The Book on AI Doc Review

Jim Sullivan

Published by eDiscovery AI
eDiscoveryAI.com

CONTENTS

INTRODUCTION: Artificial Intelligence is the Latest Frontier

It’s been over five years since Tony Reichenberger and I wrote “The Book on Predictive Coding.” At the time, I wouldn’t have believed you if you told me there would be a follow-up edition. Unfortunately, not much has changed in the world of Predictive Coding. Until now!

When I first started consulting on Predictive Coding matters, I was explaining the training process to a client, and they seemed a little overwhelmed. They told me, “I thought this was going to be a thing where I just push a button, and the entire review is complete,” I chuckled and broke the news to them. I said I’m sure we will have that kind of technology in 50 years, but until then, you will have to do it the hard way.

It turns out I was wrong. It took only 12 years.

We are entering a technological evolution unlike anything we have ever seen. The development of Artificial Intelligence is taking the world by storm. Every legal conference is practically dedicated to AI. You would be hard-pressed to find any legal discussion that can go 15 minutes without mentioning the term.

We have reached an age where **computers are capable of reviewing and classifying documents better than humans**. And that’s a big deal in eDiscovery.

What this book covers

The primary focus of this book is to provide you with everything you need to know to use Artificial Intelligence in document review. We will talk not just about technology but also about security, privacy, defensibility, and workflows.

While this book’s primary focus is Artificial Intelligence in eDiscovery, it will be discussed in context around how Predictive Coding technology and workflows have evolved and worked in the past. After all, precedent is the basis for everything in the legal industry. By comparing and contrasting the different methods, we can develop reliable and defensible workflows that aren’t much different from how things have always been done.

By the time you finish reading this book, you should be comfortable using Artificial Intelligence to help review and classify documents in eDiscovery.

The goal is to be comfortable with different strategies and workflows while understanding the proper steps to validate any process and ensure a defensible review.

I want to note there are many great uses of AI in the legal industry. Some tools help write briefs or motions, while others help with research. There are also many tools in eDiscovery designs to assist reviewers or help with investigations. This book is not about those use cases.

This book is about using technology to satisfy your legal obligations to produce relevant documents and identify the key documents necessary to win your case. The primary goal is to produce all the documents relevant to a Request for Production in the fastest, cheapest, and most accurate way possible. However, the secondary goal is less of a goal and more of a requirement. Everything done must be unquestionably defensible in the court of law. That is non-negotiable. If a method doesn't hold up in a courtroom, it's not worth writing about.

The Evolving Landscape of eDiscovery

eDiscovery technology has evolved rapidly compared to many other aspects of the legal profession. As we encounter ever-increasing data volumes, the manual document review methods have become increasingly impractical. We have made a lot of progress from the days of reviewing paper documents with hand-stamped Bates numbers.

The biggest change has come with the development of Technology Assisted Review (TAR) to aid in the identification and review of documents. We've moved on from the days of a purely linear process and now have an arsenal of tools at our disposal to identify relevant documents faster, cheaper, and more accurately.

From TAR 1.0, which allows us to classify massive volumes of data with very little work, to TAR 2.0 or Continuous Active Learning, which efficiently prioritizes relevant documents, we have made significant advances in eDiscovery technology.

With the introduction of Generative AI, there is no stopping the next wave of innovation, and it is going to be a good one.

In the following chapters, we will dive headfirst into this new frontier and discuss how Generative AI will be a game changer for reviewing and classifying documents in eDiscovery. We will explore how it works, techniques and strategies, and effective workflows to make your practice more effective. While it is certain the technology will improve quickly, it’s never too late to jump in!

1

WHY AREN'T YOU USING PREDICTIVE CODING?

Despite the massive advantages of Predictive Coding, adoption has been very uneven. A recent poll by eDiscovery Today concluded that 25.9% of respondents used predictive coding in all or most of their cases, but 36.3% used it in very few or none.

It's probably safe to say that if you are interested enough to pick up this book, you are more likely to have experience with Predictive Coding than others. Still, certainly, some of you have not had this experience.

The question is, WHY?!

Interestingly, adoption isn't sparse but uneven. That's because the people using it in every case are absolutely killing it. These people are game changers. I know of a large corporation that reduced its annual eDiscovery budget from $40M to $12M over two years due to the impressive work of ONE individual.

Now, ask yourself what is stopping you from embracing this technology. Is it because of any of these issues:

1. **Concerns about defensibility**: If anyone out there still doesn't think that using Predictive Coding is defensible, I have some news for you. I have personally supported over 1,000 cases using Predictive Coding to classify documents or exclude documents from review. And I'm not special. There have been TENS OF THOUSANDS of cases that have used Predictive Coding. And not ONCE has it been found not defensible. In the chapter about defensibility, we will get more into this, but this is no longer a valid excuse.
2. **Concerns about accuracy:** While studies have shown that Traditional Predictive Coding can generate more accurate results than humans, it's still pretty close. Using a TAR 1.0 style workflow, we could typically see a recall of 70%-80%, which isn't much better than human reviewers. AI-powered review blows that out of the water and can easily find 95%+ of the relevant documents. At the moment, we are entering a time where it's legitimately reasonable to suggest using humans is not defensible because of their consistently poor accuracy.

3. **Lack of familiarity with the technology:** I don't mean to call people out, but this is the real problem. Almost every excuse I've heard about adopting Predictive Coding stemmed from a lack of knowledge and fear of using the technology. Luckily, if you are reading this book, you are on track to becoming more familiar, gaining valuable knowledge, and eliminating those fears.

How do I know that lack of familiarity is the real issue?

For years, we evangelized the use of Predictive Coding and other workflows to add efficiency to reviews, and many clients would be excited to give it a shot. The technology was accessible, so we pushed it on "Every Case. Every Time." However, when we get on a call with the law firm to discuss the use of Predictive Coding, they always shut it down. I would hear the same thing repeatedly. It was always, "We love Predictive Coding, but just don't think it's a good fit for this case", and the client would always concede, and they would go about conducting their typical expensive linear review.

Until one day, we changed the strategy.

Anytime someone said they didn't want to use Predictive Coding, I immediately accepted that position and moved on to planning the workflow for their linear review. We were using a great workflow tool that would validate reviewers' coding and automatically routed batches to second-level reviewers, so some setup was necessary. We would walk through the reviewer protocol and set rules for what categories must be checked by first-level reviewers. We would then walk through how documents would route to second-level review.

Then, I would ask, "How do you want to batch out documents to your reviewers?"

"Would you like to batch documents in random order, by custodian, by date, or by batching the most relevant documents first?"

Using this strategy, they chose to batch the most relevant documents first **ONE HUNDRED PERCENT OF THE TIME**. I'm not kidding. Every. Single. Time. Then, we just flipped a switch to include every tagged document in training and built the CAL model nightly to allow for a prioritized review.

As part of the review reporting, we would send out a weekly report showing the number of documents reviewed and the response rate each day. As expected, the response rate immediately spiked, stayed high, and eventually dropped off. I would only do it once, but I always had to ask, “I just wanted to mention that it looks like you have a review team of XX people, and in the past week, you have only identified X number of relevant documents. I just wanted to let you know that if you were interested in stopping the review early, we can discuss how you can do that in a defensible way.”

And over half the time, they agreed to stop the review right there. The rest continued until all documents were reviewed, which was not a problem.

That’s the end of my mini rant. I just want you to know that if you aren’t using technology today, you are spending twice as much as you need to, going at 1/4th the speed, and have much lower accuracy than you would if you followed the techniques laid out in this book.

2

WHAT IS GENERATIVE AI? IS TAR THE SAME AS AI?

Let’s get some semantics out of the way:

There are lots of different names for Predictive Coding. Technology Assisted Review (TAR), Computer Assisted Review (CAR), Continuous Active Learning (CAL), and Intelligent Review, among tons of other names given to proprietary tools.

If it is a tool that uses a computer to predict the classification of documents, I refer to it as Predictive Coding in this book and in life. We’ll often refer to these models as “Traditional Predictive Coding” to distinguish them from AI-powered Predictive Coding.

We are also seeing a lot of generous uses of the term “Artificial Intelligence”, and most seem to be intended to deceive. So many vendors publish articles about their unique AI tools and then pull out a list of the TAR or CAL tools, sometimes even near duplicate identification or email threading!

The way that I see it, if you didn’t call any of these tools “Artificial Intelligence” before, the only possible reason you would call them AI now is in an attempt to confuse people and take advantage of the fact that people are excited about GENERATIVE AI. I can agree that they are technically correct to refer to machine learning as a form of artificial intelligence, but I just can’t get on board with the deception.

If there is any reference in this book to “Artificial Intelligence,” it will always be referring to Generative AI solutions using Large Language Models (LLMs). We referenced at least 10 different terms when talking about machine learning and supervised learning in our last book. Not once did we use the phrase Artificial Intelligence, so it wouldn’t be fair to begin now. Now that it’s out of the way.

What is Artificial Intelligence?

Artificial Intelligence is a broad term that refers to machines capable of performing tasks that typically require human intelligence. There are many types of AI, and lots of technologies that have existed for decades can be appropriately classified as AI.

However, what you see at every legal conference and hear on the news is something different. What everyone is referring to is Generative AI. This is the type of AI that can generate content and classifications.

What is Generative AI?

Generative AI is a type of AI that uses a large language model (LLM) designed to understand, generate, and interact with human language at a level never seen before. These models are trained on massive volumes of data to understand natural language and predict the most appropriate response.

LLMs represent a significant breakthrough in our ability to interact with machines with natural human language. This has opened incredible new opportunities across every industry. Not only can LLMs finish all your kid's homework in seconds, but it is also a game changer in everything we do. Here are a few of the unlimited use cases:

1. **Content generation and writing assistance**. Large language models can generate content and tell stories out of seemingly nothing. I often use this to create bedtime stories for my daughter. Every night generates a brand-new adventure of a princess saving the world from dragons and all kinds of bad guys. This is also great for "Give me 10 ideas for Christmas presents for my wife" or "Write a birthday card to my 99-year-old grandmother." She is turning 100 this summer, and I don't want to be the lame grandchild who doesn't include a special message in their card.
2. **Conversational AI and Chatbots**. Generative AI is incredibly good at communication and answering questions. They can serve as virtual assistants or even a life coach. If you haven't already noticed, a significant number of website chatbots are already powered by AI. While its speaking ability still needs some work, Wendy's has already begun using AI to take orders in their drive-thru windows.
3. **Education and research**. AI does an incredible job answering questions and explaining how things work. It can generate hypotheses and present ideas.

4. **Programming and code generation**. This alone would be the most significant innovation of a decade. LLMs can effectively code anything from scratch. It can write quick scripts to solve simple problems, debug existing code, explain how code works, and even build complex systems. The speed at which software can be developed and deployed is shortening daily.
5. **Healthcare**. AI can assist in analyzing patient data, medical literature, and research papers. It can find that one reference that is only available buried deep in an obscure reference book.
6. **Image, audio, and video generation**. Describe any image you want, and AI can draw it nearly perfectly in seconds. The amount of joy kids get out of this is incredible. I will warn you that every time I show a child how to generate images with AI, the prompts start with "Draw me a cat riding a unicorn and jumping over a rainbow," but it rarely takes more than 5 minutes before it turns into "draw a poop with a poop" while they roll on the floor giggling.
7. **Data extraction and analysis**. This is the one we are here for. AI can analyze text and make classifications. It can also summarize large chunks of data very quickly. Additionally, it can extract key information from documents.

While I think you get the picture, the technology is pretty incredible. If you haven't used AI before and want to take a spin, the easiest way to get started is to go to https://chat.openai.com, sign up for a free account, and ask questions. However, you should understand that you should **NEVER PUT PRIVATE DATA THROUGH THIS TOOL**. We will discuss this more in future chapters, but if you aren't comfortable with the entire world seeing what you are doing, you should not be doing it on this tool.

The difference between TAR and Generative AI

The biggest difference between TAR and the type of Generative AI we use in eDiscovery is how the machine is trained.

With TAR, we use humans to train the machine on classifications. A TAR process requires a human trainer to classify documents as positive (Relevant) or negative (Not Relevant) examples. By providing enough examples, the machine can learn the patterns, correlations, or characteristics that make a document Relevant or Not Relevant. Through training, the system "learns" what makes a document Relevant vs Not Relevant and can then score the documents accordingly. The score in most Predictive Coding tools merely measures how closely a document matches your set of trained examples.

- A TAR document with a score of 99 is not hot or important. It just means the document closely matches the positive examples you have trained.
- A TAR document with a score of 50 is not a borderline or uncertain document. It just means the document does not closely match anything you have trained.
- A TAR document with a score of 0 is not junk. It just means the document closely matches the negative examples you have trained.

TAR can help us quickly classify a large amount of data based on the examples we used to train the model.

In a typical TAR 1.0 case, I tell clients they should expect to train 10,000 documents to sufficiently train the model (assuming no rolling uploads). However, in almost all instances, we get adequate results with around 5,000 documents, which includes a control set.

It is one of the reasons some are discouraged from using a TAR 1.0 workflow. Training 5,000 documents can take over 80 hours for a highly skilled subject matter expert. That might be a lot of billable hours, but most attorneys just can't find time to add an extra 80 hours to their workload.

This is where AI changes the game.

An AI Review workflow does not require training examples. The model already understands natural human language. With AI, all you need to do is provide the instructions to the machine so it knows what to look for. That can be as easy as typing out simple instructions explaining what you want:

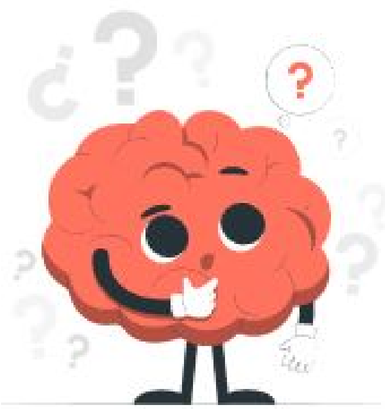
"All documents where an Acme employee suggests that pricing of widgets should be modified."

You'll notice the instructions read like a Request for Production, which is exactly what they are. In most cases, we simply copy the exact language from the Request for Production to start our instructions.

While we need to test our instructions and verify that the results are accurate, this can usually be done in an hour or two instead of taking days or weeks to train a model.

So how does this work?

The Magic of Large Language Models



Though we aren't going to turn this book into a deep dive into Large Language Models (LLMs), we can simplify the issue so we can have a high-level understanding of the process.

Large Language Models are a type of artificial intelligence that understands and generates language by predicting the next word in a sentence. Imagine you are playing a game where you must guess the next word your friend will say - that's what these models do on a much larger scale. They are trained on a massive amount of text from books, websites, and other sources.

When you ask a large language model a question or give it a prompt, it uses what it has learned to generate a response. It considers patterns of words, the way sentences are structured, and the meaning behind your request to construct a reply.

Let's walk through a simple example.

Imagine you are a large language model attempting to answer a question. You have determined that the answer will begin with the following words:

The cat likes to…

What would you suggest is the next word? There is certainly more than one possibility. I see it like

Family Feud.

“We polled 1 trillion articles to find what is most commonly the next word in this sentence.”

What is most likely to be the next word?

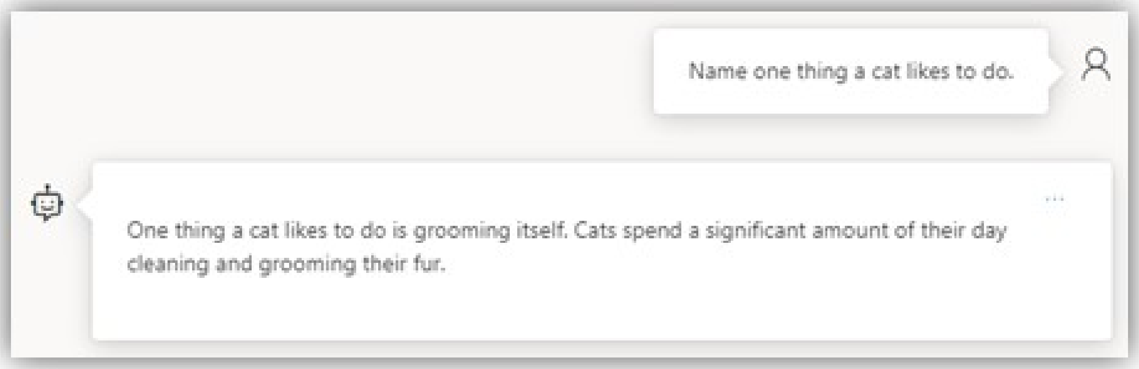
Sleep?

Eat? **Play?**

Climb?

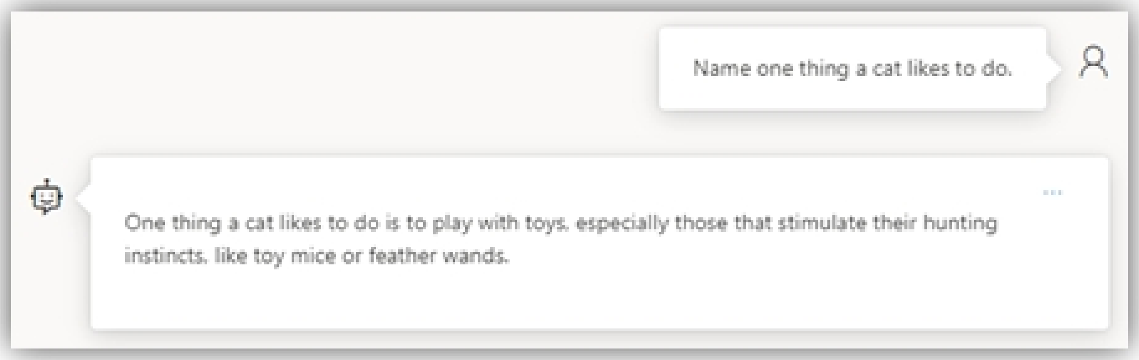
The large language model looks at all of its training data and calculates the likelihood of each word being selected. Based on the settings of the LLM, the result can be more varied or more deterministic.

This is why we will get slightly different responses when asking the exact same question:



Name one thing a cat likes to do.

One thing a cat likes to do is grooming itself. Cats spend a significant amount of their day cleaning and grooming their fur.



Name one thing a cat likes to do.

One thing a cat likes to do is to play with toys, especially those that stimulate their hunting instincts, like toy mice or feather wands.

Of course, some settings control how much variance you will see in responses. In legal applications, using more deterministic settings with less variance is better, as creativity isn't generally encouraged. A person using AI to write a fantasy novel would want to use a much less deterministic setting.

While this certainly plays a key role when generating document summaries in eDiscovery, it is much less significant when making simple classifications.

Tokens

One more thing before we get done with the boring nerd stuff. Let's talk about tokens.

In the context of a large language model, a token is the smallest unit of data the model processes. The system splits up your input text into tokens for processing. While it isn't exactly true, you can probably think of tokens as words. The number of tokens in your input and output is basically the number of words in your input and output.

The number of tokens in your input and output determines what size documents we can review with AI and how much it will cost.

As of today, it costs a significant amount of money to run an LLM, and large documents have a much higher processing cost than small documents.

If you try to have AI review a large document with thousands of pages, it could cost well over $100 for a single document.

This means that whatever vendor you are working with on an AI review, there will probably be limitations on the size of the documents you can review. Based on rumors in the industry, I believe most vendors will start out with a token limit of 32,000, which allows for processing documents with about 50-75 pages of text. But this will change quickly as technology improves, and in 5 years, it will probably no longer be an issue.

Why do I care about any of this?

You sincerely don't need to care about how Large Language Models work, but some basic background information is helpful. As we start talking about defensibility, we will see the only thing that really matters is showing that the results are accurate.

There is one thing about Large Language Models that you must know: security and privacy. We'll get into that next.

3

SECURITY AND PRIVACY OF LARGE LANGUAGE MODELS

As attorneys, should we trust AI with our data?

I’m sure you know that the answer will always be, “It depends.”

This is one of the most common attacks on the use of AI because people are very concerned about their data, as they should be. But it’s not nearly as complicated as some may make it seem.

The reasoning for this is clear: Most public AI tools do not protect your data. They do not claim to protect your data and are not trusted with confidential information.

This is not new.

Most large companies (I’m looking at you, Google) are horrible at protecting user data. As they say, “If you aren’t paying for a product, you are the product.” When you enter a query into the Google search box, you give your data to Google on their terms. They will use your data to customize ads or services offered to you and will share your data with third parties. The goal is to make money off your data. No reasonable attorney would think confidential client information should ever be shared with this service.

AI is no different.

If you enter a query into a public AI service, there are no guarantees that your data will remain confidential. In fact, it’s fair to say that you can expect that it will NOT remain confidential.

This applies to other areas related to eDiscovery. If you want to translate text from one language to another, if you simply copy the text from your database and paste it into Google Translate, you are playing with the same kind of fire. Google will not maintain your data in the manner the legal industry requires.

For some background here, this issue first occurred when it was discovered that OpenAI would use your input data to train future models. As we know how large language models work, if your data is used for training, there is no guarantee that it won’t pop up as a result of someone’s query. This is simply not acceptable in the legal industry.

When someone asks, “Does your LLM use our data to train the model?” what they really mean is, “I read an article that said OpenAI was using client data to train their model, and I want to make sure I don't have to worry about you doing the same." This is a very fair and appropriate question.

What you need to do to prevent these issues:

1. **Ask your AI service provider if your data is being shared with any 3rd parties.**
2. **Ask your AI service provider if your data is being used for training.**

That's it. That's the extent of what you need to know about privacy and security with AI.

Now, don't forget to ask the other questions that come up with any provider. You want to ensure all data is encrypted at rest and in transit, understand their data retention policies, and possibly do some investigation and penetration testing. Still, these questions would be relevant in every case and not specific to an AI review, so we won't go into them here.

4

USING AI IN ELECTRONIC DOCUMENT REVIEW

We’ve finally made it to the good stuff. Let’s talk about using AI in document review. Let’s start by clarifying what that means.

What is an AI-powered document review?

When we talk about AI-powered document review, we explicitly talk about using large language models to classify documents.

The idea is quite simple: Suppose you have a relatively simple matter where you review documents to determine if they are relevant to any of your 5 issues.

First, you describe what you are looking for to the machine in natural language. Here is a potential example:

- **Issue 1**: All documents that discuss the pricing of widgets.
- **Issue 2**: All documents that report on widget sales between 2021 and 2023.
- **Issue 3**: All documents that show profits and losses generated by widget sales in 2022.
- **Issue 4**: All documents where one party is suggesting to another that the price of widgets needs to be adjusted.
- **Issue 5**: All documents where an employee of ACME is discussing the pricing of widgets with a competitor in a way that might suggest there is a price-fixing arrangement.

As you can see, Issues 1-4 are general relevance issues, but Issue 5 is much narrower and would likely be used as a means to identify potentially key documents. The good news is that AI doesn’t care how broad or narrow your issues are!

Next, the system reviews your documents and generates output. That output will vary from system to system, but it generally includes:

1. A summary of the contents of the document.
2. A classification indicating whether the document is relevant for each of the five issues.
3. An explanation demonstrating why the document is or is not relevant to each of the issues given.

That's it! You explain what you want, and the computer tells you which documents are relevant to that query. Throw away your training sets because it will never get easier than this.

Accuracy

How good is AI at document review?

I don't say this lightly: IT IS INCREDIBLE!

Whenever anyone asks me why I believe AI Review is the future of eDiscovery, the answer is that it is too good. It's better than even the most focused subject matter experts. Humans cannot compete.

We regularly see recall scores of 95% or higher ON ISSUE CODES!

AI is a subject matter expert on everything. We once had AI review a large volume of data from Tim Kaine's time as governor of Virginia, searching for documents related to the Virginia Tech shooting. A document that was returned as relevant said,

"I'm sorry about everything that has happened. I'll see you at the memorial service tomorrow. God bless."

So, I'm looking at this document, which I categorized as Not Relevant, and could not, for the life of me, figure out what the AI was thinking. I read the explanation, and it said,

"This document is dated 3 days after the Virginia Tech shooting. Given the memorial service and the context, it is likely they are referring to the Virginia Tech shooting." It blew my mind!

In another example, we edited a long document to include a line near the signature block that said,

"Btw, did you see Peyton Manning last night? He was incredible!"

When we ran the document across AI asking for "All documents with any mention about football", of course, it came back as Relevant, with an explanation of "Peyton Manning is a football player, which makes it relevant to the issue."

It understands context. It understands sentiments. It understands slang. It understands abbreviations.

So, what other things can it do?

Relevancy Review

The most expensive and time-consuming part of document review is identifying the relevant documents. That’s where AI excels.

As you can see in the previous example, you simply develop the proper instructions. Then, the AI can review your documents with nearly perfect accuracy. With little up-front work, you can push a button to find all your relevant documents. We will go further into how to develop good instructions later in this book.

Privilege Review

Just like finding relevant documents, the AI can find privileged documents with the proper instructions. Simply describe what makes a document privileged, and it will identify documents that meet your description.

Priv Logging

Because we have explanations describing what makes a document relevant to your privilege criteria, we slightly modify the instructions. We can have the explanations formatted in a way that is perfect for generating Priv log content.

Identifying and Extracting PII and PHI

If you think about it, identifying private information is one of the easier aspects of document review. We all know regular expressions are already nearly perfect in recognizing social security numbers, so this isn’t much of a leap. But now we can identify and extract.

I will caution that instructions like

“All documents that contain private personal information”

will probably not give you a great result because the AI will return nearly everything as Relevant with an explanation that says, "This document contains an email address, which is personal information."

Instead, you have to be a little more specific, like

"All documents that contain any social security numbers, credit card numbers, bank account numbers, personal health information, or passwords."

We will go into more detail on how to craft good instructions later in this book.

Finding Key Documents

One limitation we have always had with Predictive Coding is finding the hot documents or the needle in a haystack. People would frequently tell me, "We aren't even sure if the document we are looking for exists", and I have to break the news to them that Predictive Coding isn't the right solution for them.

Using Predictive Coding tools requires you to train positive and negative examples to build a classifier. It will simply never work if you do not have enough positive examples. We see this all the time with issue codes. Predictive Coding is miserable at identifying issue codes. We usually don't even bother trying because the results will be embarrassing. Still, if we have 10 issue codes, the odds of getting 1 or 2 above 50% recall are not very good. It simply does not work.

But AI excels at this! It doesn't matter if there is one relevant document in your collection or a million. Because it is analyzing documents one at a time, it doesn't matter how rich the dataset is. When we talk about how great AI is at identifying relevant documents, we are usually talking about ISSUE CODES, which is incredible and nothing we have ever been able to do previously.

Foreign Language Review

If someone wants to use Predictive Coding on a multi-language dataset, my advice to you would be to RUN away as fast as you can.

Predictive Coding is language agnostic, but it does need to be trained on each language in your collection. If you have two languages, your effort doubles. But it's worse than that because there is rarely a single subject matter expert that can train in every language, so now you have multiple people training, which makes the training much less consistent. And the volume of relevant documents is rarely split equally between languages, so you may struggle to find enough relevant documents to train in a specific language. In most cases, the solution was simply to use TAR on your most prevalent language (usually English) and then do a linear review of whatever was left. And we know it's not cheap or easy to find a team of foreign language reviewers.

But AI doesn't care about languages. If the Large Language Model has been trained in multiple languages, it can review documents in multiple languages. Even documents in mixed languages!

We have tested documents in 20 different languages, including mixed languages, and the results show no meaningful differences.

Short Message Review

If you ignore all the discussion about AI for just a second, you might hear another issue that keeps popping up. Short messages.

You should have no problem finding a panel on short messages at any legal conference near you.

Whether it's SMS messages, chat, or another form of short message, the problems are similar. Traditional solutions don't work well with short messages.

Frequent misspellings, use of slang, and abbreviations, among other things, can make keywords and predictive coding ineffective. People are looking for better solutions.

The answer is AI. AI doesn't have a problem with spelling issues. AI understands slang and abbreviations. It is nearly perfect at doing everything that may cause trouble for traditional tools.

Image Review

What is the first step of a Predictive Coding project? Identifying the documents that lack extracted text. Without text, predictive coding does not work.

But those days are behind us. AI can recognize objects in images and classify them just as it would with a document. You also get a nice summary describing the image.

Audio Review

Why stop at images when we can do audio? AI can listen to an audio clip, summarize the contents, and classify it for whatever issues you might be interested in.

This isn’t talked about very often, but removing the limitations of needing extracted text is a really big deal.

What’s Next?

With the fast pace of improvements to AI technology, there are other things that you should see in the not-so-distant future.

The ability to review videos is just a matter of time. That might be available by the time this book is published.

The ability to apply redactions is going to be a game-changer. Being able to identify content to redact AND apply the redactions is going to be incredible.

What else? I’m sure there will be some good surprises in the next few years that blow us away.

5

THE PROCESS

While theory is great for some things, the only way to understand a process is to walk through each step. So, let's do it!

We will walk through the process of performing a relatively simple Relevancy review. This will be a linear review of the dataset:

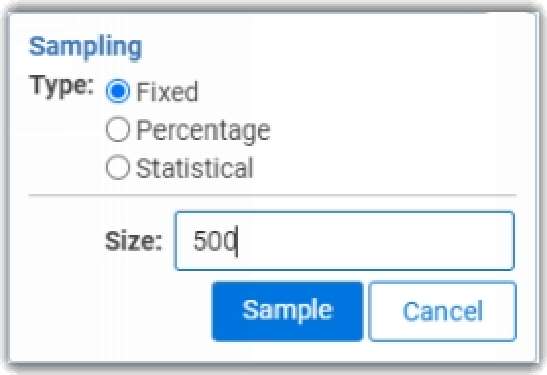
Background: For this hypothetical, we are reviewing 67,902 documents in the Jeb Bush dataset, which is publicly available from the time he was governor of Florida. We are looking for documents that discuss funding schools in Florida. Also, just for fun, we are going to see how many people in Florida write to Jeb to encourage more school funding vs. the number of people who write to suggest less school funding.

Answer Key: For convenience, we are going to use the TREC-provided answer key as a shortcut to determine which documents are relevant to each topic. It's not perfect, but it's pretty good. In a real case, we would need a subject matter expert available to review the documents to create an answer key.

Let's get started!

The first step is to create our instructions to tell the AI what we are looking for.

Round 1 Instructions



Sampling

Type: ● Fixed ○ Percentage ○ Statistical

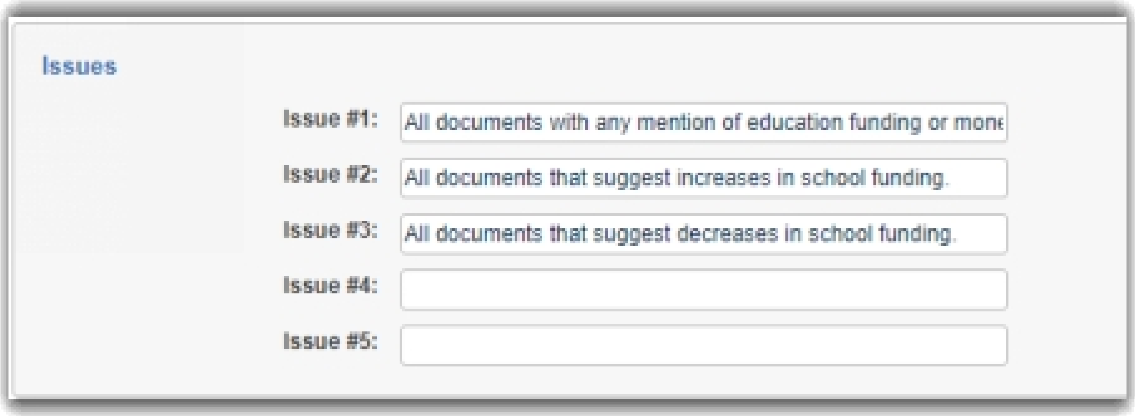
Size: 500

To do that, we need to identify the documents you need to review. From that set, we are going to take a random sample of documents. We selected 500.

Then, we run the documents through eDiscovery AI.

This is when we are presented with the opportunity to enter your instructions. Let's come up with something basic but very broad:

"All documents with any mention of education funding or money going to schools."
"All documents that suggest increases in school funding."
"All documents that suggest decreases in school funding."



Issues

Issue #1: All documents with any mention of education funding or mone

Issue #2: All documents that suggest increases in school funding.

Issue #3: All documents that suggest decreases in school funding.

Issue #4:

Issue #5:

We run these documents and look at the results…

REVIEWING DOCUMENTS

Please Hold

Round 1 Results

The results are in!

The first step is to folder up the True Positives, True Negatives, False Positives, and False Negatives. Remember, we have the TREC answer key to rely on, so we don’t have to review the documents like you would in a real matter.

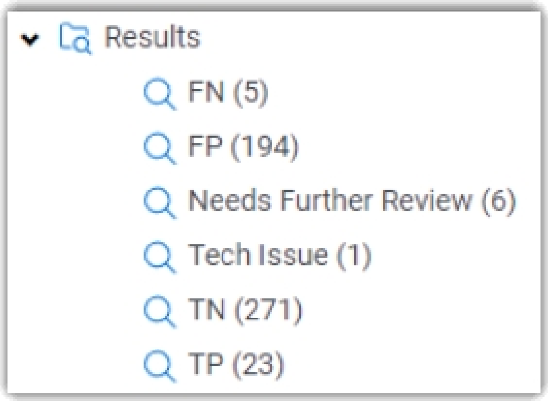
A quick guide on the confusion matrix:

True Positive = Instances where the AI classified the document as Relevant, and the Answer Key also classified the document as Relevant.

True Negative = Instances where the AI classified the document as Not Relevant, and the Answer Key also classified the document as Not Relevant.

False Positive = Instances where the AI classified the document as Relevant, but the Answer Key classified the document as Not Relevant.

False Negative = Instances where the AI classified the document as Not Relevant, but the Answer Key classified the document as Relevant.



Results

FN (5)

FP (194)

Needs Further Review (6)

Tech Issue (1)

TN (271)

TP (23)

We can use these to calculate Recall and Precision for our first run.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} = \frac{23}{(23 + 5)} = 82\%$$

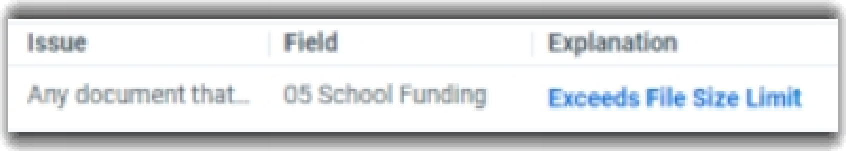
$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} = \frac{23}{(23 + 194)} = 11\%$$

The recall score is excellent for a first pass, but the precision definitely needs some work.

Now, we want to quickly review the documents to see what we can improve.

Technical Issue

There is only 1 document that is a tech issue, and a quick review shows us that it exceeds the allowed file size limit.

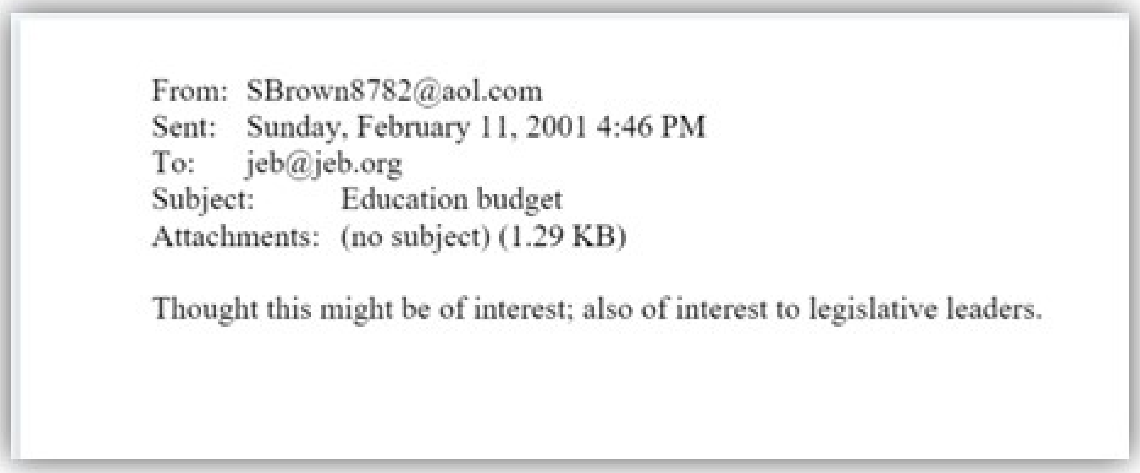


Issue	Field	Explanation
Any document that...	05 School Funding	Exceeds File Size Limit

While we could raise the file size limit and potentially review this document, we will leave it for now to save on costs.

Needs Further Review

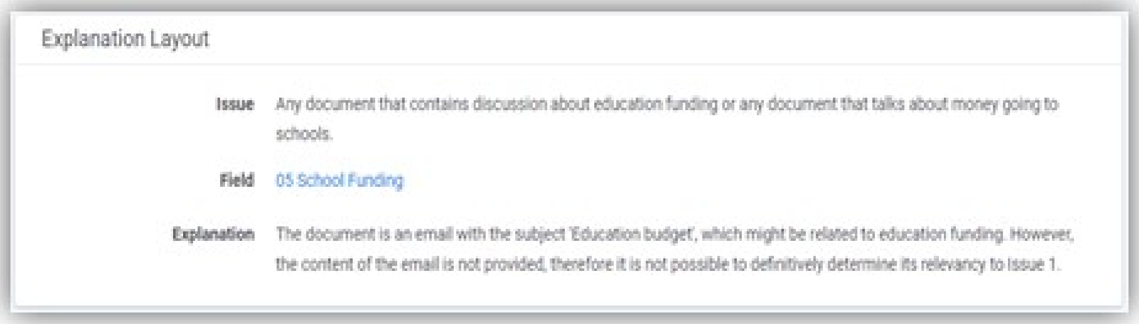
Next, we will look through the 6 docs that need further review. What we discover are documents without much content. For example:



From: SBrown8782@aol.com
Sent: Sunday, February 11, 2001 4:46 PM
To: jeb@jeb.org
Subject: Education budget
Attachments: (no subject) (1.29 KB)

Thought this might be of interest; also of interest to legislative leaders.

The explanations make it easy to understand why this was classified as Needs Further Review:



Explanation Layout

Issue Any document that contains discussion about education funding or any document that talks about money going to schools.

Field 05 School Funding

Explanation The document is an email with the subject 'Education budget', which might be related to education funding. However, the content of the email is not provided, therefore it is not possible to definitively determine its relevancy to Issue 1.

A document is usually classified as Needs Further Review because the system is uncertain how to classify it. Just like how a human reviewer might come up and ask questions about categorizing a document.

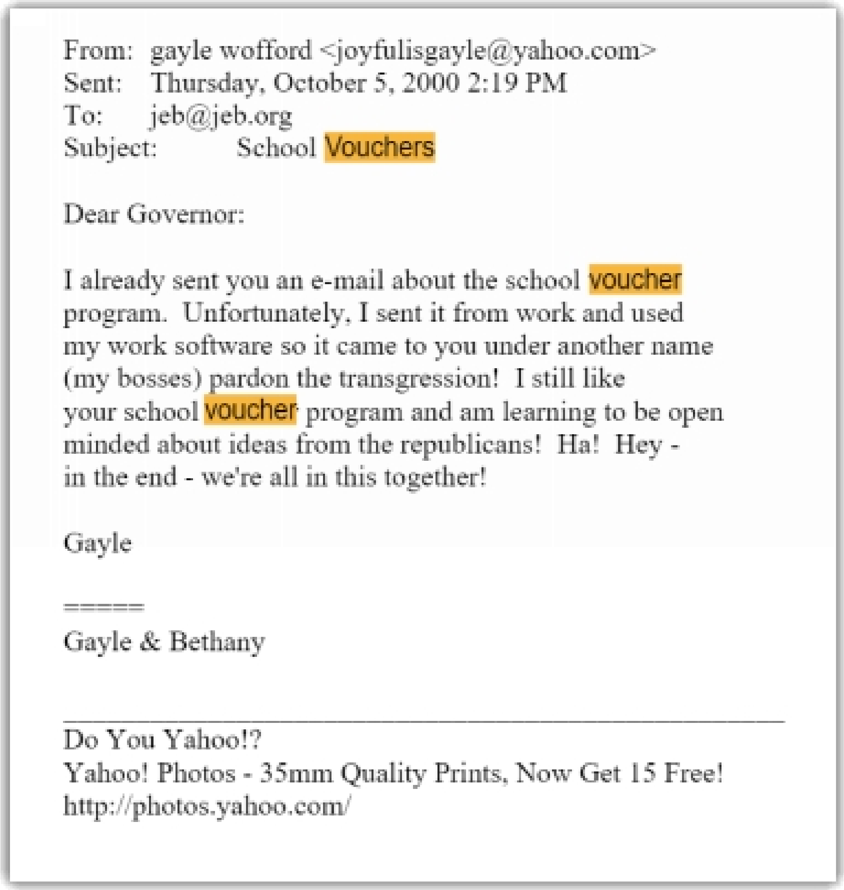
Since there is a very low volume, and all of them are borderline cases, we are OK with these being marked as Needs Further Review, but we will try to clarify a little more in our next set of instructions.

False Negatives

False negatives are the worst. Not just because they sound bad but because it means we are missing Relevant material. I'd rather have 5 false positives than 1 false negative.

The purpose here is to find out what we missed so we can update our instructions to capture them in the next round.

I added some highlighting to help us find key topics as we look through these 5 documents.



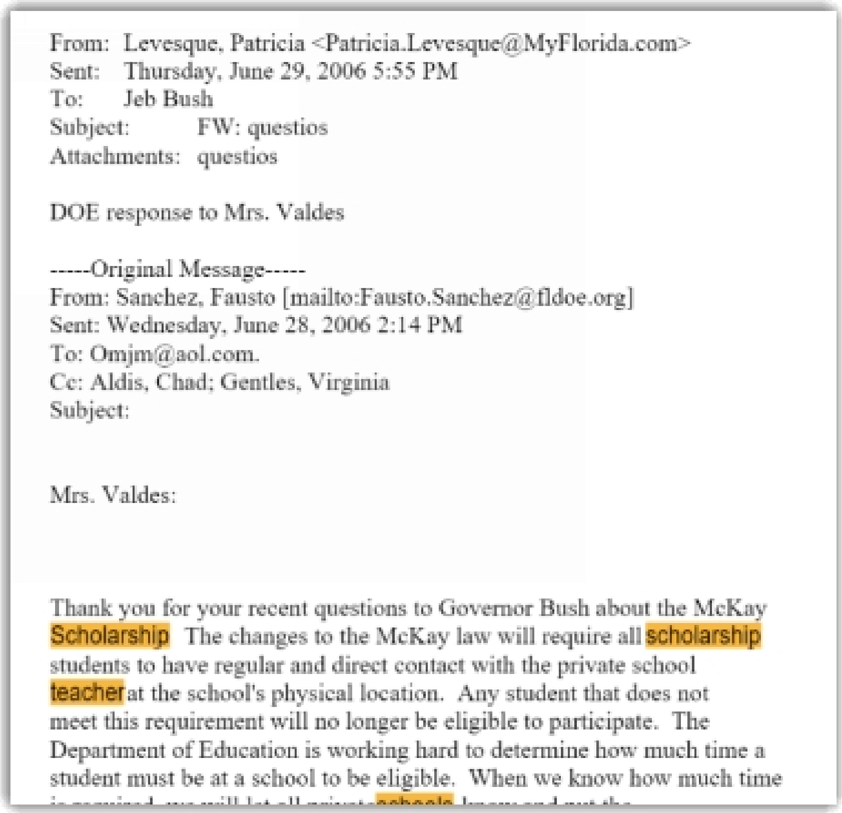
From: gayle wofford <joyfulisgayle@yahoo.com>
Sent: Thursday, October 5, 2000 2:19 PM
To: jeb@jeb.org
Subject: School Vouchers

Dear Governor:

I already sent you an e-mail about the school voucher program. Unfortunately, I sent it from work and used my work software so it came to you under another name (my bosses) pardon the transgression! I still like your school voucher program and am learning to be open minded about ideas from the republicans! Ha! Hey - in the end - we're all in this together!

Gayle

=====
Gayle & Bethany



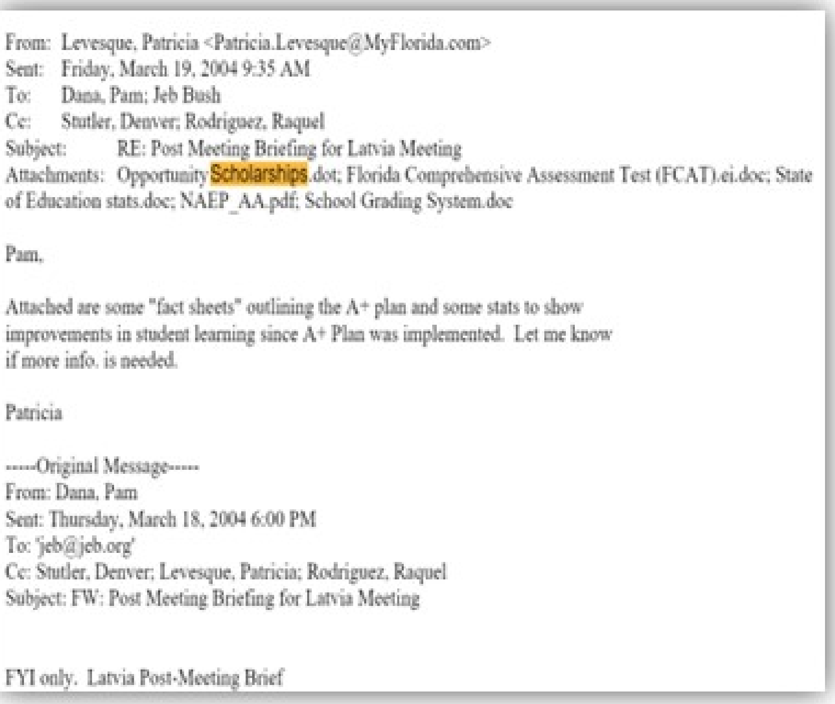
From: Levesque, Patricia <Patricia.Levesque@MyFlorida.com>
Sent: Thursday, June 29, 2006 5:55 PM
To: Jeb Bush
Subject: FW: questios
Attachments: questios

DOE response to Mrs. Valdes

-----Original Message-----
From: Sanchez, Fausto [mailto:Fausto.Sanchez@fldoe.org]
Sent: Wednesday, June 28, 2006 2:14 PM
To: Omjm@aol.com.
Cc: Aldis, Chad; Gentles, Virginia
Subject:

Mrs. Valdes:

Thank you for your recent questions to Governor Bush about the McKay Scholarship The changes to the McKay law will require all scholarship students to have regular and direct contact with the private school teacher at the school's physical location. Any student that does not meet this requirement will no longer be eligible to participate. The Department of Education is working hard to determine how much time a student must be at a school to be eligible. When we know how much time is required, we will let all private schools know and put the



From: Levesque, Patricia <Patricia.Levesque@MyFlorida.com>
Sent: Friday, March 19, 2004 9:35 AM
To: Dana, Pam; Jeb Bush
Cc: Stutler, Denver; Rodriguez, Raquel
Subject: RE: Post Meeting Briefing for Latvia Meeting
Attachments: Opportunity Scholarships.dot; Florida Comprehensive Assessment Test (FCAT).ei.doc; State of Education stats.doc; NAEP_AA.pdf; School Grading System.doc

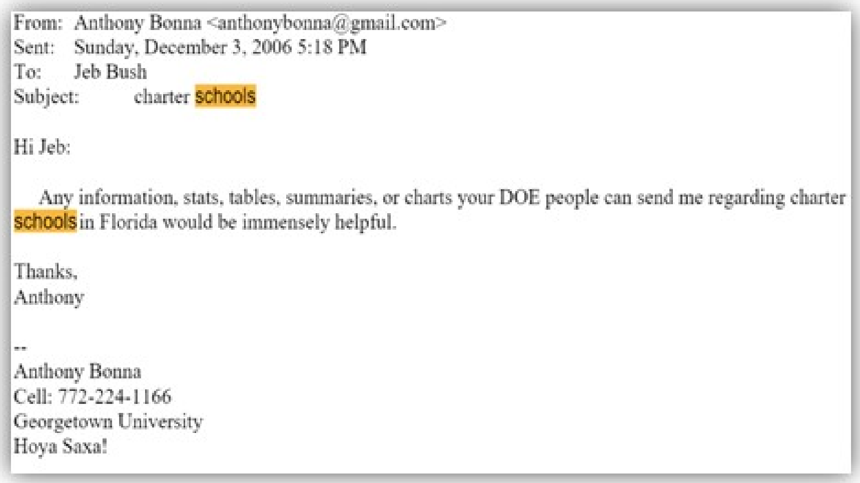
Pam,

Attached are some "fact sheets" outlining the A+ plan and some stats to show improvements in student learning since A+ Plan was implemented. Let me know if more info. is needed.

Patricia

-----Original Message-----
From: Dana, Pam
Sent: Thursday, March 18, 2004 6:00 PM
To: 'jeb@jeb.org'
Cc: Stutler, Denver; Levesque, Patricia; Rodriguez, Raquel
Subject: FW: Post Meeting Briefing for Latvia Meeting

FYI only. Latvia Post-Meeting Brief



From: Anthony Bonna <anthonybonna@gmail.com>
Sent: Sunday, December 3, 2006 5:18 PM
To: Jeb Bush
Subject: charter schools

Hi Jeb:

Any information, stats, tables, summaries, or charts your DOE people can send me regarding charter schools in Florida would be immensely helpful.

Thanks,
Anthony

--
Anthony Bonna
Cell: 772-224-1166
Georgetown University
Hoya Saxa!

We immediately see that a few topics were missed:

- School Vouchers
- Scholarships
- Charter Schools
- Teacher Salaries

This is perfect and gives us what we need to improve our instructions in the next round.

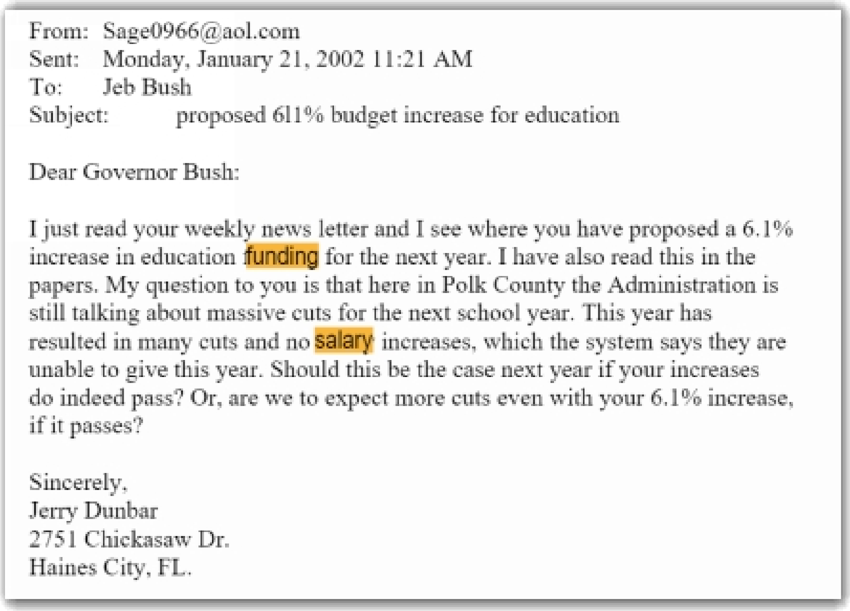
False Positives

False positives aren't quite so terrible. We didn't miss anything but labeled documents as Relevant when they were not.

A quick review of these documents determined a couple of things:

1. Almost every document returned was related to school funding.
2. The TREC answer key missed a lot of Relevant documents.

Here are a couple of the most obvious examples:

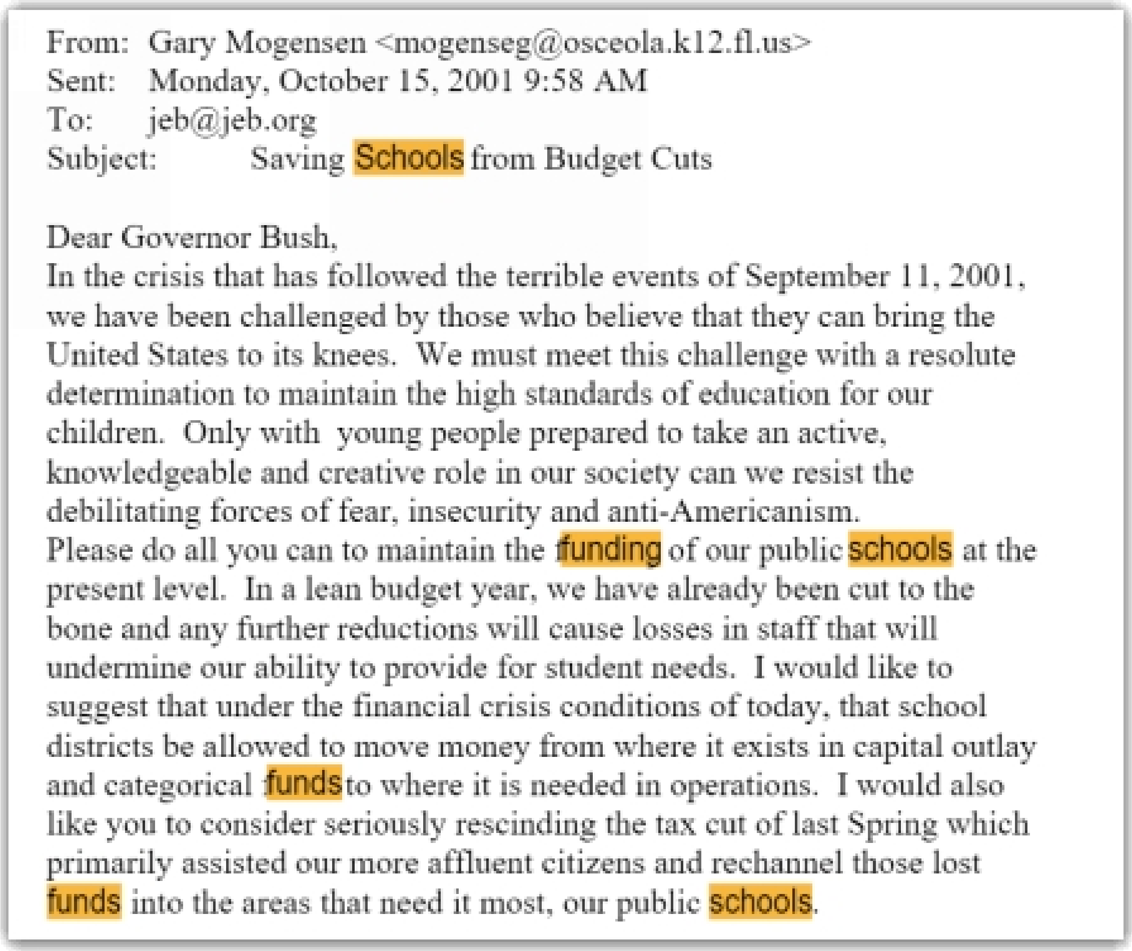


From: Sage0966@aol.com
Sent: Monday, January 21, 2002 11:21 AM
To: Jeb Bush
Subject: proposed 6l1% budget increase for education

Dear Governor Bush:

I just read your weekly news letter and I see where you have proposed a 6.1% increase in education funding for the next year. I have also read this in the papers. My question to you is that here in Polk County the Administration is still talking about massive cuts for the next school year. This year has resulted in many cuts and no salary increases, which the system says they are unable to give this year. Should this be the case next year if your increases do indeed pass? Or, are we to expect more cuts even with your 6.1% increase, if it passes?

Sincerely,
Jerry Dunbar
2751 Chickasaw Dr.
Haines City, FL.



From: Gary Mogensen <mogenseg@osceola.k12.fl.us>
Sent: Monday, October 15, 2001 9:58 AM
To: jeb@jeb.org
Subject: Saving Schools from Budget Cuts

Dear Governor Bush,
In the crisis that has followed the terrible events of September 11, 2001, we have been challenged by those who believe that they can bring the United States to its knees. We must meet this challenge with a resolute determination to maintain the high standards of education for our children. Only with young people prepared to take an active, knowledgeable and creative role in our society can we resist the debilitating forces of fear, insecurity and anti-Americanism.
Please do all you can to maintain the funding of our public schools at the present level. In a lean budget year, we have already been cut to the bone and any further reductions will cause losses in staff that will undermine our ability to provide for student needs. I would like to suggest that under the financial crisis conditions of today, that school districts be allowed to move money from where it exists in capital outlay and categorical funds to where it is needed in operations. I would also like you to consider seriously rescinding the tax cut of last Spring which primarily assisted our more affluent citizens and rechannel those lost funds into the areas that need it most, our public schools.

The great news is that nothing here stands out as Not Relevant. Of all 194 documents, the only documents that were truly Not Relevant were close calls that I had no problem including as Relevant. There were only a few that discussed schools but didn't have a direct reference to funding. We can update our instructions to fix that.

True Positives and True Negatives

I looked through several examples and didn't see anything I disagreed with. On a live matter, we would need to sample and review a larger set, but for this example, we will skip over these.

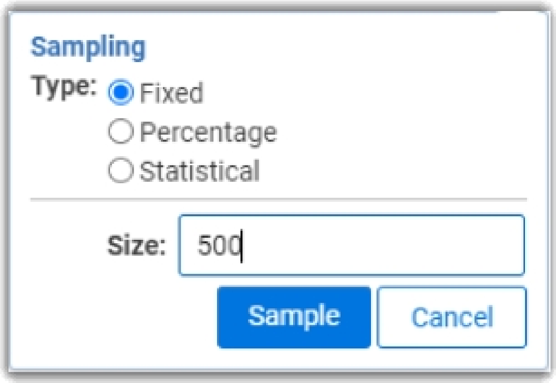
Round 2 Instructions

Now we just have to take what we learned in Round 1 and update our instructions:

“All documents with any mention of education funding or money going to schools. Any discussion of School Vouchers should be considered Relevant. Any discussion of Charter Schools should be considered Relevant. Any discussion about School Scholarships should be considered Relevant. Any discussion about teacher salaries, class size, or money for textbooks should be considered Relevant. If it does not include a direct reference to funding, it should be considered Not Relevant.”

“All documents that suggest increases in school funding.”

“All documents that suggest decreases in school funding.”



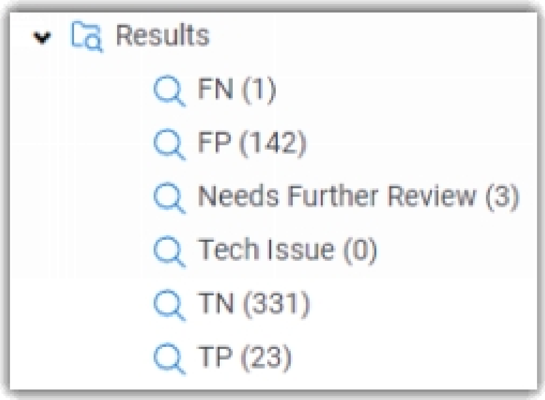
Sampling

Type: ● Fixed ○ Percentage ○ Statistical

Size: 500

Then we take another random sample of 500 documents, and we are off to the races!

Round 2 Results



Results

FN (1)

FP (142)

Needs Further Review (3)

Tech Issue (0)

TN (331)

TP (23)

The results are in!

The first step is to folder up the True Positives, True Negatives, False Positives, and False Negatives. Remember, we have the TREC answer key to rely on, so we don't have to actually review the documents like you would in a real matter.

The results already look much better.

However, we know many of the classifications on the answer key are not accurate, so we will take a look before we try to calculate our metrics.

Technical Issue

There are none to review.

Needs Further Review

Next, we need to look through the 3 that require further review.

All were generally on topic but lacked any Relevant content. I changed these to Not Relevant.

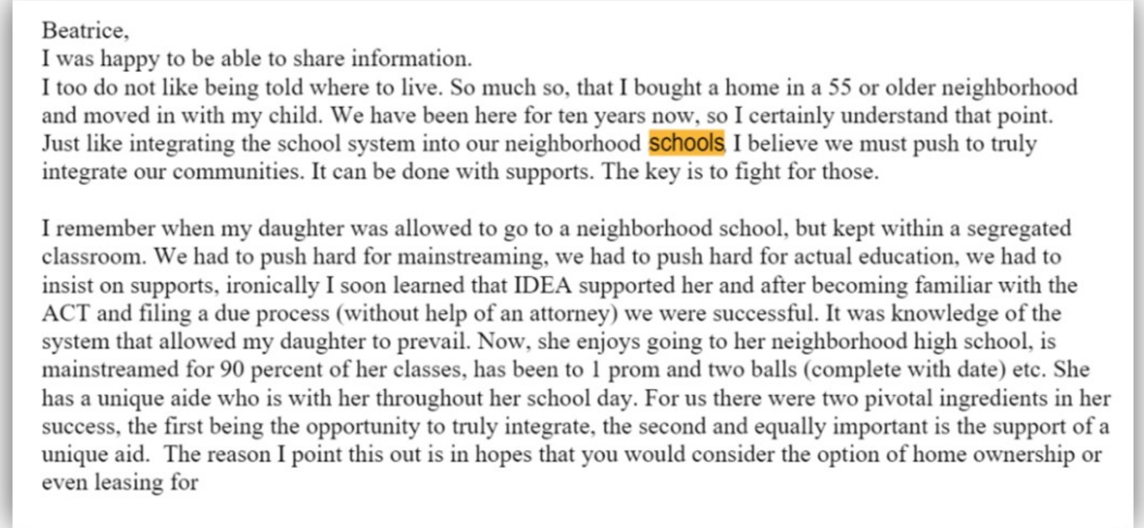
False Negatives

False negatives are still the worst, but now we only have one.

In reviewing this lengthy document, there was a small section discussing schools, but nothing I could find related to funding. There was also a reference to vouchers, but not in the context of schools.

Let's just say I would have marked this as Not Relevant, but for this exercise, we'll allow it so nobody accuses us of cheating.

Here is the (maybe) relevant portion:



Beatrice,
I was happy to be able to share information.
I too do not like being told where to live. So much so, that I bought a home in a 55 or older neighborhood and moved in with my child. We have been here for ten years now, so I certainly understand that point. Just like integrating the school system into our neighborhood schools I believe we must push to truly integrate our communities. It can be done with supports. The key is to fight for those.

I remember when my daughter was allowed to go to a neighborhood school, but kept within a segregated classroom. We had to push hard for mainstreaming, we had to push hard for actual education, we had to insist on supports, ironically I soon learned that IDEA supported her and after becoming familiar with the ACT and filing a due process (without help of an attorney) we were successful. It was knowledge of the system that allowed my daughter to prevail. Now, she enjoys going to her neighborhood high school, is mainstreamed for 90 percent of her classes, has been to 1 prom and two balls (complete with date) etc. She has a unique aide who is with her throughout her school day. For us there were two pivotal ingredients in her success, the first being the opportunity to truly integrate, the second and equally important is the support of a unique aid. The reason I point this out is in hopes that you would consider the option of home ownership or even leasing for

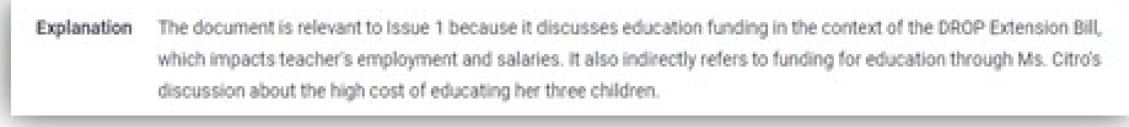
False Positives

Though the False Positives are much lower than the first pass, now we need to look through them to confirm they are accurate.

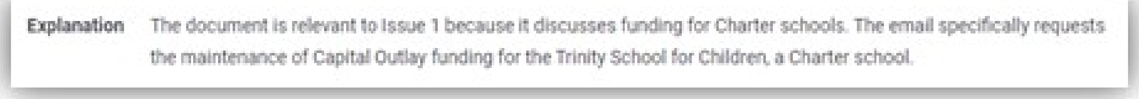
The results confirmed the TREC answer key had missed a lot of Relevant content.

This time, we reviewed every single one.

The lifesaving part was having explanations to show why each document was, in fact, relevant. A quick look at the explanation would point out the relevant portion:



Explanation The document is relevant to Issue 1 because it discusses education funding in the context of the DROP Extension Bill, which impacts teacher's employment and salaries. It also indirectly refers to funding for education through Ms. Citro's discussion about the high cost of educating her three children.

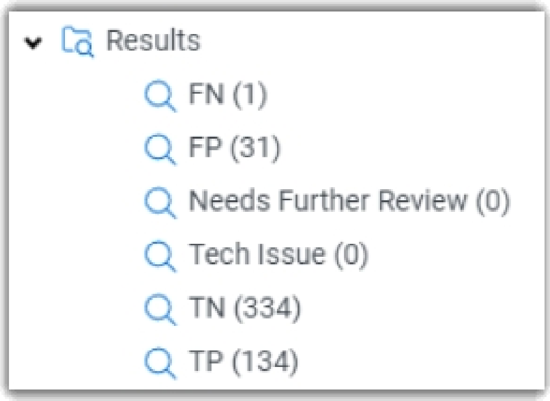


Explanation The document is relevant to Issue 1 because it discusses funding for Charter schools. The email specifically requests the maintenance of Capital Outlay funding for the Trinity School for Children, a Charter school.

True Positives and True Negatives

Again, we are going to shortcut this process a little bit. I looked through a couple of examples and didn't see anything I disagreed with. On a live matter, we would need to sample and review a larger set.

Final Results



Results
- FN (1)
- FP (31)
- Needs Further Review (0)
- Tech Issue (0)
- TN (334)
- TP (134)

We can use these to calculate Recall and Precision for our first run.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} = \frac{134}{(134 + 1)} = 99\%$$

Recall = 99%

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FN})} = \frac{134}{(134 + 31)} = 81\%$$

Precision = 81%

Now, we are at a point where we are comfortable with our instructions.

Conclusion

The remaining step is to run these revised instructions across the rest of the review set. After reviewing the rest, we can use a random sample to calculate our final recall and precision numbers.

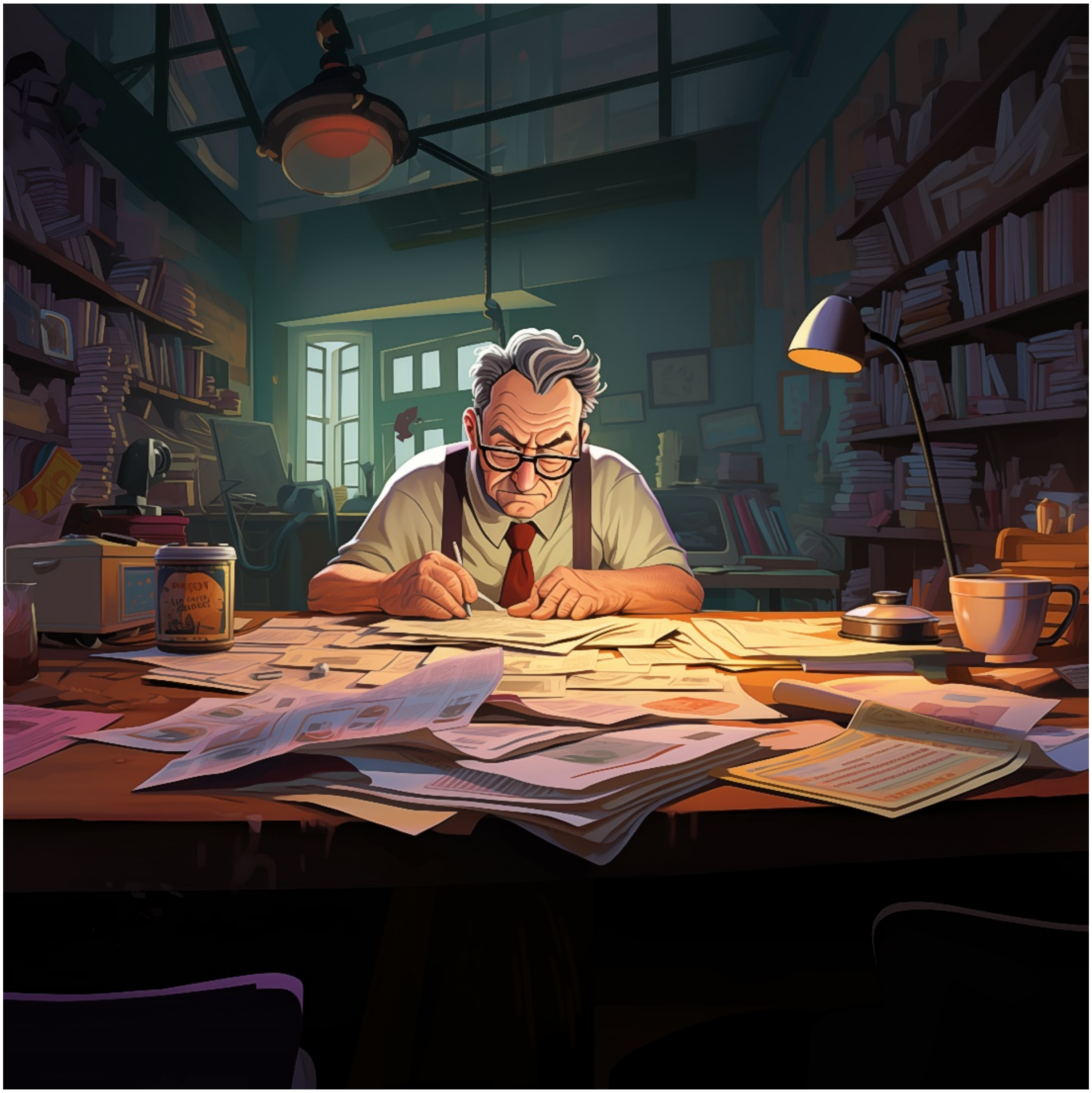
As I know you are probably at the edge of your seat waiting to hear the results of how many people support school funding in Florida, I want to tell you that 89 documents were identified as people writing in to support more funding for schools. Only 3 documents were identified as people wanting to cut funding for schools.

- 1 person opposed creating a new division of the Department of Education.
- 1 person wanted to cut spending across the board, including schools.
- 1 person wanted the old Trenton Middle School gym to be paid with private contributions and donations instead of taxpayer funds.

I think it’s safe to say that taxpayers in Florida generally supported more funds going to schools.

6

WRITING INSTRUCTIONS



Writing instructions for which documents should be considered relevant is a simple part of the process, but it might initially feel slightly unusual.

Here is the easiest way to describe writing instructions:

Write a review protocol as you would for document reviewers. This requires some investigation into the case and relevant parties. Treat the AI like a document review team. Once you are done, paste the description of each issue into the AI as your starting point.

You are basically trying to explain to a person what it is that you want as concisely as possible.

With that being said, here is a quick walkthrough of some best practices we encourage:

Imagine we are working on a case that involves hiring discrimination in the NFL. A team has been accused of discriminatory hiring practices, and we need to produce all documents relevant to that issue.

- **Start with broad instructions**: The purpose of the first round is to understand your dataset, and by being very broad, you can make sure nothing significant is being missed. So, we might start with something like "All documents that include any mention of hiring employees."
- **Review all relevant documents returned**: After the first round, review all relevant documents returned, and use that as a basis for narrowing the scope. There are two different strategies for narrowing:

We can use inclusive language and say, "All documents that include any mention of hiring coaches or management." Or we can use exclusive language and say, "All documents that include any mention of hiring. Only documents that mention the hiring of coaches or management should be considered **Relevant**. Documents that mention hiring administrative staff or players should be considered **Not Relevant**.

- **Narrow the results**: By continuing down the path of exclusions, we can narrow the results to something that aligns with our goals. The result might look something like this:

“All documents that include discussion about hiring coaches and management. Any discussion about qualifications in hiring should be considered **Relevant**. Any discussion about interview processes for coaches and management should be considered **Relevant**. Any discussion about hiring anyone other than coaches or management should be considered **Not Relevant**.”

Notice how we use “any” and “mention” as opposed to something narrower like “All documents that include discussion about hiring employees.” This would return a much narrower result.

Adding inclusion and exclusion criteria can be extremely helpful on stubborn issues.

Don’t be afraid to repeat yourself or restate the same point multiple times. Hammer it home if you must.

Define unclear terms. If you want to find documents related to Project X, start by describing what Project X is and maybe include other things it might be called.

Don’t expect it to make legal judgments or conclusions. If you ask it to find “All documents that might be relevant to an insurance claim,” you can expect that nearly all documents will be returned. Similarly, as mentioned before, instructions say, “All documents that include private personal information” will return just about everything. You are much better off being annoyingly specific, such as “All documents that contain any bank account information, passwords, credit card numbers, social security numbers, or any discussion about a person’s health or medical situation.” It does incredibly well when listing out precisely what you want in a list format.

Just remember, it understands natural language, so you should always talk to it like a person and use clear and direct language.

7

DEFENSIBILITY

Let’s talk about Defensibility!

As everyone would expect, we are hearing some concerns about AI being defensible in predictive coding and people who might be hesitant to try it out. Let’s put it all out on the table.

First, I want to start with what I think are four very important points:

1. We are talking about situations where you produce or withhold one or more documents without human review. If you are using human review but AI to QC or assist in that process, you really don’t have anything to be concerned about. As bad as humans are at reviewing documents, courts have shown they are willing to look the other way when humans make mistakes.
2. Technology-assisted review (TAR) has been used for over a decade. There have been thousands (if not tens of thousands) of matters that have used TAR to produce documents without review, and there hasn’t been a single successful challenge. While it is essential to be diligent and defensible, we must look at the facts and realize that nobody is interested in having this fight.
3. When TAR was first introduced, we jumped from a human-only review to a computer review. That’s a big step. However, now we are jumping from old computer reviews to new computer reviews. That is a much smaller step. If nobody challenged your old process, they aren’t going to want to challenge this one.
4. We are assuming you are using best practices to validate your output before production.

Here’s my theory, which lacks any scientific basis: Lawyers don’t want to challenge the use of Predictive Coding. It’s complicated, and they are afraid of looking bad. The only way to attack the use of Predictive Coding would be through an expert who was familiar with the technology. Now, find any expert in the industry and ask them if they think using Predictive Coding is a reasonable and defensible way to conduct a document review, and they will all say yes. Any expert can attack a poor implementation of Predictive Coding that isn't using proper validation techniques, but it's practically bulletproof with proper validation.

Now that is out of the way, how do we address defensibility? Let's put ourselves in the shoes of someone challenging a TAR protocol. What arguments would you make?

Argument: Producing documents without review is not approved or defensible (sorry, but I had to start with an easy one).

Response: We have been producing documents without review for over a decade. THOUSANDS of cases. Producing documents without human review has been accepted for years by courts and govt agencies. The FTC, DOJ, SEC, you name it. This was being used even before any court approved the use. Since the Da Silva Moore ruling, the courts have affirmatively accepted producing documents without review, so long as it is done correctly. I don't think anyone would seriously argue that computer-assisted review is inappropriate for electronic discovery.

Argument: AI review has not been approved by the courts.

Response: This is false. In Da Silva Moore, Judge Peck states, "By computer-assisted coding, I mean tools (different vendors use different names) that use sophisticated algorithms to enable the computer to determine relevance, based on interaction with (i.e., training by) a human reviewer." He further states, "I may be less interested in the science behind the 'black box' of the vendor's software than in whether it produced responsive documents with reasonably high recall and high precision."

Let's talk about tech briefly: There has never been a single approved "TAR Algorithm." Vendors and tools have used various machine learning models since the beginning. Many popular tools use Support Vector Machines (SVM) as their choice model. Others used a Logistic Regression algorithm. Some vendors even tried to get away with using Latent Semantic Indexing (LSI) and calling it predictive coding (they were much more successful than I expected).

AI tools today are using Large Language Models (LLMs). These machine learning models have differences, but LLMs are significantly better than the others by any measurable standard. I don’t know if I think “you can only use models that aren’t the best” is a solid legal argument.

Argument: The Da Silva Moore ruling states that training is based on interaction with a human reviewer. AI doesn’t work like this.

Response: This is a misunderstanding of AI. AI does work like this. How will the machine know what you are looking for if you don’t train it? We have an entire chapter in this book dedicated to training strategies.

In traditional Predictive Coding, the system is trained by providing thousands of positive and negative examples to help the system differentiate between characteristics that make a document relevant to any given issue. It can’t work if you can’t find enough training examples.

In AI-powered Predictive Coding, the system is trained by providing clear instructions to help the system understand what is relevant.

Are you going to argue a system must be difficult to be defensible? It can’t be good if nobody bleeds or cries? If a lawyer isn’t able to bill at least 60 hours to train, it’s not acceptable?

If you have made it this far, you’re probably invested enough in this space to know that even though this is a fun exercise, none of the above arguments matter. **The only thing that matters is how you validate the results and demonstrate high-quality output**. If someone does an excellent job with validation and can show solid results, you probably aren’t going to win even if a bunch of 3rd graders did the training.

Let’s move on to arguments you would actually use to win this type of claim:

Argument: AI is an unknown technology and is susceptible to hallucination.

Response: Look me in the eye and tell me you know more about Logistic Regression algorithms than you do about Large Language Models. Can you explain Logistic Regression to a judge but not LLMs? It doesn’t matter so long as the results are promising.

Unlike traditional Predictive Coding, LLMs are less of a black box because they can explain every classification to demonstrate how it came up with the categorization decision.

Hallucination is a real issue with AI, but it just isn't a significant factor here. With AI Review, we are using LLMs to classify a document (Relevant, Not Relevant, Needs Further Review, and Tech Issue). Hallucinations are more common when you are generating content rather than categorizing documents. In the worst-case scenario, a hallucination could only result in a misclassification of a document. And because we are validating the results, we are able to confirm that this isn't an issue.

I find it funny that hallucination is seen as a problem with AI because if you have done even an hour of document review, you will know that when you look at document after document after document, hallucination is practically a job requirement. Humans can't do that for 8 hours straight without a little hallucination.

Argument: AI isn't good enough at classifying documents to replace humans.

Response: Oh, my sweet summer child. Just wait until you see how good it is. I've seen many people defend 70% recall and 50% precision. I don't think anyone will have trouble defending scores in the 90s. This actually might open the door to the opposite effect. How will you defend 70% recall by a human review team when tools like this exist?

So, what does a defensible AI Review look like? It's a lot like any Predictive Coding review. We need to use sampling to validate the results. Let's walk through how we can do that. The general process for predictive coding has become pretty straightforward:

1. Identify the review set.
2. Train the machine.
3. Run the documents through the classifier.
4. Evaluate the results.

Believe it or not, it's no different with AI.

Let's get started!

Step 1: Identify the Review Set

We need to know which documents need to be reviewed. That should be the easy part. Then, we need to identify any documents that are bad candidates for review:

- Documents with no extracted text
- Audio files
- Images
- Huge files

For the most part, this has become an optional step. If you fail to pull out these types of documents, they are just going to get flagged as Needs Further Review or Tech Issue, but it's still a good process, so let's pull them out.

Note that I don't mention anything about foreign languages. AI doesn't care what language the document is in, so they can all be reviewed together regardless of language.

When this step is complete, we should have one single folder of the documents that need to be reviewed.

*We have AI tools to review audio and images, but you need to run those separately because the process for instructions is a little different.

Step 2: Train the Machine

While you may want to queue up the Rocky theme for this, you will find the process is pretty darn simple.

We just need to describe to the machine what we are looking for in plain, natural language. For example:

> *"Find all documents where any of our executives are discussing how to price widgets."*
>
> *"Identify any documents where a discussion took place about the company's retirement plans."*
>
> *"We are looking for any documents where an employee of Acme said something inappropriate to John Smith."*
>
> *"Can you find any documents where someone makes a statement properly suggesting we should discriminate against people who support the Green Bay Packers?"*

For a more detailed analysis of how to craft proper instructions, see Chapter 6.

We might need to take 2 or 3 iterations to come up with the perfect instructions, but through some trial and error, it will start to come naturally. Once you have your instructions, you can run them over all your documents.

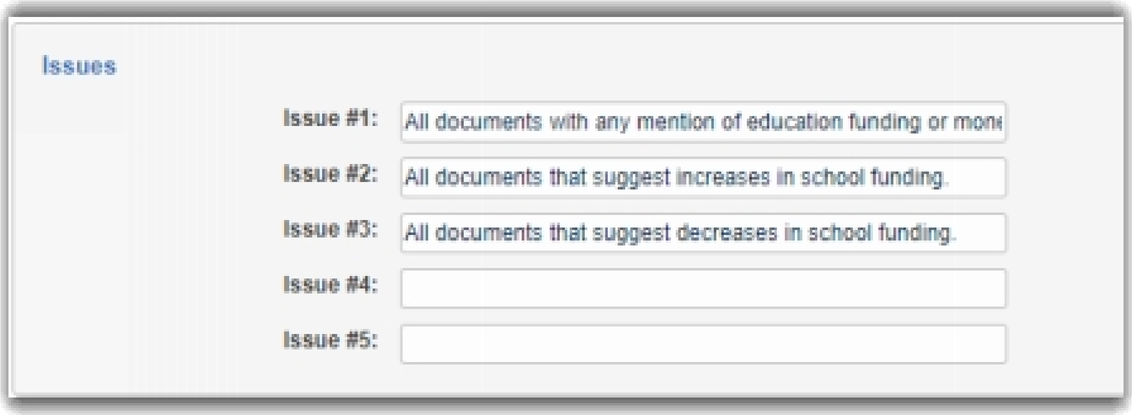
Step 3: Run Your Documents

REVIEWING DOCUMENTS

Please Hold

This part is as easy as it gets.

We simply take the instructions you crafted in Step 2 and run them across all the documents in your review set.



Issues

Issue #1:	All documents with any mention of education funding or mone
Issue #2:	All documents that suggest increases in school funding.
Issue #3:	All documents that suggest decreases in school funding.
Issue #4:	
Issue #5:	

Step 4: Evaluate Results

Everyone, get out your calculators! And no spelling funny words upside-down.

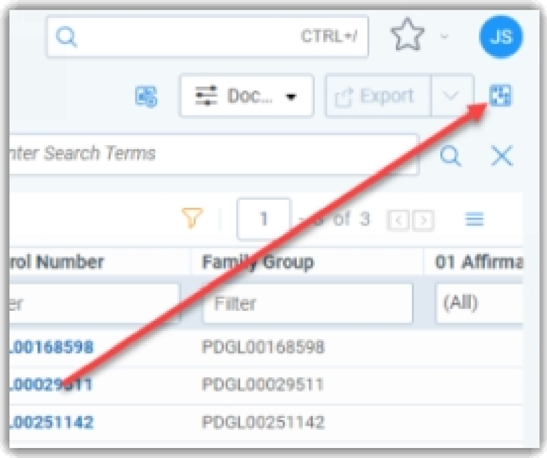
Luckily, this is an open-book test.

This step aims to determine the quality of the classifications made by AI. So how do we do that?

First, we need an answer key—something we can use to score the test. Without an answer key, we won't know if the classification made by AI is right or wrong.

I have some bad news: You need to create the answer key!

We need to take a set of documents that we know are correctly classified and compare them to the results of the AI classification. The only way to do that is with a control set using a subject matter expert.



rol Number	Family Group	01 Affirma
er	Filter	(All)
.00168598	PDGL00168598	
.00029511	PDGL00029511	
.00251142	PDGL00251142	

Let’s do it! Find the set of documents that need to be reviewed from Step 1 and generate a random sample. We don’t need to get cute with stratified or fancy sampling. Just use whatever generic sampling tool is available in your review platform.

The size of the random sample is a hot topic.

In the early days of Predictive Coding, people were terrified of being challenged, so they took some extreme steps to ensure defensibility. Many would take random samples of nearly 10k documents!

Luckily, calmer heads have prevailed, and the suggested size has gone down considerably.

We still suggest people take a statistical sample rather than something fixed.

You can calculate sample sizes for your dataset here:
https://www.calculator.net/sample-size-calculator.html

But to save you some work, here is all you need to know:

> *A sample with 95% confidence and a 1% margin of error will return about 10,000 documents.*
>
> *A sample with 95% confidence and a 2% margin of error will return about 2,400 documents.*
>
> *A sample with 95% confidence and a 3% margin of error will return about 1,000 documents.*

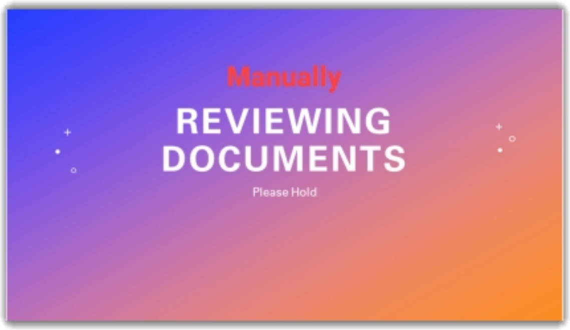
We used to recommend a 95/2 sample a long time ago but have found 95/3 to be sufficient for most cases.

Our Advice: If you can confidently describe what it means to have 95% confidence with a 3% margin of error, you are cleared to go with a 3% margin of error or greater. If you don't know or care what those words mean, go with 2% to be safe.

So now we have our sample, and we have identified a subject matter expert to review it.

It's time to review!

Have your subject matter expert look closely at every document in the random sample and have them determine which are Relevant and Not Relevant. This is a particularly important job, and the coding needs to be precise. If your answer key is all wrong, you will not have a good time.



Manually

REVIEWING DOCUMENTS

Please Hold

This step isn't complete until every document in your random sample has been classified as Relevant or Not Relevant.

Now, we can grade the results!

While this could be its own blog, the quick and dirty is that we need to determine two things:

1. ***Recall**: What percentage of all relevant documents were captured by our AI review?*

2. ***Precision:*** *What percentage of the documents deemed relevant by AI were actually relevant?*

To do this, we use a confusion matrix. Don't worry - the name is the most confusing part.

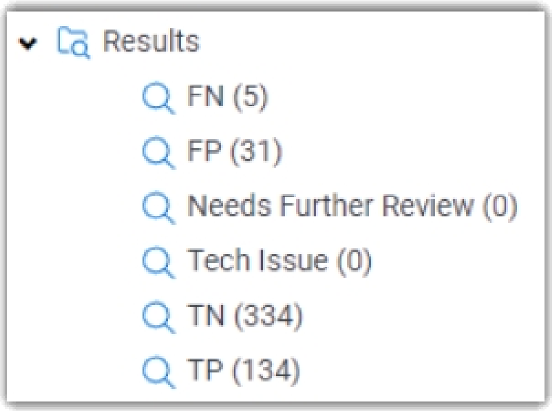
Here is a quick reminder on how to calculate the confusion matrix:

True Positive = Instances where the AI classified the document as Relevant, and the Subject Matter Expert also classified the document as Relevant.

True Negative = Instances where the AI classified the document as Not Relevant, and the Subject Matter Expert also classified the document as Not Relevant.

False Positive = Instances where the AI classified the document as Relevant, but the Subject Matter Expert classified the document as Not Relevant.

False Negative = Instances where the AI classified the document as Not Relevant, but the Subject Matter Expert classified the document as Relevant.



Results

FN (5)

FP (31)

Needs Further Review (0)

Tech Issue (0)

TN (334)

TP (134)

We can use these to calculate Recall and Precision for our first run. Let's use these numbers as an example:

$$\textbf{Recall} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} = \frac{134}{(134 + 5)} = \mathbf{96\%}$$

$$\textbf{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} = \frac{134}{(134 + 31)} = \mathbf{81\%}$$

The Big Question

Of course, if you want to know, “What is an acceptable score?” the answer will be “It depends” because that’s the answer to everything.

The process is just as important as the score, so it is vital that you have an excellent subject matter expert and document every step of the process.

That being said, here’s my opinion and definitely not legal advice:

I would not consider recall below 70% to be defensible. The lowest score I consider defensible in any situation is 70% recall and 50% precision.

In my opinion, a “good” score would be 75-85% recall and 60+ precision. With those scores, I can sleep at night. There is no chance of being challenged on the score alone.

I’ve only seen recall above 90% with precision above 75% using traditional predictive coding once, and the case was so straightforward it could have been done with keywords.

However, AI is playing on a different level. We are seeing 90%+ recall and 80%+ precision on every matter. It will be interesting to see if people demand higher scores. If you challenge 90% recall in court, you’ll get your face on the cover of eDiscovery blogs. And not in a good way.

Conclusion

That’s it! With a good process on Step 2, you shouldn’t have any surprises, and we can wrap this up with a nice little bow. Be sure to review the documents excluded in Step 1, plus any Tech Issue or Needs Further Review documents.

8

WORKFLOWS

It’s great that AI can review documents, but we need to know how to use it in real-life situations.

What problems can it solve?

The short answer: **AI can do almost anything a contract reviewer can do**.

By the next edition of this book, I’m sure we will have an entire list of new use cases, but let’s start with some basics.

Method 1: Reviewer QC

- *Skill Required: Low*
- *Time Required: Low*
- *Validation Required: None*

It is safe to say that almost all of us must answer to someone. It might be your boss, or maybe your client or their shareholders. And they are going to have to approve of the use of any new technology. And they’re scared.

So, let’s make it easy!

At this very moment, you are probably aware of, or actively working on, at least one active document review project using contract reviewers. Maybe you are doing a straight linear review, or maybe you are using CAL, but either way, they are all sitting in a room (or maybe they work from home now) and reviewing document after document all day.

How are you QCing their work?

Everyone I know says they regularly do a random sample QC of their reviewer’s documents. Sample a few hundred documents to ensure they aren’t making any significant mistakes.

But I rarely see anyone actually doing that. The standard practice is to maybe do one quick round of QC after the first week and then never again. And the odds of finishing the first round of QC is probably 50/50 at best. Because it is terrible!

Do you know how long it takes to review a random sample of 100 documents for a team of 10 reviewers? TWO FULL DAYS. What if you have 20 or even

100 reviewers?

I’m sure most people do some sampling or QC at a second level, which picks up some errors, but every reviewer knows that if they mark a document as Not Relevant, the odds of it ever getting looked at again are slim.

Ok, great. Now that I’ve complained a bunch, what are we going to do about it?

Let’s use AI!

Scenario: You have a review team of 20 reviewers working on a large review project expected to run for a few months. You manage the reviewers and are in charge of ensuring the quality of their work.

Process: Set a bi-weekly calendar alert. Every time it comes up, you go through the same process:

1. Identify the documents reviewed by each member of the team.
2. Take a random sample of 100 documents from each reviewer.
3. Run the documents through AI Review.
4. Count the number of times each reviewer had a classification that conflicts with the classification made by AI.
5. Report on the number of conflicts for each reviewer.

It may take an hour to set this up and run it the first time, but then it should be seamless. If you use a review tool like Relativity, you can set up dynamic searches to make subsequent runs in no time.

You still have to develop instructions for the AI, but they don’t have to be perfect. The goal of this exercise is to find reviewers who are making consistent mistakes, so you can pretty much copy what you see in the reviewer protocol, and you’ll be great.

It’s worth looking at some documents where the reviewer and AI disagreed. This can help you understand more about how AI works, and once you realize the AI is right every time, you might start asking yourself why use the reviewers at all!

Pros: This straightforward technique significantly adds value to any document review. Make some fancy reports, and you can sell your review services using AI to enhance the quality of your review team. This also gives you and your team a low-effort introduction to AI technology that can be expanded into other use cases.

***Cons**:* You still have to have humans review all the documents.

Summary: If you have difficulty getting a sign-off on an AI review, this is a good way to get comfortable with the new technology. The quality of your review team will increase significantly, and people will consider you a wizard.

Method 2: AI-Powered Linear Review

- *Skill Required: Low*
- *Time Required: Medium*
- *Validation Required: High*

Now, we are getting real. Let's use AI to classify documents!

This is where the real value lies. AI can review documents more accurately, faster, and cheaper than contract reviewers. And we can prove it!

Always remember to trust but verify. We always have to verify that the results are incredible. But it's like a game in many ways, and it's fun to see if you can beat your recall high score!

***Scenario**:* You have a collection of 200,000 documents in the most crucial case of your career, and finding the smoking gun document will make or break it. You need to ensure every document is meticulously reviewed and nothing is missed. Time is of the essence, and you need to get moving ASAP.

***Process**:* Let's run a simple AI review!

Before you begin any AI strategy, you should do your typical early case assessment strategies like clearing out junk and spam—anything to get the volume of documents reduced without risking any relevant documents.

You can also deduplicate your review set and propagate the results to the duplicates. There is no reason to review identical documents multiple times.

Like any review, we must identify relevant issues and develop a strategy. Then, we need to start running some test instructions.

Using the strategies from this book, we can test and refine our instructions until they get great results on a small sample. Since this case is so important, we might want to go through 3 or 4 iterations of instructions until we call things perfect.

We then identify a statistically valid random sample. We will go with a 95% confidence with a 3% margin of error sample, which returns about 1,000 random documents. We use our best subject matter expert (probably you!) and review these documents as carefully as possible. It will take about 2 days to complete the sample review, but it will be worth it because it will save us months of review.

While you are reviewing the sample, AI is doing the same as well. It's a race you won't win, but at least the results will be ready immediately.

Upon completing the review, we can calculate the metrics with AI doing the same. Your coding is considered the answer key, with the test being the document classifications. We use that to calculate recall and precision.

Note that we are performing validation BEFORE the actual review. We call this pre-validation. This allows us to avoid any risks of having AI review a ton of documents and come back with a poor result. In the next chapter, we will discuss some of the ways you can use pre-validation to improve your legal strategies.

Assuming your scores are sufficient, we can then proceed to have AI review the rest of the documents in your dataset. The review will take at most a day or two, and you will then have all your key documents identified and summarized.

There will need to be some cleanup, as there are certain to be some documents that are too large for AI to review (for now) economically. Those can go to a small review team and get started on a Priv review of the relevant families for production.

Meanwhile, you can review all the summaries of the key documents and start preparing your case.

Can you imagine having all your key documents identified and summarized within 3 days of the review getting started!?

***Pros**:* You save an incredible amount of time and money and have the most accurate classifications possible. You are likely far above 90% recall and precision and can feel comfortable that everything has been reviewed nearly perfectly.

***Cons**:* As you are reviewing as many documents as possible, your costs will be higher than with other methods. But still just a fraction of the cost of a human review.

Method 3: AI/CAL Hybrid Review

- *Skill Required: High*
- *Time Required: Medium*
- *Validation Required: Medium*

What if we combined Traditional Continuous Active Learning (CAL) with AI? We can get the benefits of a reduced review set with CAL but avoid the high cost and low accuracy of human review.

Side note: For those who aren't familiar with Continuous Active Learning, I recommend our prior book called "The Book on Predictive Coding", which can be purchased for 99 cents at thebookonpredictivecoding.com. If you don't want to do that, know that continuous active learning uses machine learning and learns from reviewed documents to push the most relevant documents to the front of the review queue.

It turns out the results are incredible!

***Scenario**:* You have to review and produce documents relevant to a Request for Production. You have some number of documents to deal with, between 1,000 and 100,000,000, and you want to review them in the most efficient way possible.

***Process**:* It's just CAL, but with AI doing the review.

As always, do your early case assessment and cull your dataset down as much as possible. Remove duplicates from the review universe for now, but they will be added back in for production.

Using the strategies from this book, we can test and refine our instructions until they get great results on a small sample. It should only take 1-2 iterations to get things going smoothly.

Next, we proceed like any CAL review. If you have any good seed documents, start with those; otherwise, we can start with a simple random sample.

Starting with the seed documents or random samples, we send them to AI for review. Once the results are back, you should quickly look at any documents tagged Needs Further Review and classify them appropriately. For now, you can set aside any documents labeled as Tech Issue. Take just the Relevant and Not Relevant documents, feed those into your CAL model, and run a training round.

At the end of the first training round, all your documents should have scores between 0 and 100. Grab the highest-scoring documents. We usually grab a few thousand documents, but you can choose the size based on how much time you want to spend doing this.

Send those documents through AI using the same instructions as before and feed the results back into the CAL model again.

You can keep this iterative process going until you run out of relevant documents for review. Once you hit a point where you are no longer encountering any relevant documents, you can stop.

Once the review is stopped, we must validate the results. This is done in two passes:

1. To validate the accuracy of the AI review, we need to use a control set to calculate recall and precision for the AI review.
2. To validate the decision to stop the review early, we need to take a random sample of all unreviewed documents to see if anything of merit was unreviewed.

We recommend that a subject matter expert perform both steps.

Like before, we will have to plan on some cleanup of the Tech Issue documents and other outliers. The good news is you can start this process immediately after the first round of documents is completed.

You have now completed the review with a level of accuracy near that of a straight AI review, with significant cost savings. While every review set is different, in our experience, CAL typically reduces the review population by 50% or more, so the savings are significant.

Pros*:* Nearly all the benefits of an AI review, with considerable cost savings. This method should return a result that is nearly as good, just slightly slower, and much less costly than the full AI-powered review option.

Cons*:* As not all documents were considered, it would be impossible to be as accurate as a linear AI review, but the results will still be vastly better than a human review. I also find the process of resubmitting batch after batch after batch to be a little bit tedious, and usually, that results in the review taking quite a bit longer.

9

LEGAL STRATEGIES

The use of AI in review is going to have a significant impact on legal strategies in document review. The biggest factor is going to be the implementation of pre-validation.

What is Pre-validation?

A typical computer-assisted review project works like this:

1. Train the model.
2. Run the model across your dataset.
3. Validate the results with a control set.

This process makes sense in traditional Predictive Coding protocols because there is a high cost of time, money, and effort to train the model but a low cost of time, money, and effort to run the model across a dataset.

With AI, there is a much lower cost of training and a higher cost of reviewing documents. So, we want to ensure we are doing it correctly before we kick off the review.

We do this with pre-validation.

The process with pre-validation looks like this:

1. Refine your instructions.
2. Validate the results with a control set.
3. Run across the rest of the dataset.

This works because, with AI, the documents are reviewed one at a time instead of analyzing the entire corpus together. The order of review does not matter at all.

Why would someone want to pre-validate?

Pre-validation gives us two very powerful new tools:

1. The ability to confirm your results will be accurate before you incur the review cost.
2. The ability to decide when to review the documents.

Suppose you have 3 months to review 200,000 documents. You could take 7 reviewers, working at a rate of 60 documents per hour, and they would finish the review in about 60 days. This will allow for 30 days for Priv and clean up before production.

Seems reasonable. But that’s the old way. Now we have AI!

Using AI with a few iterations of instructions, you could finish the review in 5 days. And now, you have 85 days for privilege review and cleanup, so you are ready for production.

That’s FANTASTIC, right?

What if the case settles right away? Now, you just incurred all the costs of the relevancy review for no reason.

This is where pre-validation comes in.

With pre-validation, we can refine our instructions and prove they are accurate BEFORE incurring any review costs.

How does that look?

Using the same process as usual, a subject matter expert would refine the instructions until they were comfortable with the result. They would then run those instructions across a small random sample of documents and use that sample as a control set to calculate recall and precision. Assuming the recall and precision are acceptable, we can just STOP.

We now know the instructions have been validated and can estimate the result. But it doesn’t make any difference if we review the rest of the documents right away or 10 years from now.

So, you could sit and wait to see if your case settles.

In this hypothetical, I would suggest we wait 60 days. If the case settles, you just saved a tremendous amount of money for your client. Their review costs will be pennies compared to the human review option.

And if the case doesn’t settle? Just click the button to review all the documents using your pre-validated instructions. As long as we are using a static large language model and haven’t changed it since our original validation, the results will be guaranteed.

In the same spirit, what if you have an incredibly important case and need to find all your key docs ASAP? Then, run the documents immediately after validation and have the entire relevancy review with key docs identified in 5 days. That’s a pretty big competitive advantage.

I doubt we have even scratched the surface of potential legal strategies that will benefit from using AI in document review. Still, hopefully, this example can show you one of the benefits to consider.

CONCLUSION

Even though the book isn’t very long, we covered a lot of information pretty quickly. We hope you find this information helpful and that it encourages you to dive into the world of using AI to improve your review workflows.

I hope it is evident that we love this stuff. If you have any questions or comments, or you want to talk over any strategies you are considering, we can always be reached at support@ediscoveryai.com

Happy AI-Reviewing!

BIOGRAPHY

Jim Sullivan is the co-founder of eDiscovery AI. He is an attorney who has worked in eDiscovery since graduating from law school in 2007 and would never miss an opportunity to nerd out about technology. Jim has consulted on thousands of Predictive Coding projects and frequently speaks at legal conferences about the use of technology in eDiscovery. He lives in Minnesota with his wife and a 9-year-old monster.