D})$ they incur. The best system is the one with the lowest $C^o(\mathcal{D})$ cost.

5.5.1 Uncertainty Ranking

In Uncertainty Ranking or UR we first annotate for responsiveness the τ_r documents whose $\Pr(c_r|d)$ is closest to 0.5 (i.e., the ones whose responsiveness is most uncertain). A document is then deemed responsive if the reviewer has annotated it as such, or (for the documents which have not been manually annotated for responsiveness) if $\Pr(c_r|d) > 0.5$. We then annotate for privilege, among the documents that have been deemed responsive, the τ_p documents whose $\Pr(c_p|d)$ is closest to 0.5. A document is then deemed privileged if the reviewer has annotated it as such, or (for the documents which have not been manually annotated for privilege) if $\Pr(c_p|d) > 0.5$. This baseline is similar to MINECORE in that the class assigned to a test document may result from the reviewers' manual annotation work, or from the automated classifiers, or from a combination of them. However, neither annotation costs nor misclassification costs play a role in UR.

5.5.2 Relevance Ranking

In Relevance Ranking or RR we first annotate for responsiveness the τ_r documents with the *highest* $\Pr(c_r|d)$, and we then annotate for privilege, among the documents that the reviewers have deemed responsive in the previous phase, the τ_p documents with the *lowest* $\Pr(c_p|d)$. Unlike MINECORE and UR, RR assumes that only the documents that have been certified responsive and nonprivileged by the reviewers are going to be produced (documents certified responsive and privileged by the reviewers are entered on the privilege log, while all other documents are withheld); as a result, the two rankings (by $\Pr(c_r|d)$ and $\Pr(c_p|d)$) attempt to top-rank the documents that have the highest chances of meeting the requirements (responsiveness *and* nonprivilege) for disclosure.

5.5.3 Active Learning via Uncertainty Sampling

In the design of MINECORE our focus has been on cases in which we have already built the best classifier that we can, and in such cases we would not expect further gains from active learning. In our experiments, however, we have simply trained on a fixed set of documents, and it is possible that active learning might indeed give further gains.

This motivates our choice to include ALvUS and ALvRS (see below) as additional baselines. In ALvUS, an interactive process asks the reviewer to annotate for responsiveness the k documents in $\mathcal{D}$ for which $\Pr(c_r|d)$ is closest to 0.5; at this point, this set $\mathcal{D}'$ of k documents is added to the training set, the posterior probabilities $\Pr(c_r|d)$ of the documents d annotated as responsive are set to 1, h_r is retrained, and $\mathcal{D}/\mathcal{D}'$ is classified for responsiveness again; this process is repeated (using the newly computed $\Pr(c_r|d)$ values) until exactly τ_r documents have been annotated.[2] After this, an identical process is used for privilege, substituting h_p and τ_p for h_r and τ_r in the above. At the end, a document $d \in \mathcal{D}$ is assigned to c_P iff $\Pr(c_r|d) > 0.5$ and $\Pr(c_p|d) \leq 0.5$; to c_L iff $\Pr(c_r|d) > 0.5$ and $\Pr(c_p|d) > 0.5$; and to c_W otherwise. ALvUS is similar to MINECORE and UR, in that the class assigned to a test document may result from the reviewers' manual annotation work, or from the automated classifiers, or from a combination of them. In the experiments reported in this chapter we use $k = 1000$, which was found to work well by [26].

Note that the comparison between MINECORE and ALvUS is only partially fair, since ALvUS is much more expensive computationally, given that it requires $\lceil \tau_r/k \rceil + \lceil \tau_p/k \rceil$ retraining operations (unlike MINECORE, which requires none).

5.5.4 Active Learning via Relevance Sampling

A variant of the previous baseline is obtained if the active learning process asks the reviewer to annotate for responsiveness the k documents in $\mathcal{D}$ for which $\Pr(c_r|d)$ *is highest* (and the ones for which $\Pr(c_p|d)$ is *lowest* when the reviewer annotates for privilege). The rest of the process is as in ALvUS; in particular, here too we use $k = 1000$. At the end,

[2] To be more precise, in the last iteration fewer than k documents may be annotated, so as to make the total number of documents annotated equal to τ_r. For example, if $\tau_r = 3267$ and $k = 1000$, 1000 documents will be annotated in each of the first three rounds, while in the final round only 267 documents will be annotated.

a document $d \in \mathcal{D}$ is assigned to c_P iff it has been manually annotated as responsive and nonprivileged; it is assigned to c_L iff it has been manually annotated as responsive and privileged; it is assigned to c_W otherwise. Unlike ALvUS, ALvRS thus assumes that, unless a document has been under the scrutiny of *both* the junior reviewer (for responsiveness) and the senior reviewer (for privilege), it is withheld. Among e-discovery researchers and practitioners, ALvRS is known as "continuous active learning" (CAL) [26, 27, 30]; was originally introduced in [46], where it was indeed called "Relevance Sampling".[3] The latter paper is also the work in which ALvUS was introduced first, under the name of "Uncertainty Sampling".

Note that for every baseline system other than FA and FM we compute the cost $C^o(\mathcal{D})$ that the baseline incurs when manually annotating exactly τ_r documents for responsiveness and, if possible,[4] τ_p documents for privilege, where τ_r and τ_p are the values used in the MINECORE system. This policy may be biased in favour of MINECORE, since τ_r and τ_p are optimal settings for MINECORE whereas other systems might have yielded lower overall costs with either more or less manual reviewing. However, none of the baseline systems we test have an apriori way of analytically setting the optimal number of documents to manually review. This means that our comparisons are, if not to post-hoc optimal systems, at least to reasonable systems.

5.6 Experiments

In this section we describe a number of experiments that we have conducted to test the cost-effectiveness of MINECORE.

[3] CAL, as described in [26, 27, 30], is actually a simpler variant of ALvRS since it deals with one classification task only (i.e., responsiveness), instead of the two cascaded tasks (i.e., responsiveness and privilege) that ALvRS deals with.

[4] In some cases a baseline system might deem responsive *fewer* than τ^p documents, which means that fewer than τ^p documents (i.e., all the ones deemed responsive) would be annotated for privilege; in this case the comparison between this baseline system and all other systems (including MINECORE) is still fair, though, since this system will incur a smaller annotation cost (for privilege) than MINECORE.

5.6.1 Test Collection

One problem that hinders the evaluation of MINECORE is that, in the world of e-discovery, at present, there is no publicly available test collection that is annotated by both responsiveness and privilege. The TREC 2010 Legal Track included a privilege topic and several responsiveness topics, but each topic was independently sampled so there are very few privilege annotations on documents that were annotated for relevance. Chapter 3 further discusses the issues with the TREC 2010 collection.

One solution is to generate such an annotated collection ourselves: however, this would be a major feat in terms of annotation cost, since it takes real lawyers to do this annotation, and real lawyers (especially senior ones, whom we would need in order to annotate for privilege) can be extremely expensive. We bypass this problem by running "simulated" experiments, on a collection unrelated to e-discovery in which documents can belong to more than one class, and by repeatedly picking two classes to play the role of c_r and c_p, respectively.

As a test collection we have chosen RCV1-v2, a standard, publicly available benchmark for text classification first presented in [47] and consisting of 804,414 news stories produced by Reuters from 20 Aug 1996 to 19 Aug 1997.[5] RCV1-v2 ranks as one of the largest corpora currently used in text classification research. RCV1-v2 is multi-label, i.e., a document may belong to several classes at the same time, which makes it suitable for our purposes. In [47] the collection is partitioned into a training set of 23,149 documents and a test set of 781,265 documents, the latter being split into four chunks of 199,328, 199,339, 199,576, 183,022 documents, respectively. In the experiments reported in this chapter we have used the 23,149 training documents as the training set Tr, and the first chunk of 199,328 test documents as the test set Te.

In the topic hierarchy of RCV1-v2 there are 103 classes, of which 101 have at least one positive training example. Since we experiment with pairs of classes (representing c_r and c_p), we could in principle experiment with $101^2 = 10,201$ different pairs. Aside from representing a substantive computational load, this would also mean experimenting

[5] `http://trec.nist.gov/data/reuters/reuters.html`

with classes whose prevalence is, for many e-discovery scenarios, not realistic. We have therefore limited our experiments to pairs (c_r, c_p) such that the test set prevalence of c_r (i.e., $\Pr(c_r|Tr)$) is in [0.03,0.07] and the prevalence of c_p in the responsive documents (i.e., $\Pr(c_p|c_r, Tr)$) is in [0.01,0.20]. These values are representative of some e-discovery settings, and they yield a sufficient number of positive labels for our experiments. For each of the 24 responsiveness classes that meet the responsiveness prevalence criterion we have randomly selected 5 privilege classes that meet the privilege prevalence criterion. This gives rise to 120 class pairs, which is the set we use for the experiments described in this chapter.

5.6.2 The learning algorithm

For all the experiments reported in this chapter we have used Support Vector Machines (SVMs) as the classifier, since they have consistently delivered strong performance in text classification. We have used the well-known SVM^{light} implementation for which we have used the default parameter values [38,39]. Concerning the vector representations fed to the SVM learner, we have used the ones made available during the creation of the Reuters collection [47]. We refer to that work for details on the preprocessing techniques that were used to generate them.

SVMs return confidence scores that are not posterior probabilities; these scores must thus be converted into posterior probabilities, since MINECORE essentially depends on the availability of posterior probabilities. Given that the returned scores are a monotonically increasing function of the classifier's confidence in the fact that the document belongs to the class, this conversion may be obtained by applying to the scores a logistic function, since such a function has a sigmoidal shape. We obtain well-calibrated posterior probabilities (defined as the posterior probabilities $\Pr(c|d)$ such that, given class c and set s, $\sum_{d \in s} \Pr(c|d)$ is equal to the class prevalence $\Pr(c|s)$) by using a *generalized* logistic function and optimizing its slope parameter; for this optimization we follow exactly the same process as described in [16], to which we refer the reader for details.

5.6.3 Cost structures

In order to use realistic misclassification costs and annotation costs, we have chosen to elicit our cost structures from e-discovery experts. We have been able to obtain the help of three senior members of the e-discovery community; two lawyers and an technical expert in technology-assisted review, each of whom have extensive experience with actual e-discovery cases in their practice.

We asked the two lawyers to think of an actual case they may be familiar with, and to articulate the cost structure that characterizes that case. To be sure to understand their cost values, we conducted a 60 minute call with each of the two lawyers. During the call, the lawyers explicated their rationale behind choosing the cost values. We took a different approach to gather the cost structure inputs from an e-discovery professional who is an expert in TAR. We developed a questionnaire (For details please refer Appendix B) with a total of eight questions. Answers to all of the eight questions were made mandatory since a partially filled out questionnaire would be less useful to us. Each question except the first has three possible answers. The task was to pick a single answer by ticking one of the three boxes, and then to fill in the requested relative cost value. The expert attempting to answer the questionnaire was allowed to make any assumption about the type of case and the amounts at stake in the case, but required to make the same assumptions for every question.

Through this process we obtained 3 cost structures, which are detailed in Table 5.2. Each individual cost is expressed in US$. Note that the values indicated by the 3 experts are in some cases markedly different (e.g., there is a factor of 150 between the values of λ_{LP} indicated by two of the experts); this need not be taken as evidence of disagreement among the experts for decisions on the same task, since different experts were free to choose different legal cases to have in mind when arriving at these estimates. Rather than trying to reconcile these cost structures in any way, we have thus run 3 experiments, one for each of the cost structures, on the assumption that MINECORE should be able to cater to different application needs.

Table 5.2: Cost structure values in US$.

	λ_r^a	λ_p^a	λ_{PL}	λ_{PW}	λ_{LP}	λ_{LW}	λ_{WP}	λ_{WL}
CostStructure1	1.00	5.00	600.00	5.00	150.00	3.00	15.00	15.00
CostStructure2	1.00	5.00	100.00	0.03	10.00	2.00	8.00	8.00
CostStructure3	1.00	5.00	1000.00	0.10	1.00	1.00	1.00	1.00

5.6.4 Experimental protocol

The experimentation protocol we adopt is the following. As groundwork, we train our binary classifiers via the chosen binary learner using the 23,149 training documents, and apply them to the 199,328 test documents (the test set Te thus plays the role of our universe $\mathcal{D}$). For each document $d \in Te$, the classifier for class c generates a confidence score, from which we obtain a posterior probability $\Pr(c|d)$ via probability calibration.

At this point, we run each of the seven methods (MINECORE plus the six baseline methods) for each of the cost structures (see Table 5.2) we have elicited from the experts.

In particular, for the risk minimization method, we first simulate the manual annotation process for responsiveness: for all $d \in \mathcal{D}$ such that $E_{y_r y}[\Delta^{or}(d)] < 0$ we set $\Pr_2(c_r|d)$ to 1 if d is responsive and to 0 if d is nonresponsive. We then do the same for privilege: for all $d \in \mathcal{D}$ such that $E_{y_p y}[\Delta^{op}(d)] < 0$ we set $\Pr_3(c_p|d)$ to 1 if d is privileged and to 0 if d is nonprivileged. We then compute the total cost of the process via Equation 5.21.

5.7 Results

In this section we present the results of testing MINECORE against the 6 baseline methods presented in Section 5.5, on the 120 class pairs described at the end of Section 5.6.1; we have run each such experiment for each of the 3 cost structures discussed in Section 5.6.3.

In Table 5.3 we exemplify, on a sample cost structure (CostStructure1), what the results look like. The table reports, the class prevalences of c_r and c_p, the values of τ_r and τ_p that MINECORE returns, and the $C^o(\mathcal{D})$ value (expressed in thousands of US$) resulting from each of the seven methods for 80 class pairs (due to space constraints). For each of the 6 baseline methods, we also report the increase in $C^o(\mathcal{D})$ value with respect

Table 5.3: Results obtained from CostStructure1

	c_r	c_p	$\Pr(c_r)$	$\Pr(c_p\|c_r)$	τ_p	τ_r	FA		FM		UR		RR		ALvUS		ALvRS		RM
							$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$
1	M12	M14	3%	1%	3257	1100	26	+13%	227	+865%	29	+22%	34	+45%	30	+28%	33	+41%	23
2	M12	CCAT	3%	5%	1738	1997	49	+36%	227	+533%	58	+63%	60	+68%	65	+82%	59	+65%	36
3	M12	M132	3%	7%	2889	1201	60	+38%	227	+424%	68	+57%	68	+57%	65	+51%	67	+54%	43
4	M12	E21	3%	11%	2048	2063	72	+44%	227	+353%	85	+71%	84	+68%	87	+73%	83	+66%	50
5	M12	M131	3%	18%	2726	1400	180	+30%	227	+64%	192	+39%	189	+36%	196	+41%	177	+29%	139
6	M132	GPOL	3%	1%	2254	1227	30	+25%	229	+859%	33	+39%	38	+59%	34	+44%	36	+54%	24
7	M132	CCAT	3%	2%	1794	2300	41	+26%	229	+596%	52	+58%	55	+66%	50	+51%	54	+66%	33
8	M132	M12	3%	6%	2360	1828	37	+12%	229	+588%	43	+30%	48	+45%	41	+25%	47	+42%	33
9	M132	M131	3%	7%	2506	1685	68	+29%	229	+332%	79	+49%	78	+48%	78	+47%	73	+38%	53
10	M132	GCAT	3%	15%	2258	1152	41	+25%	229	+592%	46	+40%	49	+48%	49	+47%	48	+46%	33
11	M131	CCAT	3%	1%	1141	2797	52	+34%	231	+490%	67	+71%	67	+72%	65	+65%	66	+70%	39
12	M131	M132	3%	6%	1709	1528	63	+27%	231	+365%	78	+56%	72	+44%	65	+30%	69	+40%	50
13	M131	E12	3%	7%	1309	2066	83	+36%	231	+280%	94	+55%	93	+52%	103	+69%	93	+53%	61
14	M131	ECAT	3%	9%	822	3334	95	+61%	231	+291%	111	+88%	112	+90%	112	+88%	108	+83%	59
15	M131	M12	3%	15%	1465	1823	75	+34%	231	+313%	80	+44%	82	+47%	91	+63%	82	+47%	56
16	E12	M11	3%	1%	8371	437	55	+32%	232	+458%	45	+7%	47	+14%	46	+10%	46	+12%	42
17	E12	GDIP	3%	3%	7135	1334	73	+22%	232	+290%	73	+22%	74	+25%	72	+21%	77	+29%	60
18	E12	E212	3%	4%	7135	1336	71	+30%	232	+323%	71	+29%	75	+36%	71	+29%	74	+35%	55
19	E12	M131	3%	7%	7639	1467	87	+35%	232	+261%	92	+42%	96	+49%	102	+58%	93	+45%	64
20	E12	E21	3%	13%	5589	1769	99	+33%	232	+210%	110	+47%	112	+49%	111	+48%	114	+52%	75
21	C21	C17	4%	1%	5862	1101	78	+18%	235	+254%	76	+14%	79	+19%	72	+9%	75	+13%	66
22	C21	C15	4%	3%	4610	1651	84	+11%	235	+211%	88	+16%	90	+19%	85	+13%	87	+15%	75
23	C21	ECAT	4%	5%	3084	2184	93	+10%	235	+180%	104	+24%	104	+24%	95	+13%	102	+23%	84
24	C21	C31	4%	18%	2037	2298	104	+15%	235	+159%	117	+29%	116	+27%	130	+43%	121	+32%	91
25	C21	M141	4%	20%	7052	389	103	+15%	235	+162%	99	+10%	101	+12%	98	+9%	99	+10%	90
26	E212	GPOL	4%	2%	2527	3592	47	+3%	236	+416%	62	+35%	67	+47%	58	+27%	66	+46%	46
27	E212	E12	4%	4%	2357	1410	40	+8%	236	+543%	45	+23%	48	+30%	46	+25%	49	+32%	37
28	E212	M12	4%	8%	2312	1805	70	+31%	236	+342%	78	+47%	78	+47%	85	+60%	80	+52%	53
29	E212	MCAT	4%	9%	2059	3171	73	+23%	236	+297%	90	+51%	91	+53%	95	+59%	89	+50%	59
30	E212	C17	4%	19%	1967	2574	61	+11%	236	+327%	74	+34%	74	+35%	82	+48%	75	+37%	55
31	GCRIM	E212	4%	1%	6001	815	44	+46%	237	+693%	39	+31%	49	+65%	37	+23%	46	+54%	30
32	GCRIM	C15	4%	2%	4533	3118	57	+18%	237	+390%	68	+41%	76	+56%	71	+47%	73	+53%	48
33	GCRIM	C18	4%	3%	4909	2088	48	+24%	237	+513%	53	+37%	61	+58%	50	+29%	59	+52%	39
34	GCRIM	GDIP	4%	6%	3891	2930	80	+40%	237	+316%	91	+59%	96	+69%	86	+52%	96	+69%	57
35	GCRIM	GPOL	4%	20%	2352	4572	105	+42%	237	+221%	124	+68%	129	+74%	128	+74%	129	+74%	74
36	C24	GDIP	4%	1%	9416	1624	77	+27%	240	+294%	67	+11%	71	+17%	62	+1%	66	+9%	61
37	C24	C15	4%	2%	6552	2979	89	+18%	240	+218%	94	+24%	100	+32%	90	+20%	94	+25%	75
38	C24	C31	4%	5%	4318	3803	106	+21%	240	+172%	122	+39%	126	+43%	118	+34%	122	+39%	88
39	C24	MCAT	4%	10%	7329	2090	118	+26%	240	+156%	124	+32%	129	+38%	126	+34%	128	+37%	94
40	C24	C21	4%	20%	3390	4063	142	+39%	240	+136%	159	+56%	154	+51%	204	+100%	192	+88%	102
41	GVIO	C21	4%	1%	4604	4661	63	+2%	242	+291%	77	+25%	88	+42%	63	+2%	74	+20%	62
42	GVIO	C24	4%	1%	6015	2405	63	+24%	242	+374%	66	+29%	75	+48%	65	+27%	65	+27%	51
43	GVIO	CCAT	4%	6%	3490	3540	84	+23%	242	+253%	101	+48%	103	+51%	87	+28%	100	+47%	68
44	GVIO	ECAT	4%	6%	3156	4464	92	+26%	242	+231%	120	+64%	116	+59%	119	+63%	119	+64%	73
45	GVIO	GCRIM	4%	13%	4667	2560	94	+36%	242	+251%	106	+54%	107	+55%	110	+61%	110	+59%	69
46	C13	M12	5%	1%	18998	452	104	+49%	247	+252%	79	+12%	79	+13%	78	+11%	76	+9%	70
47	C13	C15	5%	4%	12039	2243	128	+25%	247	+141%	130	+27%	130	+27%	132	+29%	134	+31%	102
48	C13	GPOL	5%	6%	10068	3383	127	+18%	247	+130%	136	+26%	138	+29%	137	+27%	140	+30%	107
49	C13	M14	5%	7%	16228	1283	116	+27%	247	+170%	105	+15%	108	+18%	105	+15%	108	+18%	91
50	C13	MCAT	5%	14%	11256	2488	135	+20%	247	+118%	147	+30%	150	+32%	152	+34%	148	+31%	113
51	GDIP	C31	5%	1%	5321	5393	94	+27%	249	+238%	112	+52%	123	+66%	100	+35%	115	+56%	74
52	GDIP	E12	5%	2%	6244	4334	82	+19%	249	+261%	95	+38%	106	+53%	87	+25%	97	+41%	69
53	GDIP	CCAT	5%	5%	4060	4562	110	+32%	249	+200%	130	+57%	135	+63%	118	+42%	132	+59%	83
54	GDIP	ECAT	5%	17%	3049	6279	150	+50%	249	+148%	178	+77%	184	+83%	199	+98%	181	+81%	101
55	GDIP	GPOL	5%	19%	3209	5248	182	+37%	249	+88%	202	+53%	198	+50%	228	+72%	214	+62%	133
56	C31	C151	5%	4%	13069	2079	124	+28%	252	+159%	118	+21%	121	+25%	114	+17%	116	+19%	97
57	C31	C15	5%	10%	10168	2824	142	+23%	252	+119%	153	+33%	154	+34%	150	+30%	154	+34%	115
58	C31	ECAT	5%	11%	6230	3660	158	+21%	252	+93%	175	+34%	178	+36%	187	+43%	181	+39%	131
59	C31	C21	5%	12%	6961	3970	164	+14%	252	+75%	203	+41%	196	+36%	239	+66%	218	+51%	144
60	C31	M14	5%	20%	13516	1873	138	+13%	252	+105%	145	+18%	152	+24%	152	+24%	155	+25%	123
61	C181	C151	5%	2%	8194	4348	86	+19%	253	+249%	95	+32%	109	+50%	92	+28%	104	+44%	72
62	C181	GCAT	5%	5%	6513	4458	105	+34%	253	+221%	118	+50%	130	+65%	114	+45%	122	+55%	79
63	C181	C152	5%	10%	5277	6378	137	+42%	253	+160%	159	+64%	172	+78%	171	+76%	172	+77%	97
64	C181	C15	5%	11%	4647	6369	152	+42%	253	+135%	175	+63%	187	+74%	186	+73%	188	+75%	107
65	C181	C17	5%	12%	6612	4805	120	+29%	253	+171%	137	+47%	145	+56%	159	+71%	149	+60%	93
66	M141	ECAT	5%	1%	1054	5162	54	+21%	253	+467%	81	+81%	81	+81%	80	+80%	80	+78%	45
67	M141	GCAT	5%	4%	1258	3819	68	+39%	253	+417%	87	+77%	89	+81%	96	+95%	86	+76%	49
68	M141	C24	5%	5%	1320	4978	80	+44%	253	+359%	102	+84%	98	+79%	102	+85%	104	+89%	55
69	M141	C31	5%	12%	809	6315	129	+87%	253	+268%	151	+119%	148	+115%	166	+141%	164	+139%	69
70	M141	C21	5%	13%	1047	4413	107	+46%	253	+246%	117	+60%	106	+46%	114	+56%	125	+72%	73
71	M11	ECAT	5%	2%	2790	2704	64	+26%	254	+396%	77	+51%	80	+57%	76	+49%	77	+51%	51
72	M11	C152	5%	4%	1797	5573	121	+45%	254	+205%	144	+74%	149	+79%	155	+87%	150	+80%	83
73	M11	M132	5%	5%	3613	2058	68	+43%	254	+438%	74	+58%	81	+71%	66	+40%	79	+67%	47
74	M11	M13	5%	5%	3349	2883	89	+38%	254	+295%	100	+57%	106	+65%	87	+35%	101	+58%	64
75	M11	CCAT	5%	10%	1561	5486	125	+51%	254	+205%	149	+80%	154	+86%	158	+90%	153	+84%	83
76	E21	C31	5%	1%	5196	4511	78	+23%	254	+302%	94	+49%	104	+65%	87	+38%	103	+63%	63
77	E21	M12	5%	5%	6477	2506	89	+29%	254	+265%	95	+36%	103	+48%	108	+56%	107	+54%	70
78	E21	MCAT	5%	7%	4821	4715	106	+24%	254	+195%	131	+52%	134	+56%	133	+55%	132	+54%	86
79	E21	E12	5%	8%	4539	3272	107	+33%	254	+214%	123	+52%	121	+49%	136	+68%	129	+60%	81
80	E21	GPOL	5%	15%	4176	5495	129	+30%	254	+156%	154	+55%	155	+56%	172	+73%	156	+58%	99
				Median values	4598	3145	94	+29%	248	+235%	106	+47%	107	+52%	104	+47%	108	+52%	73

to MINECORE (a positive increase means that the baseline generates higher costs than MINECORE).

Table 5.3 shows the result obtained by using a sample cost structure (here: CostStructure1); $C^o(\mathcal{D})$ denotes the cost incurred by the method while Δ denotes the percentage increase in cost with respect to MINECORE (e.g., +30% means that the cost of the method is 30% higher than that of MINECORE). For readability we indicate costs in thousands of US$, rounding them to the closest unit; e.g., $272,456 would be indicated as 272. MINECORE is here shortened as "RM" (for "Risk Minimization"), Fully Manual is shortened as "FM", Fully Automatic as "FA", Uncertainty Ranking as "UR", Relevance Ranking as "RR", Active learning via Uncertainty and Relevance as "ALvUS" and "ALvRS" respectively. The last row represents median values across the 120 class pairs. The table reveals that for this cost structure (here: CostStructure1), MINECORE is the least expensive of the seven methods for all 120 class pairs. An overall view of the relative merits of the 7 methods can be obtained by looking at the bottom row of the table, which reports median values computed across the 120 class pairs (throughout this chapter we look at medians, rather than at averages, in order to reduce the impact of outliers). In terms of the median values, the 2nd best method is (surprisingly enough) the FA method, which is 29% more expensive than MINECORE. Other methods are even more expensive, up to 235% more than MINECORE; among these other methods one can note a slight advantage of the uncertainty-based methods (UR and ALvUS) over the relevance-based ones (RR and ALvRS), while there seems to be no substantial difference between the methods which are based on active learning (ALvUS and ALvRS) and the ones which are not (UR and RR).

The values of τ_r range in the [809,18998] interval, corresponding to [0.41%,9.53%] of the total set of 199,328 documents; those of τ_p range instead in the [389,7942] interval, corresponding to [0.20%,3.98%] of the total set. This shows two important facts. First, MINECORE sanctions that only a small minority of the documents (max 9.53% of the total for responsiveness, max 3.98% for privilege) should be manually reviewed; this is in line with what e-discovery practitioners expect. Second, MINECORE requires many fewer documents to be manually annotated for privilege than for responsiveness; this is a

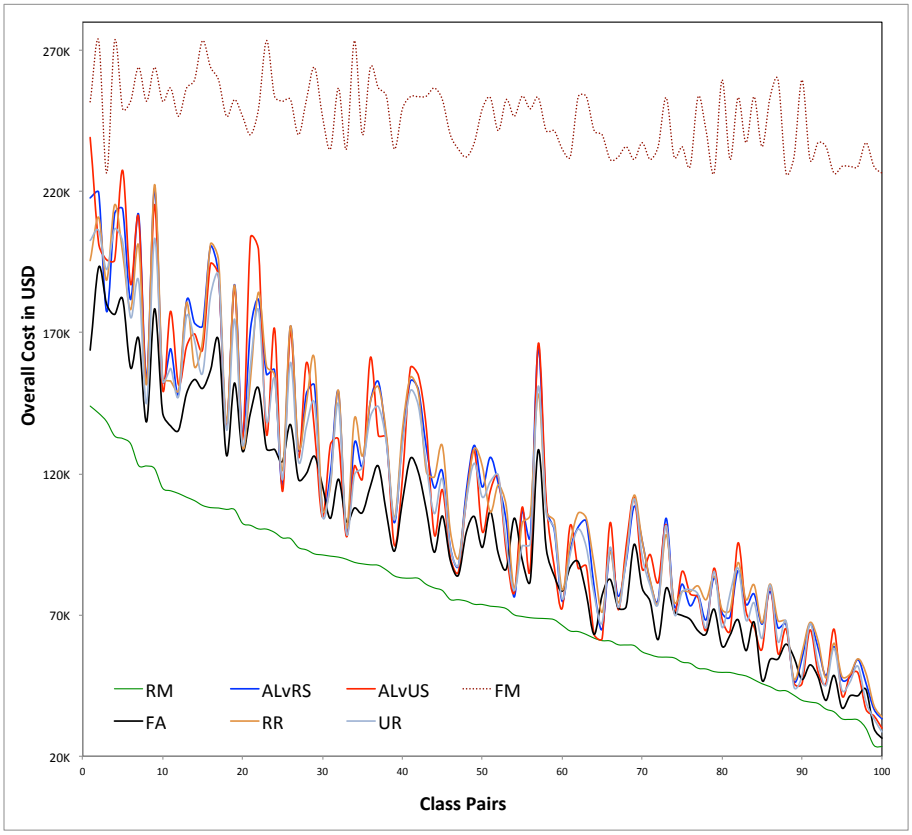

Figure 5.6: Overall costs with CostStructure1 as input

consequence (a) of the fact that many documents are ruled out from further consideration on responsiveness grounds alone, and are not further checked for privilege; and (b) of the fact that manually reviewing for privilege is more expensive, and thus more strongly discouraged by MINECORE, than manually reviewing for responsiveness.

Figure 5.6 shows the overall costs with CostStructure1 for the 7 methods across the 120 class pairs, with the x axis sorted by decreasing cost for MINECORE (here shortened as "RM"). First, the cost of the FM baseline is quite high, varying in a narrow range in a manner that strictly depends on the prevalence of the responsiveness class. Second, none of the baselines other than FM, while all systematically better than FM, are systematically better or systematically worse than all the other ones, which is shown by the fact that the relative plots keep intersecting each other. Third, MINECORE systematically outperforms all others, often by a substantial margin.

Table 5.4 shows the results obtained on a sample class pair (category GPOL as

c_r and category CCAT as c_p) using the different cost structures of Table 5.2. This is a comparison among the results obtained for the different cost structures on a representative class pair.[6]

It is immediately obvious that the cost structure has a lot of influence (i) on how many documents get manually reviewed, both for responsiveness and for privilege, (ii) on the total costs incurred by the various methods, and (iii) on the difference in cost between these methods and MINECORE. In general CostStructure2 results in much smaller numbers of manually reviewed documents than CostStructure1; this is because (see Table 5.2) the misclassification costs are much smaller than in CostStructure1, which makes manual annotation less cost-effective.

CostStructure3 is also an interesting limiting case, in that it results in $\tau_r = \tau_p = 0$; that is, MINECORE decrees that no document is worth manually annotating, and that the decisions of the automatic classifiers should be used, which means that in this case MINECORE coincides with FA. The reason for this behavior lies in the fact that the misclassification costs in Λ_m are (relatively to the annotation costs in Λ_a) very low, too low to justify *any* amount of manual annotation. In general, if the costs in Λ_m are low and the costs in Λ_a are high, low values of τ_r and τ_p (sometimes as low as 0) will result, since manual annotation is discouraged. Conversely, if the costs in Λ_m are high and the costs in Λ_a are low, high values of τ_r and τ_p (sometimes as high as $|\mathcal{D}|$) will result, and MINECORE will suggest manual annotation for all documents in $\mathcal{D}$. In general, the higher (resp., lower) the ratio between the costs in Λ_m and those in Λ_a, the closer to FM (resp., FA) MINECORE is going to be performance-wise. MINECORE is especially advantageous with respect to both baselines when the cost structure justifies the notion that some (but not all) of the documents in $\mathcal{D}$ are worth annotating manually.

Figure 5.7 shows the percentage increase (with respect to MINECORE) in the overall cost $C^o(\mathcal{D})$ resulting from the 6 baseline methods for each of the 120 class pairs according to the 3 different cost structures. Pairs are listed on the x axis by decreasing cost brought

[6] In this example responsiveness is simulated by RCV1-v2 class GPOL ("DomesticPolitics") while privilege is simulated by class CCAT ("Commercial/Industrial"); this class pair was chosen as representative since it is the one for which the median increase in overall cost (+47%) between MINECORE and a high-performing baseline (ALvUS) is obtained.

Table 5.4: Results obtained GPOL(as R)-CCAT(as P) class pair

	τ_p	τ_r	FA		FM		UR		RR		ALvUS		ALvRS		RM
			$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$
CostStructure1	6169	6885	177	+32%	273	+105%	207	+55%	215	+61%	196	+47%	212	+59%	**93**
CostStructure2	918	1189	57	+3%	273	+397%	63	+14%	64	+16%	57	+3%	63	+14%	**55**
CostStructure3	0	0	15	+0%	273	+1714%	15	+0%	15	+0%	15	+0%	15	+0%	15

Table 5.5: Results from all cost structures

	FA		FM		UR		RR		ALvUS		ALvRS		RM
	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$	Δ	$C^o(\mathcal{D})$
CostStructure1	94	+29%†	248	+235%†	106	+47%†	107	+52%†	104	+47%†	108	+52%†	**73**
CostStructure2	24	+2%†	248	+893%†	26	+10%†	26	+11%†	25	+4%†	25	+7%†	**24**
CostStructure3	10	+0%	248	+2416%	10	+0%	10	+0%	10	+0%	10	+0%	10

about by MINECORE. For better comparison all figures are displayed across the range [-15%,+145%] on the y axis. In the FM figure (top right) this makes the CostStructure2 and CostStructure3 curves, and most of the CostStructure1 curve, plot above the ceiling. It extends the comparison shown in Table 5.4 to the full set of class pairs. As can be seen, all of the baselines generally incur substantially higher costs than MINECORE with CostStructure1; this difference is instead far smaller for CostStructure2 (as noted above, there is no difference between MINECORE and the other baselines – except FM – for CostStructure3).

Finally, Table 5.5 shows the median (across the 120 class pairs) overall cost obtained by each method with each cost structure. This table reveals the results obtained by using the different cost structures of Table 5.2. The results in a given row are the median of the 120 results obtained with the tested 120 class pairs. **Boldface** indicates the best method, while † indicates a statistically significant ($p < 0.01$) increase in overall cost with respect to MINECORE, as determined by the Wilcoxon test.

For CostStructure2, MINECORE does better by this median measure than all of the baseline methods by smaller margins than are achieved for CostStructure1. For both of those two cost structures, the costs generated by each baseline method is statistically significantly higher according to a Wilcoxon signed rank test for paired samples over the 120 class pairs, at $p < 0.01$. Concerning CostStructure3, similarly to what happened for the pair showcased in Table 5.4, MINECORE evaluates both τ_r and τ_p to 0 for all class pairs, making MINECORE and all the other methods (aside from FM) coincide with FA.

Incidentally, one cannot help noticing how the FM fully manual baseline is, by a very wide margin and according to all three cost structures, the worst of all systems. This is a further confirmation of a fact first noted in [36], which reasserts that technology-assisted review is nowadays unavoidable in e-discovery.

A first thing to observe is that, in MINECORE, a document can end up being manually annotated only for responsiveness, only for privilege, for both responsiveness and privilege, or for neither responsiveness nor privilege.

A second thing to observe is that Phases 2 and 3 are structurally identical, since Phase 2 does for responsiveness what Phase 3 does for privilege. One might thus wonder if we could switch their order without negatively impacting (or perhaps even positively impacting) $C^o(\mathcal{D})$. The answer is no, and the reason lies in the fact that, in typical e-discovery scenarios, λ_p^a is higher or much higher than λ_r^a (we indeed imposed the constraint that $\lambda_r^a < \lambda_p^a$). This has the consequence that it makes sense to employ the expensive (as characterised by λ_p^a) senior reviewers for annotating documents that the cheap (as characterised by λ_r^a) junior reviewers have "pre-filtered".

A third observation which is in order is about ranking. During Phase 2 MINECORE clearly separates the set (let us call it $\mathcal{D}_2^{man}$) of the τ_r documents that should be annotated from the set (let us call it $\mathcal{D}_2^{aut}$) of the $(|\mathcal{D}| - \tau_r)$ documents that should not be annotated (the same happens at the end of Phase 3). If the human reviewer annotates all and only the former, one might wonder why is ranking useful at all. While ranking is indeed unnecessary in theory, it is useful in practice, for two reasons:

- The choice of which documents to put in $\mathcal{D}_2^{man}$ and which to put in $\mathcal{D}_2^{aut}$ is far from perfect, since it relies on automatically generated posterior probabilities. As a result, the human reviewer might find out, at the very moment s/he is invited to stop annotating, that s/he was still finding many mislabeled documents, and s/he might thus want to annotate some more documents in order to be on the safe side;

- If, for some reason, the reviewer stops annotating before the stopping condition is reached, the fact that s/he has annotated by following the ranked list guarantees that the cost-effectiveness of her work has been maximized.

As a result, we indeed assume that rankings are generated (and followed by the human reviewers) in both Phase 2 and Phase 3.

5.8 Summary

During e-discovery, the party performing the review may incur costs of two types. Annotation costs (deriving from the fact that human reviewers need to be paid for their work) and misclassi cation costs (deriving from the fact that failing to correctly determine the responsiveness or privilege of a document may adversely affect the interests of the parties in various ways). Relying exclusively on automatic classification would minimize annotation costs but could result in substantial misclassification costs, while relying exclusively on manual classification could generate the opposite consequences.

Thus, we develop a risk minimization framework called MINECORE, that seeks to strike an optimal balance between these two. In our MINECORE model the documents are first automatically classified for both responsiveness and privilege. In the next step, some of the automatically classified documents are annotated by human reviewers for responsiveness (typically by junior reviewers) and for privilege (typically by senior reviewers), with the overall goal of minimizing the expected cost (i.e., the risk) of the entire process.

Risk minimization is achieved by optimizing, for both responsiveness and privilege, the choice of which documents to manually review. We present a simulation study in which categories from a standard text classification test collection (RCV1-v2) are used to mimic responsiveness and privilege topic. Our findings indicate that MINECORE can yield substantially a lower total cost than any of a set of strong baselines we propose.

In our work, we have assumed that lawyers will be able to conceptualize unit annotation costs and unit misclassification costs in comparable units. Although this has proven to be a useful, one important insight from our experience is that people often find it difficult to quantify uncertain costs using the same units in which they would express costs that would be incurred. We have assumed for the purposes of our work that some model of costs and risks exists and can be formalized, but in practice the process of designing such models may not be as simple as asking an attorney to assign values to the elements

in one of our cost structures. We have also assumed that both costs and risks accumulate linearly. We are confident that our framework will give lawyers more to discuss, since adopting our approach would mean that they would ultimately need to agree on both the cost structure and the way in which error probabilities are estimated.

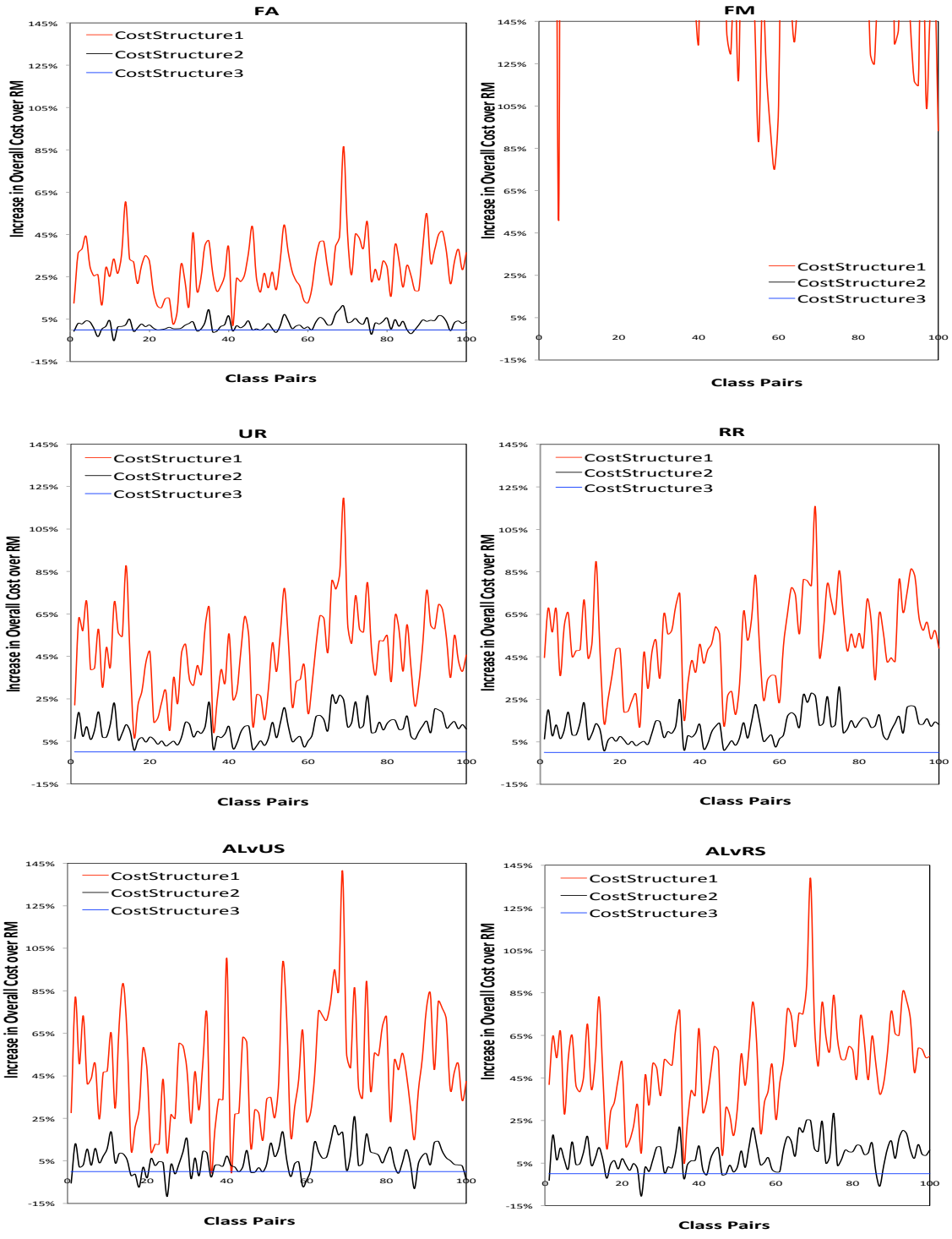

Figure 5.7: Percentage increase in the overall cost

Chapter 6: Conclusions

E-discovery practices are indeed well suited for the interplay between the humans and the computerized models to identify which documents in a collection are responsive to a production request, and to identify the documents that should be withheld on the basis of privilege. This dissertation can help to inform the legal community that the adoption of predictive coding technique is actually a good option in some litigation cases. Our research aims to provide multiple contributions to the current e-discovery practice. It provides multiple proofs of concept to encourage affordable e-discovery procedures by isolating and studying its key components.

Research in e-discovery has been hampered by the lack of publicly available test collections. The only test collection that is publicly accessible for e-discovery privilege classification was created during the TREC 2010 Legal Track. Before we start to design a classifier, as a first step, we ask *Is it possible to perform an unbiased classifier evaluation?* We start by answering this question because designing any system without having an evaluation plan does not provide any discernible value. Hence we first build and release a useful set of documents to enable unbiased classifier evaluation. To create a labeled unbiased set for evaluation purpose, we remove the selection bias (introduced by the sampling process during TREC 2010) by re-sampling from the biased document categories. We maintain the sampling probability of our re-sampling process to be approximately the same as the rate used during the creation of the test collection. This process resulted in a total of 252 documents as our held-out test set with senior assessor's judgments making it the Gold standard for evaluation.

We ask if the TREC 2010 Legal Track test collection is reliable and reusable? Chapter 3 explain the issues in the context of TREC Legal Track 2010 test collection creation

process. Since pooling was widely adopted, we identify two types of errors in the collection; sampling errors and measurement errors. One way of understanding the measurement errors is by studying the classifier's sensitivity to assessor errors. To do so, we utilize a subset of document families and also the entire set of documents families that were selected for re-assessment by senior attorneys as test sets. We focus separately on estimates of recall and precision. The recall and precision values derived are point estimates, and are subject to random variation. Hence we also provide an indication of the expected range of variability around a point estimate, and account for it when comparing the two scores. We compute 95% confidence interval to identify the range within which the point estimates lies in the entire population. We plot the point estimates and the confidence intervals using the judgments from senior assessors. As the senior assessors' judgments sample is less than 8% of the size of the full set of official judgments, our results yields fairly large confidence intervals, but the comparison does offer useful insights. A standard way of performing analyses to assess the samples is through system ablation study. We removed the results from a system that participated in the stratification process and then re-score all systems, including the ablated system, then observe the effect on system comparisons. Comparing the post-ablation to pre-ablation results, we see that confidence intervals for precision increase for the ablated system which could be attributed to the difference in sampling probabilities of the strata. We conclude that assessor errors do adversely affect absolute estimates of recall, and we have suggested future work on statistical correction for the effects of those errors. For the task of identifying privileged documents, it is known that the recall measure is more important. Thus, this initial result is promising.

Now that we have an evaluation plan for a classifier, as our next step we proceed to design a classifier to identify privileged documents. We build multiple binary classifiers utilizing the email content and metadata features. We further investigate the extent to which the remaining privilege judgments in the TREC Legal Track 2010 test collection obtained by the human reviewers are useful for training. As the difference in reviewer's expertise adversely affect the absolutely estimates in recall, our research questions *RQ3a and RQ3b* aim to analyze the influence of annotator expertise and sample selection bias, on classifier training. For studying the effect of training the classifier on different sets of

judgments depending on the annotators' expertise, we develop two classifiers; one with judgments from junior level annotator as training set and the other one with judgments from senior assessors as training set. We then compare the classifier performance for recall and precision values with 95% confidence intervals. We observe a significant increase in the recall measure of the classifier trained on document set with senior assessors' judgments. The problem of selection bias exists in the TREC Legal Track 2010 collection because of the fact that the participating teams were allowed to challenge the judgments of the junior annotator (for details refer Chapter 2) leading to some chosen sample to be reviewed by senior assessor. To study the effect on the bias caused by that chosen sample, we again build two classifiers; one with those documents that were not chosen for adjudication as training set and the other one that were chosen for adjudication as training set. By comparing the classifier results, we conclude that training classifiers on documents that are not chosen for adjudication yields good result. We explain the findings above by collectively analyzing the classifiers' privilege predictions on the unbiased test set.

After completing the task of building a binary classifier for identifying privileged documents, we reached out to some lawyers to understand how to present the results from a system. We wanted to learn which features helps them to perform the privilege review. We ask the research questions outlined in Chapter 1, section 1.2 as *RQ4a, RQ4b and RQ4c*. Our aim was to get an understanding about how best to present the results from the classifier. As our first step, we thought to highlight the actors in the email communication. We presented three types of metadata information to the lawyers doing the review; actor privilege importance score which we call as propensity, actor's organization information and actor's role information in that organization. We developed an algorithm to score the importance of specific email addresses with the goal to determine their propensity to engage in privileged communication. Both recursive and heuristic techniques are used to estimate the propensity score, ultimately resulting a coverage of 94% of the email addresses. Since litigation is time-sensitive, we provided a graphical display of privileged communication temporal patterns. The last type of information from the automation process that was presented to the lawyers during review was the importance of the term to identify privilege. Only the top 10% of the important terms were highlighted to avoid

clutter.

We categorized the findings from *RQ4a, RQ4b and RQ4c* as quantitative and qualitative results. The results to measure the accuracy and speed are quantitative. The qualitative results are from our interview and from our usability questionnaire. To draw some conclusions about the accuracy of privilege review process, we first select an informative set of judgments as benchmark against which review accuracy can be measured. From our analysis, either senior attorney could reasonably be chosen as a benchmark against which the other participant's judgments could be measured for accuracy. Using one of the senior attorney's judgments as benchmark, we conclude that there is a statistically significant improvement in recall. This improvement was noticed across all users except one (who was a novice user). This is a promising result. However, when we measure the our system performance for speed, a paired t-test found no detectable difference in average review speed across the two conditions. This could be attributed to our thinking that lawyers were more careful in reviewing a document when more information was provided to them. During this study, we also evaluated our research prototype for its usability. Our usability questionnaire assigned a higher rating to the overall review experience. Person highlighting feature of the system was reported to be useful (to at least some degree) by five of the six participants, whereas term highlighting and the date graph were each reported to be useful to some degree by only two of the six participants.

By performing a user-study with the lawyers, we acquired some important pieces of information. One of the main lessons we learn was that the users were open to adopt predictive coding techniques that help them perform the privilege review. The second conclusion we drew was that, there was no measurable change in the review time. Since time is proportional to money during privilege review, the final questions we answer in our dissertation are about the overall costs incident upon the entire e-discovery process.

As one answer to the questions we raise in *RQ5a and RQ5b*, we develop a risk-based minimization framework. This framework is based on utility theory and relies on cost-sensitive ranking. We formalize our problem on the basis that costs and risks exists and can be characterized. Additionally, that misclassification costs do not exist in isolation (e.g., privilege only), but rather at a two-stage level (i.e., responsiveness and privilege).

Hence the two stages are best addressed jointly. Our semi-automated system assumes that a document might be produced to the requesting party even if it has not been manually reviewed to be responsive and nonprivileged. When deciding which document should be manually reviewed, we use our ranking algorithm to determine which document is expected to bring about the smallest cost when produced. Manual annotation time and effort is sparingly utilized to bring about a reduction in the number of documents to be reviewed for responsiveness and privilege. A threshold based stopping criteria is used to indicate when the reviewers should stop annotating.

We gathered inputs for our cost structure from three e-discovery experts. We develop an algorithm that utilizes the classifier results and the cost structure to determine which document needs a manual review. Then, we ask our final question; will our risk-minimization framework help us save some money for any given litigation. Chapter 5 discusses the methodology, experimentation and the results of our risk-minimization framework by introducing a new evaluation measure.

Our conclusions are supported by experimentation. For experiments we need a collection that has judgments for two classes (responsive class and privilege class). We need labeled documents that are; (1) responsive and privileged, (2) responsive and not-privileged, (3) not-responsive and privileged and (4) not-responsive and not-privileged. We overcome the problem of the lack of a such a test collection by running simulated experiments on a extensively labeled collection unrelated to e-discovery in which documents can belong to more than one class.

We build two binary classifiers, utilize their posterior probabilities with the values from the cost structure to determine which document needs a manual annotation. We propose multiple effective baseline methods for comparison. Some of our baseline methods are human-in-the-loop systems, i.e., their predictions are obtained via some combination of manual annotation work and automatic classification. We run our simulations by picking 120 pairs of classes to play the role of responsive class and privileged class. We obtain the results for seven different methods for each of the 120 pairs of classes. Our models were tested on a collection of nearly 200,000 documents with three different cost structures as inputs.

From our findings it is clear that cost structure has a lot of influence (1) on how many documents get manually reviewed, both for responsiveness and for privilege, (2) on the total costs incurred by the various methods, and (3) on the difference in cost between the baseline methods and our semi-automated system. Our results show that all of the baselines generally incur substantially higher costs when compared to our model. The empirical evidence with statistical significance tests show that our semi-automated process systematically achieves a reduction in the overall cost of the e-discovery process for two out of the three litigation cases.

6.1 Contributions

This dissertation work shows a positive synergy between the lawyers and machines. Although this research work is specific to the domain of e-discovery, the contribution below could be applied to any domain where the relevant content is intermixed with sensitive information (like personal and organization emails, medical records, government records,etc.). The work done in this dissertation can be divided into three categorical contributions; System contributions (S), Practical contributions (P) and Research contributions (R).

In addition to the the research question and answers, the *System* contributions highlight the frameworks built with the view to enable other researchers to replicate and continue the work we started.The *Practical* contributions highlight the value of this dissertation work in the e-discovery industry and the *Research* contributions highlight the domain-specific research advances.

The contribution of this thesis includes:

6.1.1 System Contributions

- *S1* - Release of 252 unbiased families[1] from the TREC Legal Track 2010 collection with domain-expert annotations for privilege that could be use as a held-out test set and for evaluation. (Chapter 3, section 3.2)

[1] A family is an email message along with its attachments.

- *S2* - Development of a multiple binary classifiers for predicting families which have privileged content. (Chapter 3, section 3.3)
- *S3* - Development of an algorithm to score the importance of people (in privileged context) in email communications. (Chapter 4, section 4.2.1)
- *S4* - Development of a methodology to compute term importance utilizing word entropy. (Chapter 4, section 4.2.4)

6.1.2 Practical Contributions

- *P1* - Development of a research prototype to enable lawyers to perform privilege review. (Chapter 4, section 4.2.6)
- *P2* - Release the code for a review application to enable lawyers to quantify the e-discovery outcome errors in terms of US Dollars.

6.1.3 Research Contributions

- *R1* - Representing e-discovery outcomes as a ternary classification problem. (Chapter 5, section 5.1)
- *R2* - Introducing the idea of quantifying the different kinds of erroneous e-discovery outcomes in terms of US Dollars. (Chapter 5, section 5.1 and section 5.6.3)
- *R3* - Developing a semi-automated process with risk-based ranking algorithm to determine which document deserves to be reviewed by a human. (Chapter 5, section 5.4 and section 5.4.1)

6.2 Limitations

A number of important points should be kept in mind when interpreting the experiments and results reported in this dissertation. In particular, we would like to highlight the following limitations:

1. We develop a classifier to predict privileged documents. We utilize the test collection created during the TREC 2010 Legal Track to train and evaluate our classifier. During evaluation we use the senior annotator judgments on the 252 families in the test set as gold standard. These 252 labeled families were a result of our re-sampling procedure explained in chapter 3, section 3.2. We were limited to a total of 252 families due to the lack of randomly chosen families that had been judged by the senior assessor.

2. The TREC Legal Track 2010 collection lacked positive training examples especially those that are labeled by the senior attorneys. Our classifier was trained on a limited number of positive labeled examples.

3. This work takes the first step to understand the users' needs by building an interactive user-interface to perform user study. We recruited users who had a law degree due to the nature of the task. In our user study explained in chapter 4, we were limited to only six users who were willing to participate in our study.

4. In our user study discussed in chapter 4, we were limited to only 61 labeled families while evaluating each user's accuracy because only 61 families were reviewed by all the participants.

5. Our work in chapter 5 assumes that human reviewers do not make mistakes, i.e., the judgment of our human reviewers always coincides with the ground truth.

6. In chapter 5, the evaluation metric used to measure the overall cost, is assumed to be a linear function.

7. Experiments in chapter 5 use a test-collection which is not topically related to e-discovery. This is due to the lack of publicly available test collection of documents that are annotated by both responsiveness and privilege.

8. In chapter 5, we quantify classifier errors in terms of cost value in US Dollars. We represent the misclassification cost values as a 3 by 3 contingency matrix with 6 non-zero positive values. In our work, we limit the structure of the input cost matrix to

a 3 by 3 dimension.

9. Experiments in chapter 5 were limited to only three input cost matrices.

10. We limited our experiments to category pairs such that the test set prevalence of responsive documents between 3% - 7% and the prevalence of privileged documents in the responsive set is between 1% - 20%.

6.3 Future Work

Our experience working on this thesis also suggests several directions as future work.

1. Our initial efforts in this dissertation focused on building a binary classifier to classify for privileged email communications. More experiments need to be conducted to improve the accuracy of our classifier. In our work we build a classifier with an acceptable recall measure. We stress on the fact that privilege review is a recall problem. As a part of future work, the first thing that we suggest is to improve the overall accuracy of privilege classifiers. We also suggest the use of sophisticated features to build the classifier. We suggest employing neural network architecture that, given a sentence, outputs a host of language processing predictions; such as; part-of-speech tags, chunks, named entity tags, semantic roles, semantically similar words and the likelihood that the sentence makes sense (grammatically and semantically) using a language model. These kind of features that exploit the language could have high potential in predicting privileged communications especially between the attorney and the client.

2. Our privilege review platform was designed thinking about the problem that began with the idea that modeling attributes of people such as their roles and their propensity to engage in privileged communication might be particularly important for the privilege review task, and our results provide support for that belief. Our results also indicate that dates, while important for relevance review, may be of less value for privilege review (at least in the way we are doing things now). We have noted that the increases in recall that we observed were often accompanied by substantial

declines in precision. A further study will be needed to better characterize this effect and to control for it in future experiments.

3. During our user study, we noted no evidence of improvements in review speed, although of course even our most expert participants were novice users of the particular interface that we presented them with. In future work we may therefore consider longitudinal studies that would allow us to see how the same users behave at different points in their personal learning curve.

4. Our experience after the user study suggests to run small-scale studies to tune specific components. As examples, we could ask; what types of multi-word expressions should be considered for highlighting? how many terms or multi-word expressions should be highlighted? how many categories of term highlighting are useful? Studies along those lines might ultimately lead to test collections that could be used as a basis for tuning and evaluating specific system components; for that we will also need to give thought to intrinsic measures for evaluating the performance of individual components.

5. Another productive research direction would be to explore whether we might productively replace expensive attorneys in some early studies. Would utilizing law students be suitable? Law librarians? Crowd-sourcing services such as Mechanical Turk? Surely we can go some distance in this direction; the key question is how far can we productively go without compromising the accuracy.

6. Our classifiers use SVM as the learning algorithm. As a part of future work we suggest researchers to extend to use other types of learning algorithms like; Logistic Regression, Transductive SVMs, Random Forest, Gradient Boosting Machines, Neural Networks etc. in place of the standard SVMs.

7. We propose to model the cost function as a nonlinear cost functions in place of the unit costs we currently use.

8. Our risk-minimization work in chapter 5 assumes that manual reviewers do not make mistakes, i.e., the judgment of our human reviewers always coincides with the ground

truth. In future, we suggest experiments that would study the effect of reviewer's errors.

9. Finally, we should note that this work could be extended in other settings where search amidst sensitive content is needed.

6.4 Implications

Today in e-discovery, automation for relevance review has been a topic of discussion. The decision of whether predictive coding can be employed during production is a choice. Attorneys owe it to their clients to become familiar with this newer technology and to consider whether it should be used. It is likely that predictive coding will become more widely used in the near future as parties gain confidence in its accuracy and as we show some preliminary evidence that it truly reduces costs at least in some litigation cases. As the technology-assisted review tools are deployed and adopted, it is natural to expect larger cases to be tackled. With an increase in the number of relevant documents in the collection, automation of privilege review is going to be one predictable consequence. It therefore seems timely to begin to think seriously about how and to what extent use of predictive coding systems could help the e-discovery process.

As the volume of digital information grows every year, the need to adopt automation becomes more and more urgent. The answer to the question *how can the costs of manual review be controlled?* has become a commonplace.

The most promising alternative available today for collections with high prevalence resulting in large-scale manual review process is the use of predictive coding and other computerized categorization strategies that can rank electronic documents by using an algorithm that determine which document is, responsive, and/or privileged. Manual review is still required during production. Empirical research suggests that predictive coding is at least as accurate as humans in traditional review. Additionally, there is evidence that significant number of manual review hours could be reduced depending on the nature of the documents and other factors, which would make predictive coding one answer to the critical need of significantly reducing review costs. It is certainly not the sole answer, and

any cost savings may be negligible unless litigants first take a holistic approach. But, assuming that best practices have been followed throughout the e-discovery life cycle, these new techniques presented may be one practical approach.

Our conclusions about one way to reduce the overall production expenditures are shaped by the topic prevalence, algorithm and cost structures we included in our analysis. Tasks involving pre-processing of the collection could present a greater cost burden for the producing parties when volumes of digital data are huge. Conversely, computer applications for conducting review are unlikely to be economically viable options when dealing with smaller document sets, in which any savings in attorney hours might be overshadowed by machine-training costs. Our attempt is thus to encourage the legal community to make the choice that is the best option for the litigation. Our hope is that the work in this dissertation will help inform the e-discovery community about how to adapt the review practices to address concerns about the costs of production.

Appendices

.1 Appendix A

IRB Approval Letter

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	August 4, 2015
TO:	Douglas Oard
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[784030-1] Development and Evaluation of Search Technology for Discovery of Evidence in Civil Litigation
REFERENCE #:	
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	August 4, 2015
EXPIRATION DATE:	August 3, 2016
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category # 7

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a project design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

Prior to submission to the IRB Office, this project received scientific review from the departmental IRB Liaison.

This submission has received Expedited Review based on the applicable federal regulations.

This project has been determined to be a Minimal Risk project. Based on the risks, this project requires continuing review by this committee on an annual basis. Please use the appropriate forms for this procedure. Your documentation for continuing review must be received with sufficient time for review and continued approval before the expiration date of August 3, 2016.

Please remember that informed consent is a process beginning with a description of the project and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Unless a consent waiver or alteration has been approved, Federal regulations require that each participant receives a copy of the consent document.

Please note that any revision to previously approved materials must be approved by this committee prior to initiation. Please use the appropriate revision forms for this procedure.

All UNANTICIPATED PROBLEMS involving risks to subjects or others (UPIRSOs) and SERIOUS and UNEXPECTED adverse events must be reported promptly to this office. Please use the appropriate reporting forms for this procedure. All FDA and sponsor reporting requirements should also be followed.

All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to this office.

Please note that all research records must be retained for a minimum of seven years after the completion of the project.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Maryland College Park (UMCP) IRB's records.

.2 Appendix B

Questionnaire for Gathering Cost Values

Questionnaire on Annotation Costs and Misclassification Costs in e-Discovery

Douglas W. Oard, Fabrizio Sebastiani, Jyothi K. Vinjumur

Premises:

- The questionnaire contains 8 questions. You should answer them all, since partially filled questionnaires will be much less useful to us.
- Each question except the first has 3 possible answers. Pick your choice by ticking one (and only one) of the 3 tick boxes. If your answer is either the 1st or the 2nd, you should also fill the additional box with a number higher than 1.
- You may make any assumption about the type of case and the amounts at stake in the case, but make the same assumptions for every question.
- **We heartily thank you for your effort; your contribution is of critical importance to our research in automating the e-discovery process.**

Assumptions:

- Documents that are responsive and nonprivileged are produced (**P**) to the requesting party;
- Documents that are responsive and privileged are reported on the privilege log (**L**) and not produced;
- Documents that are nonresponsive are withheld (**W**) by the producing party (i.e., they are not produced);
- "Mistake X is Z times more serious than mistake Y" can be interpreted as "The overall cost that the producing party incurs by making many mistakes of type X is Z times higher than the cost the same party would incur by making as many mistakes of type Y".

Question # 1 *Which of the following best describes your background:*

- ☐ *Senior attorney who has supervised e-discovery reviews*
- ☐ *Attorney who has participated in e-discovery reviews as a reviewer*
- ☐ *Other highly knowledgable e-discovery expert*
- ☐ *Other attorney*
- ☐ *Other (please describe):* _____

Question # 2 *Consider two types of mistakes:*

LP *Situation:* *Document is responsive and nonprivileged (it should thus be produced)*
Mistake: *Document is erroneously reported on the privilege log and not produced*

PL *Situation:* *Document is responsive and privileged (it should thus be reported on the privilege log and not produced)*
Mistake: *Document is erroneously produced*

Is mistake LP more serious than mistake PL?

- ☐ *Yes, mistake LP is* ☐ *times more serious than mistake PL.*
- ☐ *No, mistake PL is* ☐ *times more serious than mistake LP.*
- ☐ *They are equally serious.*

Question # 3 *Consider two types of mistakes:*

LW *Situation:* *Document is nonresponsive (it should thus be withheld)*
Mistake: *Document is erroneously reported on the privilege log (and not produced)*

WL *Situation:* *Document is responsive and privileged (it should thus be reported on the privilege log and not produced)*
Mistake: *Document is erroneously deemed nonresponsive (and thus withheld)*

Is mistake LW more serious than mistake WL?

- ☐ *Yes, mistake LW is* ☐ *times more serious than mistake WL.*
- ☐ *No, mistake WL is* ☐ *times more serious than mistake LW.*

☐ *They are equally serious.*

Question # 4 *Consider two types of mistakes:*

WP *Situation: Document is responsive and nonprivileged (it should thus be produced)*
Mistake: Document is erroneously considered nonresponsive (and thus withheld)

PW *Situation: Document is nonresponsive (it should thus be withheld)*
Mistake: Document is erroneously produced

Is mistake WP more serious than mistake PW?

☐ *Yes, mistake WP is* ☐ *times more serious than mistake PW.*

☐ *No, mistake PW is* ☐ *times more serious than mistake WP.*

☐ *They are equally serious.*

Question # 5 *Consider two types of mistakes:*

LP *Situation: Document is responsive and nonprivileged (it should thus be produced)*
Mistake: Document is erroneously reported on the privilege log and not produced

LW *Situation: Document is nonresponsive (it should thus be withheld)*
Mistake: Document is erroneously reported on the privilege log (and not produced)

Is mistake LP more serious than mistake LW?

☐ *Yes, mistake LP is* ☐ *times more serious than mistake LW.*

☐ *No, mistake LW is* ☐ *times more serious than mistake LP.*

☐ *They are equally serious.*

Question # 6 *Consider two types of mistakes:*

LW *Situation: Document is nonresponsive (it should thus be withheld)*
Mistake: Document is erroneously reported on the privilege log (and not produced)

WP *Situation: Document is responsive and nonprivileged (it should thus be produced)*
Mistake: Document is erroneously considered nonresponsive (and thus withheld)

Is mistake LW more serious than mistake WP?

☐ *Yes, mistake LW is* ☐ *times more serious than mistake WP.*

☐ *No, mistake WP is* ☐ *times more serious than mistake LW.*

☐ *They are equally serious.*

Question # 7 *Consider the following type of mistake:*

WP *Situation: Document is responsive and nonprivileged (it should thus be produced)*
Mistake: Document is erroneously considered nonresponsive (and thus withheld)

Is the cost of annotating a document for responsiveness higher than the cost brought about by a mistake of type WP?

☐ *Yes, the cost of annotating a document for responsiveness is* ☐ *times higher than the cost brought about by a mistake of type WP.*

☐ *No, the cost brought about by a mistake of type WP is* ☐ *times higher than the cost of annotating a document for responsiveness.*

☐ *The two costs are equal.*

Question # 8 *Is the cost of annotating a document for responsiveness higher than the cost of annotating a document for privilege?*

☐ *Yes, the cost of annotating a document for responsiveness is* ☐ *times higher than the cost of annotating a document for privilege.*

☐ *No, the cost of annotating a document for privilege is* ☐ *times higher than the cost of annotating a document for responsiveness.*

☐ *The two costs are equal.*

If you wish you may add your name and contact (and possible additional comments) here:

Name : _____

Contact : _____

Comments : _____

_____

_____

_____

Bibliography

[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] Armstrong v. Bush, 1989.

[4] In re Sealed Case. 877(No. 89-5102):976, 1989.

[5] Kleen Products, LLC v. Packaging Corporation of America, 2011.

[6] Global Aerospace Inc. v. Landow Aviation, LP, 2012.

[7] Moore v. Publicis Groupe SA, 2012.

[8] EORHB, inc. v. HOA Holdings LLC, 2013.

[9] Alan Agresti and Brent A Coull. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, 1998.

[10] Ron Artstein and Massimo Poesio. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[11] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[12] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. Relevance Assessment: Are Judges Exchangeable and Does it Matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674. ACM, 2008.

[13] Jason R Baron. The TREC Legal Track: Origins and Reflections on the First Year. In *Sedona Conference*, volume 8, pages 251–253, 2007.

[14] Jason 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.

[15] Jason R Baron, David D Lewis, and Douglas W Oard. TREC 2006 Legal Track Overview. In *TREC*, 2006.

[16] Giacomo Berardi, Andrea Esuli, and Fabrizio Sebastiani. Utility-theoretic ranking for semi-automated text classification. *ACM Transactions on Knowledge Discovery from Data*, 10(1):Article 6, 2015.

[17] Alina Beygelzimer, Varsha Dani, Tom Hayes, John Langford, and Bianca Zadrozny. Error Limiting Reductions between Classification Tasks. In *Proceedings of the 22nd International Conference on Machine learning*, pages 49–56. ACM, 2005.

[18] Paul E Black. Ratcliff/Obershelp pattern recognition. *Dictionary of Algorithms and Data Structures*, 17, 2004.

[19] David C Blair and Melvin E Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299, 1985.

[20] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 619–620. ACM, 2006.

[21] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information retrieval*, 10(6):491–508, 2007.

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

[23] Ben Carterette, Virgil Pavlu, Evangelos Kanoulas, Javed A Aslam, and James Allan. Evaluation Over Thousands of Queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 651–658. ACM, 2008.

[24] Ben Carterette and Ian Soboroff. The effect of assessor error on ir system evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–546. ACM, 2010.

[25] Jianlin Cheng, Amanda Jones, Caroline Privault, and Jean-Michel Renders. Soft Labeling for Multi-pass Document Review. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law, DESI V Workshop*, 2013.

[26] Gordon V Cormack and Maura R Grossman. Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 153–162. ACM, 2014.

[27] Gordon V. Cormack and Maura R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. CoRR abs/1504.06868, 2015.

[28] Gordon V. Cormack and Maura R. Grossman. Multi-faceted recall of continuous active learning for technology-assisted review. In *Proceedings of the 38th ACM Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 763–766, Santiago, CL, 2015.

[29] Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. Overview of the TREC 2010 legal track. In *Proc. 19th Text REtrieval Conference*, page 1, 2010.

[30] Gordon V. Cormack and Mona Mojdeh. Machine learning for information retrieval: TREC 2009 web, relevance feedback and legal tracks. In *Proceedings of the 18th Text Retrieval Conference (TREC 2009)*, Gaithersburg, US, 2009.

[31] Yanlei Diao, Hongjun Lu, and Dekai Wu. A comparative study of classification based personal e-mail filtering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 408–419. Springer, 2000.

[32] Pedro Domingos. Metacost: A General Method for Making Classifiers Cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 155–164. ACM, 1999.

[33] Edna Selan Epstein. *The Attorney-Client Privilege and the Work-Product Doctrine*. ABA 2001.

[34] Gregory 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.*, 15:1, 2008.

[35] Manfred Gabriel, Chris Paskach, and David Sharpe. The challenge and promise of predictive coding for privilege. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law, DESI V Workshop*, 2013.

[36] Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3):Article 5, 2011.

[37] Bruce Hedin, Stephen Tomlinson, Jason R Baron, and Douglas W Oard. Overview of the trec 2009 legal track. Technical report, National Archives And Records Administration, 2009.

[38] Thorsten Joachims. Making large-scale SVM learning practical. In Bernhard Schölkopf, Christopher J. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, chapter 11, pages 169–184. The MIT Press, Cambridge, US, 1999.

[39] Thorsten Joachims. Estimating the generalization performance of a SVM efficiently. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 431–438, Stanford, US, 2000.

[40] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI 2007)*, pages 877–882, San Francisco, US, 2007.

[41] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 2003.

[42] Anne Kershaw. Automated document review proves its reliability. *Digital Discovery & e-Evidence*, 5(11), 2005.

[43] J Richard Landis and Gary G Koch. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement Among Multiple Observers. *Biometrics*, pages 363–374, 1977.

[44] Renaud LAPLANCHE, Joaquin DELGADO, and Matt TURCK. Concept search technology goes beyond keywords. *Information outlook*, 8(7), 2004.

[45] David D Lewis. The trec-4 filtering track. *Harman [7]*, pages 165–180, 1995.

[46] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 3–12, Dublin, IE, 1994.

[47] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[48] Adjoa Linzy. Attorney-client Privilege and Discovery of Electronically-Stored Information, the. *Duke L. & Tech. Rev.*, 2011.

[49] Giuseppe Manco, Elio Masciari, Massimo Ruffolo, and Andrea Tagarelli. Towards an adaptive mail classifier. In *Procedings of Italian Association for Artificial Intelligence Workshop*, 2002.

[50] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Workshop on Link Analysis, Counterterrorism and Security*, pages 33–44, 2005.

[51] Stefano Mizzaro. Relevance: The whole history. *Journal of the Association for Information Science and Technology*, 48(9):810–832, 1997.

[52] Alistair Moffat and Justin Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2, 2008.

[53] Robert C Moore and William Lewis. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224. Association for Computational Linguistics, 2010.

[54] Doug Oard, Fabrizio Sebastiani, and Jyothi Vinjumur. Minimizing the expected costs of review for responsiveness and privilege in e-discovery. *Manuscript under Review*, 2018.

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

[56] Douglas W Oard, Bruce Hedin, Stephen Tomlinson, and Jason R Baron. Overview of the trec 2008 legal track. Technical report, University of Maryland, College Park; College of Information Studies, 2008.

[57] Douglas W Oard, Jyothi Vinjumur, and Fabrizio Sebastiani. When is it rational to review for privilege? 2017.

[58] Douglas W Oard and William Webber. Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, 2013.

[59] Nicholas M Pace and Laura Zakaras. *Where the money goes: Understanding litigant expenditures for producing electronic discovery*. RAND Corporation, 2012.

[60] George L Paul and Jason R Baron. Information inflation: Can the legal system adapt. *Rich. JL & Tech.*, 13:1, 2006.

[61] Emily Pronin. Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, 11(1):37–43, 2007.

[62] Herbert L Roitblat, Anne Kershaw, and Patrick Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the Association for Information Science and Technology*, 61(1):70–80, 2010.

[63] Tanay K. Saha, Mohammad Al Hasan, Chandler Burgess, M. Ahsan Habib, and Jeff Johnson. Batch-mode active learning for technology-assisted review. In *Proceedings of the 3rd IEEE International Conference on Big Data (Big Data 2015)*, pages 1134–1143, Santa Clara, US, 2015.

[64] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *Journal of the Association for Information Science and Technology*, 58(13):2126–2144, 2007.

[65] Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, Vancouver, CA, 2008.

[66] Jitesh Shetty and Jafar Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, pages 74–81. ACM, 2005.

[67] Karen Sparck Jones and Cornelis Joost van Rijsbergen. Information Retrieval Test Collections. *Journal of documentation*, 32(1):59–75, 1976.

[68] Katrin Tomanek and Udo Hahn. A comparison of models for cost-sensitive active learning. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1247–1255, Beijing, CN, 2010.

[69] Stephen Tomlinson, Douglas W Oard, Jason R Baron, and Paul Thompson. Overview of the trec 2007 legal track. In *TREC*. Citeseer, 2007.

[70] Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Proceedings of the 15th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 2262–2269, Miami, US, 2009.

[71] Jyothi K Vinjumur. Evaluating Expertise and Sample Bias Effects for Privilege Classification in E-discovery. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 119–127. ACM, 2015.

[72] Jyothi K Vinjumur and Douglas W Oard. Finding the privileged few: Supporting privilege review for e-discovery. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[73] Jyothi K Vinjumur, Douglas W Oard, and Amittai Axelrod. An AID for Avoiding Inadvertent Disclosure: Supporting Interactive Review for Privilege in E-discovery. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 53–62. ACM, 2016.

[74] Jyothi K Vinjumur, Douglas W Oard, and Jiaul H Paik. Assessing the Reliability and Reusability of an e-discovery Privilege Test Collection. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2014.

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

[76] Ellen M Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 355–370. Springer, 2001.

[77] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 1. MIT press Cambridge, 2005.

[78] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and Topic Discovery from Relations and Text. In *International Workshop on Link discovery*, 2005.

[79] William Webber. Re-examining the Effectiveness of Manual Review. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval for E-Discovery Workshop*, page 2, 2011.

[80] William Webber, Douglas W Oard, Falk Scholer, and Bruce Hedin. Assessor error in stratified evaluation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 539–548. ACM, 2010.

[81] William Webber and Laurence AF Park. Score adjustment for correction of pooling bias. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 444–451. ACM, 2009.

[82] William Webber and Jeremy Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 929–932. ACM, 2013.

[83] William Webber, Bryan Toth, and Marjorie Desamito. Effect of written instructions on assessor agreement. In *Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1053–1054. ACM, 2012.

[84] Wenpu Xing and Ali Ghorbani. Weighted Pagerank Algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.

[85] Emine Yilmaz and Javed A Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 102–111. ACM, 2006.

[86] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. A New Rank Correlation Coefficient for Information Retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 587–594. ACM, 2008.

[87] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610. ACM, 2008.

[88] Dong Zhang, Daniel Gatica-Perez, Deb Roy, and Samy Bengio. Modeling interactions from email communication. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 2037–2040. IEEE, 2006.

[89] Justin Zobel. How Reliable are the Results of Large-scale Information Retrieval Experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. ACM, 1998.