IMPLEMENTING FEDERAL-WIDE COMMENT ANALYSIS TOOLS

CDO Council Special Projects
Final Recommendations
June 2021

Contact Us: Ted.Kaouk@usda.gov and Christopher.Alvares@usda.gov

TABLE OF CONTENTS

EXECUTIVE SUMMARY

COMMENT ANALYSIS: AI/MACHINE LEARNING PILOT

PURPOSE

The Comment Analysis pilot has shown that a toolset leveraging recent advances in Natural Language Processing (NLP) can aid the regulatory comment analysis process. We developed tools that help comment reviewers identify the topics and themes of comments, as well as group comments that are semantically similar. Tools like these offer significant value by creating efficiencies through novel insights and streamlined processing of comments, reducing duplicative, upfront development efforts across government, and ultimately realizing cost savings for agencies and the USG.

TOOLS DEVELOPED

DEDUPLICATION / SEMANTIC SIMILARITY

Allows reviewers to **focus on unique submissions** and **group similar comments** by identifying and grouping:

1. identical or nearly identical form letters and comments on **exact text match**
2. comments with similar ideas by detecting **semantic meaning of comments**

TOPIC MODELING / CATEGORIZING

Helps reviewers **organize comments for distribution to SMEs** by identifying themes (or categories) in comments under a regulation and **clustering comments into these themes** using unsupervised topic modeling.

FINDINGS

HOW ARE THESE TOOLS DIFFERENT FROM WHAT EXISTS TODAY?

Existing **deduplication** methods detect duplication by comparing **text similarity**. They look for literal matches. Our semantic similarity system does that, but it can also identify **semantic meaning**, catching synonyms and paraphrasing that current systems would miss.

For **topic modeling**, we are using Hierarchical Latent Dirichlet Allocation (HLDA), which can **provide more sophisticated topic groupings** than current categorization systems that rely on keywords and produce results that are too broad.

MODELS BEST-SUITED FOR COMMENT ANALYSIS

Our team found that the BigBird model (released in 2021) outperforms SBERT with RoBERTa (2019) on longer datasets and is best suited for identifying duplicate and semantically similar comments.

HLDA models the relationships between topics as well as the topics independently, producing better topic groupings and visualizations than LDA. We used hLDA implementations released in 2017 with SBERT models released in 2020.

RECOMMENDATIONS

GENERALIZABLE TOOLSETS PROVIDE VALUE

The tools developed were customized from open-source, state-of-the-art models and are now in a generalizable format to be used by any federal agency to improve their comment analysis process. We have not trained models on specific agency language because we have found success in generalizable models that work to a degree of performance across agencies, regardless of domain. Agencies should evaluate results and time savings of the standard toolset.

AGENCY CUSTOMIZATION OF PILOT TOOLSET

There are ways for agencies to customize the generalizable tools further as needed. Through step-by-step instructions included in the full report of final recommendations, agencies can leverage the pilot toolset and further tailor them to fit mission-specific requirements.

POTENTIAL TIME SAVINGS

From a small sample time study for which we are still collecting data, we estimate the time savings for every

1,000 comments

DEDUPLICATION TIME SAVINGS

45 hours

TOPIC MODELNG TIME SAVING

80 hours

EXECUTIVE SUMMARY

COMMENT ANALYSIS: FUTURE CONSIDERATIONS

PROCESS

As part of the CDO Council Comment Analysis pilot, the analytics team met with over 50 stakeholders across 7 stakeholder agencies including OMB and GSA. The team identified the value in continuing development on additional use cases that would build on the accomplishments of this short-term pilot and further improve the comment analysis process using NLP. This includes supporting the USG and individual agencies in tailoring and scaling the pilot's generalizable topic and deduplication/semantic similarity models, developing additional use cases and enhancements identified by stakeholders in human-centered design sessions, and identifying claims of inequity or environmental impact.

FROM PILOT TO PRODUCTION

Further refine and scale the pilot models and prototype

- ✓ The semantic similarity and hLDA models were fine-tuned and performed well using a sample of around 500 comments under each rule from various agencies. GSA and agencies can scale this using steps in the final recommendations report.
- ✓ Agencies can evaluate results of the base toolset and weigh the cost/benefit of customizing it, using steps in the full report.
- ✓ GSA may also consider the opportunity to offer customization of these pilot toolsets as a shared service to agencies.
- ✓ The team also developed a clickable prototype that demonstrates how these tools can be integrated into a comment analyst's current workflow.
- ✓ Further investment in development of the clickable prototype and code should be explored. In the interim, agencies can access GSA's GitHub to leverage and scale to production.

FEATURES TO ENHANCE CURRENT MODELS

Facilitates user-centered design and agile, iterative development

- ✓ **Interactive topic modeling** allows comment processors to nudge and train models in real-time to **fine tune topic labels** to fit agency specific topics
- ✓ Topic modeling in individual comments
- ✓ Automatically generate and assign topic labels
- ✓ Document summarization to apply different AI/ML NLP techniques to **assist comment processors** in generating comment summaries
- ✓ Attachment analysis to help reviewers more quickly analyze and understand attachments
- ✓ Identifying **bot comments** and/or **fraudulent comments** to reduce effort on comment processors
- ✓ Additional use cases and enhancements in full report

EQUITY AND CLIMATE IMPACT ANALYSIS

Enables analytics innovation in support of Administration Executive Orders and priorities.

- ✓ **Identify** stakeholders with interest in and/or relevance to current administration priorities
- ✓ **Supervised topic modeling** of themes re: claims of inequity, bias, disparate impact, climate or environmental impact
- ✓ Uses inputs, labels, crowd-sourced annotators, and potentially synthetically generated text to train model
- ✓ Sentiment analysis/emotion classification to **identify intensity of claims** in comments
- ✓ **Improved understanding of regulations**, programs, and processes that may cause disparate impact to constituents

PROJECT OVERVIEW

INTRODUCTION

THE PILOT

The CDO Council Comment Analysis pilot has shown that recent advances in Natural Language Processing (NLP) can effectively aid the regulatory comment analysis process. The proof-of-concept is a standardized toolset intended to support agencies and staff in reviewing and responding to the millions of public comments received each year across government.

THE NLP TECHNOLOGY

The team leveraged state-of-the-art neural network models based on empirically validated research published in NLP journals and conferences. These models are trained on massive amounts of text and bring with them knowledge of the English language, as well as world knowledge. While they can be further fine-tuned to agency needs, the base models demonstrated a high degree of performance across participating agencies, regardless of technical content or comment type. This flexibility allows any federal government agency to adopt the "generalizable" models and software to improve their comment analysis process.

THE CDO COUNCIL TOOLSET

We developed tools that help comment reviewers identify the topics and themes found within comments under a regulation, as well as identify and group comments that are duplicates or near duplicates and those that are semantically similar, meaning they contain similar ideas. The team also developed a clickable prototype that demonstrates how these tools can be integrated into a comment analyst's current workflow.

Tools like these offer significant value by creating efficiencies through novel insights and streamlined processing of comments, reducing duplicative development efforts across government, and ultimately realizing cost savings for agencies and the broader USG.

THE REPORT

Included is an overview of the base NLP toolset and prototype, as well as recommendations for government-wide implementation and customization.

PROJECT OVERVIEW

PROJECT BACKGROUND

SCENARIO	BACKGROUND	PAIN POINT
The federal government publishes tens of thousands of documents each year in the Federal Register, with over 800,000 total documents since 1994, which garner millions of submissions from the public (comments and other matter presented).	Agencies have a legal obligation to consider all relevant submissions and respond to those which, significantly, would require a change in the proposed rule. To discern relevance, significance, and disposition, human review is needed.	The capacity for human review often can't meet the demand for high-volume comment events. Initial screening and classification allows regulatory officials to focus on relevant submissions and respond to groups of significant comments that address the same topic. Some agencies perform independent, tailored analyses to assist with this initial screening.

OPPORTUNITY

The CDO Council recognized an opportunity to leverage recently advanced Natural Language Processing (NLP), which would be more efficient than these independent analyses. A generalizable toolset could provide effective comment grouping with less upfront effort, and this toolset could be shared and reused by rule makers across government to aid and expedite their comment analysis.

USDA proposed a proof-of-concept pilot project to the CDO Council. Supported by CXO funding, the USDA and CDO Council worked with stakeholder agencies across the government to develop a proof-of-concept NLP toolset that demonstrates how agencies government-wide can benefit from efficiencies in pre-screening and categorization of comments. This would allow rulemaking staff to focus on the most relevant comments and provide single responses to groups of similar comments on the same topic.

This project sought to identify the key challenges and use cases where a generalized solution across agencies would provide the most value by working with several federal agencies and analyzing government-wide data currently on Federal Docket Management System (FDMS), the back-end to public-facing Regulations.gov.

The proof of concept would demonstrate that NLP can be applied to standardize and streamline the comment analysis process across agencies, creating a decision support tool to aid staff in reviewing and responding to public comments.

PROJECT OVERVIEW

CURRENT LANDSCAPE

AGENCY COMMENT ANALYSIS MATURITY

Agencies have made significant progress in leveraging data as a strategic asset to enable effective decision-making. Our team worked with several agencies that have developed processes and tools to support their comment analysis review efforts.

USDA Food & Nutrition Service (USDA FNS)

USDA FNS contracts an outside vendor to review, analyze, and organize the comments received on its proposed dockets. As the organization responsible for Supplemental Nutrition Assistance Program (SNAP), as well as other areas of public interest such as diet guidelines, they receive hundreds of thousands of comments a year, lending to this need. Their vendor produces a report organizing comments and highlighting aspects the agency deems important, such as the topics addressed throughout comments, the number of unique submissions compared to form letters, and whether the comments were submitted by individuals, organizations, legal entities, and more.

Environmental Protection Agency (EPA)

The EPA has a program that performs comment processing, including using deduplication software that distinguishes unique comments from mass mail campaign (or form letter) comments. The tool is accessible to each of the program or regional offices, but beyond that, each program office is responsible for performing its own analysis – and some have developed their own tools to aid this process.

USDA Forest Service (USDA FS)

The USDA Forest Service (USDA FS) developed a system used to solicit comments from the public via webform submissions; track comments submitted via the system or via email or hard copy; track issues raised in those comments and the agency response to them; and more, such as identifying form letters and flagging letters with keywords that require further review.

THE MAJORITY OF AGENCIES WE SPOKE WITH use a combination of Excel, Word, and SharePoint to conduct manual reviews and analysis of the comments they receive, whether they receive thousands (or millions) of comments or as few as 10.

PROJECT OVERVIEW

To ensure they meet legal obligations around responding to comments, solutions across agencies range from..

workarounds to **integrated data analytics**

Piloting cross-agency decision support tools **advances an integrated strategy** for comment analysis across the Federal government.

REGULATIONS.GOV, FDMS, AND NON-PARTICIPATING AGENCIES

The Federal Docket Management System (FDMS) serves as the back-end of the public-facing Regulations.gov, and many agencies utilize FDMS to collect and perform some processing of comments.

FDMS **Regulations.gov**

Managed by GSA

Earlier versions of FDMS included features explored in this pilot, including **de-deduplication** and **auto-categorization**. While de-duplication was discontinued in the most recent release of FDMS, GSA intends to restore that feature at a future date.

Note: Most agencies use FDMS/Regulations/gov. However, several agencies, including **two** of our stakeholders, do not.

Federal Communications Commission (FCC) & **Federal Energy Regulatory Commission (FERC)**

FCC & FERC…

- Maintain their own websites and back-end systems for comment collection and processing; similar to Regulations.gov in function
- Separate comment submissions into two different workflows:
    1) individuals filing shorter comments or
    2) organizations or those filing more detailed submissions.

FCC's workflows direct individuals to one submission form and organizations to another in order to comment on their regulations. FERC requires organizations or those filing longer comments to register for an account before they can submit, and individuals must request a link to the comment submission form.

PROJECT OVERVIEW

Reasons agencies use proprietary systems rather than government-wide shared services, such as FDMS

- FDMS became available after proprietary systems were stood up
- Agency autonomy over features
- Cost

Natural language processing tools, such as those developed in this pilot, have the **potential to be leveraged across participating and non-participating agencies** alike.

PREVIOUS NLP WORK IN THE REGULATORY COMMENT ANALYSIS SPACE

Some organizations have attempted similar efforts to discover how the federal-wide comment analysis process could be aided by machine learning and natural language processing. GSA has conducted studies, and issued reports, with recommendations for how federal rulemaking could be improved. For example, GSA's 10x program explored this topic in 2020 and concluded that the technology was not mature enough, the level of effort was too high, and the regulatory environment posed too much risk to move forward with machine learning for the comment analysis use case. While the CDO Council concurred that there are challenges associated with the regulatory environment, recent advances in NLP with neural network and transfer learning techniques exist that were not considered in the 10x study. For example, some of the models used in the CDOC pilot were released as recently as 2021, allowing for increased accuracy while at the same time lowering model creation cost and effort due to advanced techniques and development of a base toolset. In this pilot, we explored the level of effort required to implement these tools and share recommendations on this in later slides.

IN APRIL 2021, another report prepared for the Administrative Conference of the United States (ACUS) identified agency best practices for handling the challenges of mass, computer-generated, and fraudulent comments. These challenges are enabled by the online public

comment process, and the report offers recommendations for new technologies, coordination and training, docket management, and increasing transparency. This report seeks to align with several of their recommendations.

It is encouraging that many organizations, public and private, have made strides in improving the regulatory comment analysis workflow and are invested in further advancing and aligning these efforts alongside recent advancements in NLP.

PILOT APPROACH AND OUTCOMES

PILOT APPROACH AND OUTCOMES

PILOT APPROACH

Throughout this pilot, we worked with stakeholders from 7 agencies across the government: **The USDA Forest Service (FS), USDA Food and Nutrition Service (FNS), Environmental Protection Agency (EPA), Federal Communications Commission (FCC), Federal Energy Regulatory Commission (FERC), General Services Administration (GSA),** and **Office of Management and Budget (OMB)**.

Some of these agencies participated regularly in stakeholder meetings, informing our understanding of their current comment analysis processes and pain points. Others provided oversight and subject area expertise throughout our pilot activities.

1 Discovery

- Desk research: Prior studies on NLP and comment analysis improvements
- MS Forms Survey to comment analysis SMEs and stakeholders across 9 federal agencies to further understand comment analysis pain points and potential NLP use cases

2 Design

- Human-centered design thinking session to validate survey insights and prioritize pilot use cases
- Requirements gathering and refinement sessions

3 Development

- Iterative development of models, incorporating feedback during twice-weekly check-ins with CDO Council representatives and bi-weekly and ad-hoc sessions with stakeholder agencies

4 Delivery

- Demo of NLP tools/prototype
- Share code repository and instruction package with CDO Council and GSA
- Delivery of final report, including recommendations for further development of the proofs-of-concept and government-wide implementation

PILOT APPROACH AND OUTCOMES

KEY TAKEAWAYS FROM RESEARCH & DISCOVERY

In attempting to further understand the comment analysis process and the use cases that would provide agencies with the most value, we **identified stakeholder pain points** and **areas where an NLP tool could be used** to assist comment reviewers in their analysis. We overlayed these pain points on the five key phases of comment processing, noting that most pain points applicable to this pilot fell in the collecting, sorting, and synthesizing phases: when comments are screened and analyzed.

COLLECTING → **SORTING** → **SYNTHESIZING**

Overwhelming number and length of responses **Handling the number of duplicate/ similar form letters and comments**	**Reviewing and redacting info (e.g., PII) in comments and attachments**	**Needing to sort comments and parts of comments by significance, technical accuracy, topic and/or commenter type**	**System limitations: Inability to analyze attachments, sort correctly, etc.**

"Takes time to sort out non-substantive comments"

"Attachments are an issue because they have to be opened individually, and then converted electronically in order to analyze and redact information."

"Sorting comments by topic; teasing out comments that cut across topics."

"Difficulty tagging type of comment: individual, advocacy group, BOT, campaign, elected official, other."

PILOT APPROACH AND OUTCOMES

In addition to pain points, we learned what comment analysis stakeholders like about their current comment analysis process, what tools they currently use to complete their analysis, and what they consider a "substantive comment" as opposed to a non-substantive one. Survey responses and stakeholder discussions indicated that comments including scientific or economic evidence; specialized, expert knowledge; legal flaws in proposed rules; and alternative solutions or enhancements to the rule, could result in modifying a rule. However, it is important for comment reviewers to assess and confirm the technical accuracy of such evidence or recommendations.

TYPES OF SUBMISSIONS LEADING TO A RULE MODIFICATION

SCIENCE, EVIDENCE, & EXPERT KNOWLEDGE
Comments containing substantiated facts have large influence in changing of rule

REFERENCE TO LAW, REGULATION, & POLICY VIOLATION
Comments pointing to a legal flaw in a proposed law warrant close attention

ALTERNATIVE SOLUTION or ENHANCEMENT
Comments proposing a new solution or suggesting betterments based on new or missed information

SURVEY RESPONSES

"Comments based on significance & impacts of effects back up by best available science of the proposed action."

"Comments that clearly show we have violated or will violate law, regulation and sometimes policy are those most likely to get us to modify a rule, proposed action etc."

"They are relevant, they contain new or nuanced information, they are specific, they suggest solutions/alternatives, they contain supporting rationale."

PILOT APPROACH AND OUTCOMES

In addition to our human-centered discovery efforts, the CDOC analyzed public comment data exported from Regulations.gov for comments made between **October 10, 2019, and March 12, 2021**, which included:

over 5,900 dockets	*with*	**over 2.9 million comments**

These comments represented all government agencies that had rules with comments posted on Regulations.gov within that time period.

After extracting the comment data from Regulations.gov, the team performed several rounds of exploratory data analysis (EDA) to understand the dataset and inform the deduplication and topic modeling algorithms. EDA included running token count distribution, rapid automated keyword extraction (RAKE), tf-idf, Pandas profiling, as well as analysis of comment attachments to identify the types of files and counts of each type. We also tested several tools for the ability to parse attached PDFs. Detailed EDA findings are in the appendix.

PILOT ENVIRONMENT

State-of-the-art NLP models use GPUs (graphics processing units) to train because they can process many pieces of data simultaneously, and they excel at matrix operations, making them a perfect fit for machine learning applications. For this pilot, we used Floydhub, a lightweight cloud GPU platform built specifically for data science teams. For our pilot purposes, the use of Floydhub was very successful because it provides a very low-cost platform to train and run the various NLP models being tested by our team. In the future, for agencies to use the pilot tools in production, the algorithms and code should be shifted to a FedRAMP-compliant environment.

See the Recommendations section for production environment suggestions for government-wide implementation.

PILOT APPROACH AND OUTCOMES

PILOT USE CASE SELECTION

The pain points offered in survey responses informed the use cases that we proposed to stakeholders. We provided stakeholders with nine use case options.

USE CASE	IMPACT
Deduplication of Comments	Using techniques for identifying identical or nearly identical comments, such comments can be removed or grouped. This saves reviewers time by allowing them to focus on unique submissions, especially when there are a high number of form letters. Note: We later added detection of semantically similar comments to this use case, upon model development and testing.
Redaction	By identifying comments containing profanity, obscenities, and threats, such comments can be filtered or redacted. This could also apply to PII and other sensitive information. This will block toxic or sensitive information and reduce the amount of manual oversight required, allowing reviewers to focus efforts on substantive comments.
Categorizing	The aspects identified from text analysis can be used to place comments into clusters or groups based on similarities. Comments with similar ideas can also be grouped. This helps to organize the comments, allowing reviewers to focus their efforts more efficiently during the review process.
Attachment Analysis	Sort comments based on the presence of an attachment. This allows reviewers to more quickly identify SME comments that may have significant impact and require more time-consuming analysis.
Document Summarization	At an individual level, this analysis tool breaks down text into aspects (attributes or components of a policy) and highlights the commenter's opinion toward each. This helps reviewers to see critical points, especially in long responses.
Opinion Identification	Using a text analysis technique, we can identify the specific aspects of a rule or regulation that the public feels strongly toward. This provides reviewers insight into public opinion on an aggregate level.
Identifying Fraudulent Comments	Identifies false information included in comments and/or commenters posing as someone they are not. This ensures comments are authentic and allows reviewers to focus efforts on valuable content.
Identifying Bot Comments	Uses metadata to analyze comments and identify computer programs automatically submitting comments. This ensures comments are authentic and allows reviewers to focus efforts on valuable content.
Identifying Commenter	These techniques help identify commenters based on prior known entries or by identifying likely attributes of individual commenters when previous entries are unknown.

PILOT APPROACH AND OUTCOMES

We conducted human-centered design thinking sessions in which we gathered input and feedback from comment analysis SMEs from seven participating agencies. We discussed the use case options and asked stakeholders to prioritize the proposed use cases and propose additional use cases (found in the appendix) based on their agency's needs. The team conducted additional requirements gathering sessions to further understand and refine the use cases throughout development.

	DE-DUPLICATION / SEMANTIC SIMILARITY	TOPIC MODELING (CATEGORIZING)
USE CASE	• Identify, group, and/or remove identical or highly similar form letters and comments • Identify and group comments with similar ideas by detecting semantic meaning of comments in addition to exact text matches	• Identify themes that emerge from comments under a rule or regulation, without providing pre-set topics (unsupervised topic modeling) • Cluster comments into these themes or categories based on similarities in their wording or ideas
IMPACT	This saves reviewers time by allowing them to focus on unique submissions and group similar comments. Existing methods detect duplication by comparing **text similarity**. They look for literal matches. Our semantic similarity system does that, but it can also identify **semantic meaning**, catching synonyms and paraphrasing that current systems would miss.	This helps reviewers organize comments for distribution to subject matter experts and efficient summarization and response. Our team is using Hierarchical Latent Dirichlet Allocation (HLDA), which can **provide more sophisticated topic groupings** than current categorization systems that rely on keywords.
MODELS SELECTED	• BigBird – released in 2021 • Used with Sentence Transformers (SBERT) – released in 2019; pilot evaluated models released in 2020	• hLDA – first described in 2003; pilot used implementation from 2017 • Used in combination with BERTopic – released in 2020

PILOT APPROACH AND OUTCOMES

At the time of this report's delivery, FDMS includes an auto-categorization feature that is used by some agencies, but which "automatically populates categories of documents based on the key concept of the documents" according to FDMS training documentation. Our survey responses indicated that the feature created categories that were too broad, such as "water" or "United States," according to one agency commenter.

Agencies found significant value in the de-duplication feature from FMDS version 4, which was discontinued in the latest FDMS release. At the time of this report, agencies can request the FDMS Help Desk deduplicate comments and receive a report with the results, as a workaround until the feature is reinstated on FDMS.

Our pilot sought to not only provide a stand-in tool for these features but to improve upon them using neural network tools that were recently released. By using the most advanced available techniques, the tools developed in this pilot can inform FDMS modernization efforts as well as provide access to tools for agencies that do not use FDMS/Regulations.gov or may not have the resources for development of these kinds of tools. Doing so directly lends to the pilot goal of creating base toolsets to advance comment analysis government-wide.

DEVELOPMENT – INNOVATING DEDUPLICATION

The CDO Council toolset modernizes the current deduplication functionality available to agencies by also recognizing semantic similarity. Current deduplication systems detect duplication by comparing text similarity –they look for literal matches in text, which is very beneficial for identifying form letters that are identical or slightly modified. Our semantic similarity tool can detect text similarity, but it can also identify semantic meaning, catching synonyms and paraphrasing that current systems would miss.

Our team explored various open-source models and algorithms that were pretrained to identify identical or nearly identical text. Through this research and testing, we recognized a need for identifying semantically similar text – meaning the tool can identify comment pairs that may not include the same exact text (like form letters) but that include similar ideas, grouping those comments automatically so that reviewers can review the groupings and more quickly distribute them to subject matter experts for review and response.

PILOT APPROACH AND OUTCOMES

To arrive at this conclusion:

1. We explored the top, state-of-the-art language models for Semantic Textual Similarity (the measurement of how close the ideas are in two pieces of text) and Paraphrase Identification (the detection of whether a piece of text is a rewrite of another).

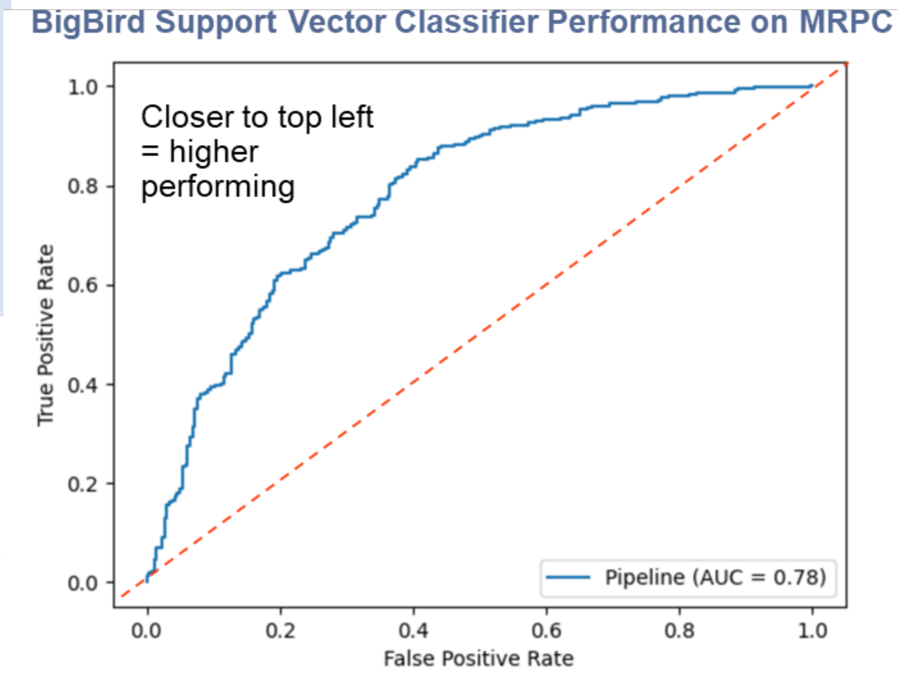
2. We tested several high-performance models, such as SBERT with RoBERTa that was pretrained for semantic text similarity and paraphrase tasks. SBERT with RoBERTa performed well for initial development tasks without additional training. However, RoBERTa (released in 2019) has a length restriction on the size of text it can process. We recognized that other models may better fit the lengthy text found in Regulations.gov comments.

3. We ultimately selected a model called BigBird, which is a newer model (released in 2021) by Google, designed to improve the processing of longer documents through memory management. This lends to it being best suited for longer text, such as lengthy Regulations.gov comments.

4. We fine-tuned the BigBird model on a paraphrase identification dataset called "Microsoft Paraphrase Research Corpus (MRPC)" so it can better find semantically similar text within comments.

5. On MRPC, which has sentence-long data, we found that BigBird outperformed SBERT with RoBERTa using multiple performance metrics (e.g., F1 score, AUC score).

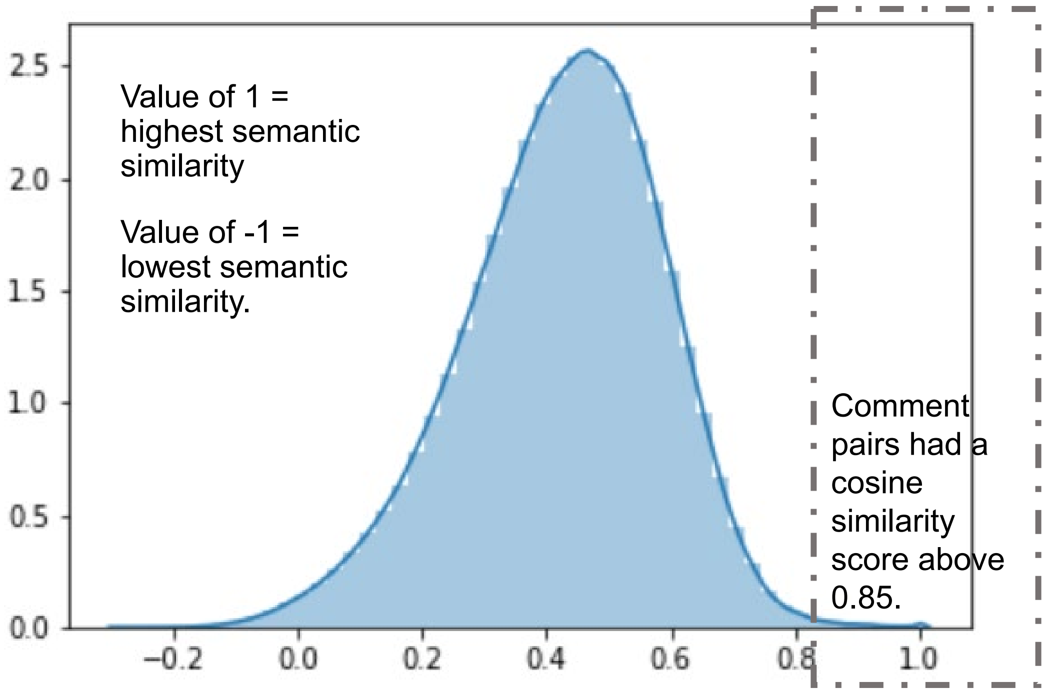
6. On even longer datasets (e.g., lengthy comments on Regulations.gov), due to how each language model deals with long text, BigBird with SBERT appears to outperform SBERT with RoBERTa.



PILOT APPROACH AND OUTCOMES

To compare a pair of text (such as two comments) after the text has been encoded with a language model, we compare them with a metric called cosine similarity.

- Modern language models, such as BigBird and RoBERTa, are helpful within NLP because they turn freeform text into a quantitative value. Text is otherwise a qualitative value, and these language models turn the text into a semantically meaningful vector of numbers. This numeric representation is called an "embedding."
- BigBird and RoBERTa can be combined with another technology, SBERT (released in 2019, though our team evaluated SBERT models released in 2020), to turn groups of text into a single, semantically meaningful embedding, making follow-up analysis manageable.
- Cosine similarity measures the similarity between two embeddings, where similar embeddings have a cosine similarity closer to 1 and dissimilar embeddings have a cosine similarity further away, with -1 being the worst value the metric can score.
- Exact duplicates would have a cosine value of 1.
- When comparing comment pairs, the system produces a cosine metric.
- After testing with sample EPA and USDA comments, we suggest a cosine metric above .85, which indicates that a comment pair is very semantically similar.



PILOT APPROACH AND OUTCOMES

To further understand the cosine similarity metric, take the following example.

“Tokens” are closely synonymous with “words.” One token is differentiated from the next token by white space between them. In a comment pair that is semantically similar but not textually similar, like the one below, the “token” similarity score is much lower than the cosine similarity metric, or “semantic” similarity score.

<u>Token/Word similarity example</u>

“You and I like toast” vs. **“We like toast”**

- 2/6 (33 %) token similarity: “like" and "toast" are similar in both
- Quote 1 has dissimilar tokens: "You", "and", "I" .
- Quote 2 has a dissimilar token: "we“
- **Semantically**, they are very similar. There is a cosine similarity of “.9183,” indicating they are very similar semantically.

For more detail on tokens, see the appendix.

PILOT APPROACH AND OUTCOMES

The example below is an excellent example of semantic similarity, identified by the SBERT with RoBERTa model. These examples were an initial proof of concept that our approach can detect semantic similarity between comments.

Furthermore, we saw similar results with the BigBird model, which we found outperforms SBERT with RoBERTa. The BigBird examples can be seen in the prototype section.

- In the example below, we analyzed a proposed USDA FNS rule on SNAP utility allowances for heating and cooling with approximately 5.9k comments.
- We utilized SBERT using RoBERTa sentence embedding tuned for semantic similarity tasks to identify comment pairs.
- This method also provides the opportunity to identify exact duplicates, which would have a cosine value of 1.

ROW 48

For the people in our state who are struggling the most to buy food and pay utilities, SNAP benefits are already too low. This proposal would lower those already insufficient benefits even more. Does that sound like a humane thing to do? To make life even harder on the people who already have it the hardest? To actively take money away from older adults, people with disabilities and all the most vulnerable people in our state? No one should feel comfortable passing this proposal.

I find it unbelievable that there is consideration of cutting SNAP benefits even further. If enacted, low-income families, those with disabilities, and the elderly would have even more difficulty paying for both food and utilities, which would put them at great risk for poor health outcomes. Our country must stop harming the least of these in order to benefit the wealthy.

ROW 52

Please do not cut funding for SNAP or reduce or undermine the ability for families to qualify for SNAP. SNAP is a critical resource for needy families and cuts to the program would cause harm to the children and families who rely on SNAP. Cuts would disproportionately impact the elderly and people with disabilities. The proposed cuts are unacceptable and dangerous.

Reducing SNAP benefits will harm families and children. Do the right thing!

PILOT APPROACH AND OUTCOMES

Using the BigBird model, this deduplication tool could **save reviewers time** by allowing them to **focus on unique submissions** and **group comments with similar meaning and ideas**. This way, a group of semantically similar comments can be routed to a subject matter expert more quickly for their review and response. This streamlines the process by allowing the comment processor to review groups of similar comments.

The processor would still be able to review and respond to each comment if they choose but could also respond to groups of semantically similar or duplicate comments at once. The semantic similarity function builds on and differentiates from existing deduplication tools or previous FDMS functionality; it is a key capability that stakeholders confirmed would be helpful in their current process. See the Prototype section for how these results are incorporated into a comment analyst's workflow.

DEVELOPMENT – ADVANCED TOPIC MODELING

The most applicable technique, based on stakeholder interest in surfacing themes at the regulation level without pre-labeling, is unsupervised hierarchical topic modeling (using hierarchical Latent Dirichlet Allocation, or hLDA). hLDA allows us to identify the hierarchy of themes, improving the identification of distinct topics by producing a tree structure of parent/child relationships.

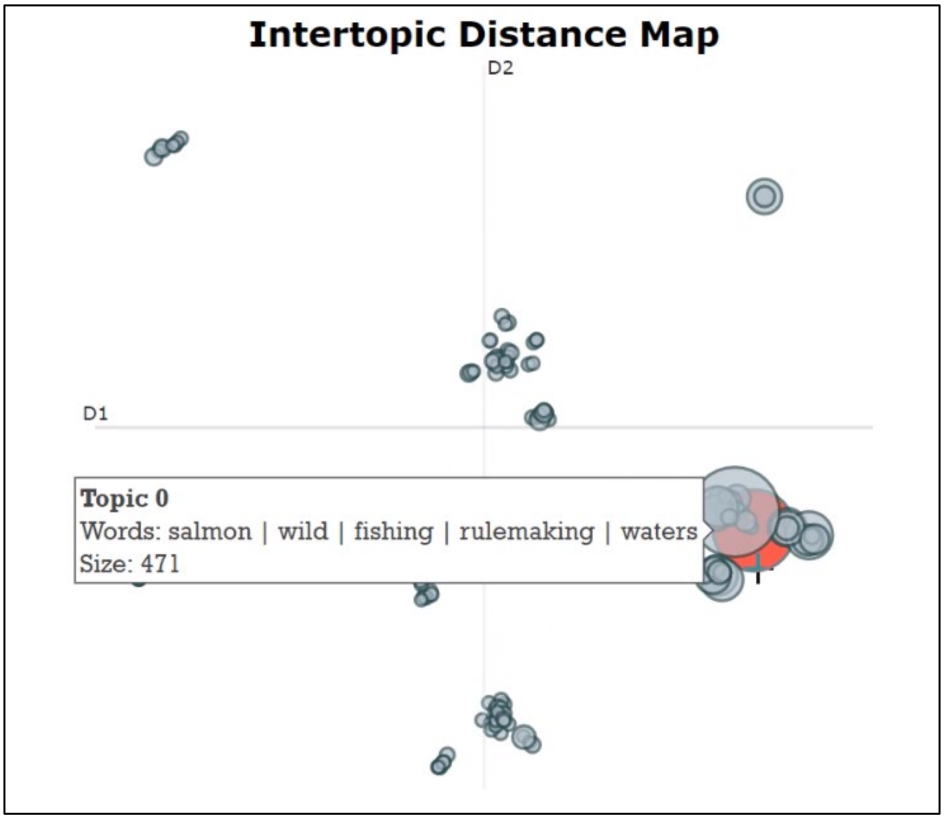
hLDA was first described in 2003, although the team leveraged an open-source implementation released in 2017. Additionally, it was tuned in combination with BERT pretrained word embeddings (released in 2020).

- We explored both open-source, pretrained language models and models that we trained ourselves on regulatory comments to determine the vocabulary challenges with splitting topics.
- The BERTopic model selected was pretrained on the entirety of English-language Wikipedia and Brown Corpus, which is comprised of open-source books. These are standard language models used in the NLP field to convert text into vectors (numeric values).
- The team developed automated workflows to deploy the models, import and clean data, run the models, and extract results on both the agency level and the regulation level to better examine the variations in topics within and across Regulations.gov dockets.

PILOT APPROACH AND OUTCOMES

The results of the hierarchical topic models are groups of topics that allow reviewers to easily identify initial themes in comments, as well as group comments for distribution to subject matter experts for review and response. The results were incorporated into a tool that shows how comment processors can use these results in their workflow (see the Prototype section).

An example of the visual output of the hLDA topic model is below, showing many distinct topics clustered into 5-6 broader topic groupings. The team performed unsupervised topic modeling using BERT sentence embeddings on a random sample of 5,000 comments from a USDA Forest Service Oil and Gas Resources rulemaking docket.



Currently, the data scientist selects the number of topics that the comments are grouped into by assessing how distinct the topics are when there are 4, 6, 8, and 10 groupings (for example). In future efforts, this could be improved by creating an option for the comment reviewer to select the number of topics once they assess the results and how distinct the topic areas are that are identified by the tool.

The team arrived to hLDA by first using Latent Dirichlet Allocation (LDA). For more details on our initial results using LDA, see the appendix.

PILOT APPROACH AND OUTCOMES

IDENTIFYING CLAIMS OF INEQUITY, BIAS, OR DISPARATE IMPACT

In an interest to further the goals of the Executive Order on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government, the team created a list of keywords, or a "seed list," that might be found in a comment claiming disparate impact caused by the proposed regulation. These keywords were selected from the text of the Executive Order and included a shortlist of words that were intended to maximize precision and reduce polysemy (the coexistence of many possible meanings for a word or phrase). This is best practice for designing a list of words that you intend to become automated.

The team used this list to scan the results of the stakeholder agency topic models for keywords that might indicate the comments contained **claims of inequity, bias, or disparate impact**. They placed a flag every time a comment included one of the seed list words and then arranged the results of the topic models to identify the frequency of flags within topics.

The team then used an open-source tool called OptimSeed, which uses **query expansion** to create a larger seed list by automatically nominating new terms for the list and removing ones from the original list that did not produce the intended results.

A larger seed list improves **recall**, meaning the results are less likely to miss anything that could be related to claims of inequity. However, it also increases **polysemy** (when words have two or more distinct meanings), which decreases the **precision** of the results. That said, OptimSeed also removes unhelpful words from the list. Query expansion therefore refines an arbitrary list of words into one that more accurately returns the desired results.

polysemy	the coexistence of many possible meanings for a word or phrase
precision	the % of documents which are relevant
recall	the % of relevant documents that are successfully retrieved

Future equity analysis initiatives, considerations, and ideas can be found in the appendix.

PILOT APPROACH AND OUTCOMES

PROTOTYPE

The team developed a front-end, clickable prototype to demonstrate how the deduplication, semantic similarity, and topic modeling tools developed in this pilot could be integrated into a comment analyst's workflow as a decision support system. Informed by comment analysts' pain points and requirements, the prototype incorporates user-centered design into the visualization of these NLP tools.

Although the following slides demonstrate the deduplication/semantic similarity and topic models separately, ideally these features would be integrated into one fluid workflow with an order of operations. A possible workflow could look like:

1. Run the models so that the tool first identifies duplicate, or nearly duplicate, form letters and mass mail campaign comments.
2. These comments can be grouped and/or sorted from the queue of comments to be reviewed.
3. Then, semantically similar comments could be identified and grouped.
4. The topic model could run on the remaining pool of comments to attribute a topic to the groupings or create new groupings based on the findings of the hierarchical model.
5. The groupings could then be distributed to the appropriate subject matter experts for additional review and response.

This decision support system (DSS) could be integrated with FDMS, and any interested agency could leverage the code for integration with their respective tools. Upon integration, the tool could run nightly (or on another specified interval) to pull new comment submissions into the user interface, which would display upon user login to the tool.

Because regulations have very domain-specific information, concepts and relationships, human review is still required to discern whether and how a comment is relevant to the rule being commented on. This prototype shows how an initial pre-screening tool can allow regulatory SMEs to more quickly focus on the most relevant comments and respond to a group of similar comments on the same topic.

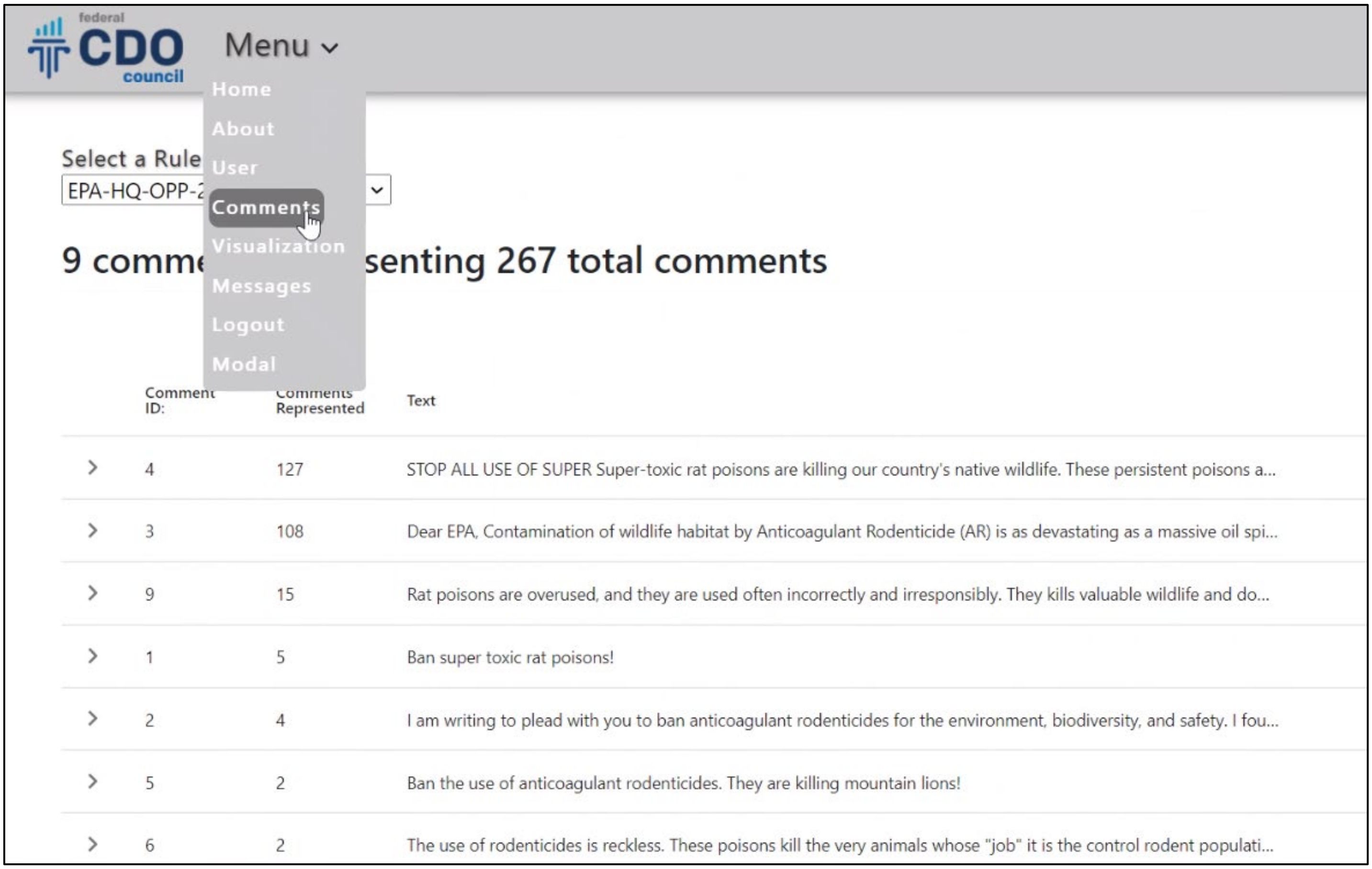
We recommend further exploration of this prototype, workflow, and the supporting models. Future efforts could include validation and evaluation of the prototype and models through end-to-end testing with stakeholder agencies across government.

PILOT APPROACH AND OUTCOMES

Given the short timeframe of the pilot, not all features of a fully production-ready tool were incorporated into the prototype. For example, in the menu below, the Comments tab would be renamed to reflect the Deduplication and Semantic Similarity Visualization. In addition to enhancements noted in the following slides, we also recommend further investment in user-focused iteration on the prototype.

DEDUPLICATION AND SEMANTIC SIMILARITY VISUALIZATION

In the current prototype, upon login, the user has access to the following menu. The user can select the Deduplication and Semantic Similarity tab (here, called the Comments tab) in order to begin the recommended workflow steps.



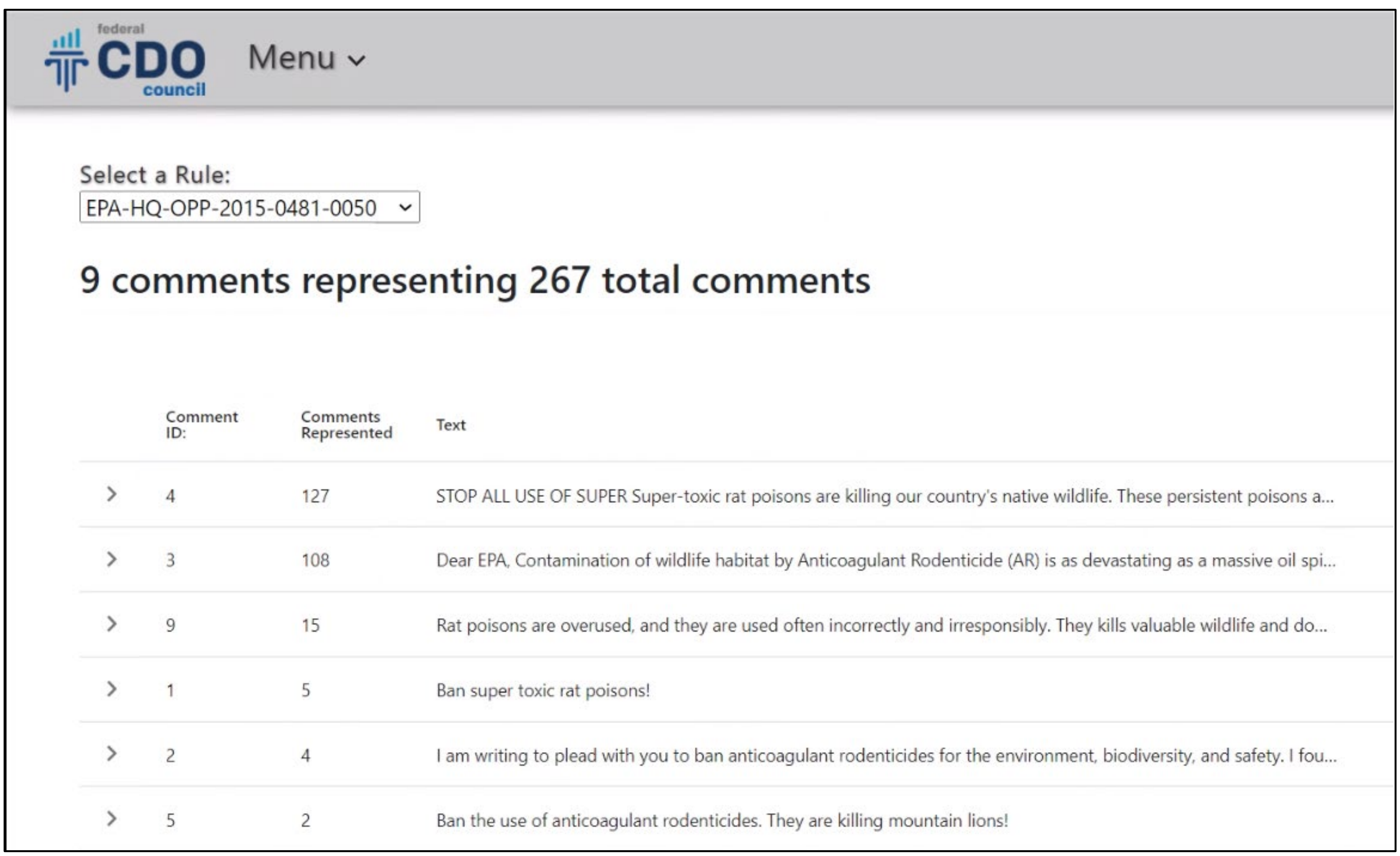
PILOT APPROACH AND OUTCOMES

VIEWING DUPLICATE AND SIMILARITY GROUPINGS

Once the comment analyst or docket manager selects the Deduplication and Semantic Similarity menu tab, they are brought to the screen below. This feature allows them to select a rule and quickly view groups of comments that have been identified as semantically similar.

The semantic similarity model groups comments based on a cosine similarity score, discussed in earlier slides. Recall that a cosine similarity score analyzes the similarity of sentence embeddings to assign a score between -1 and 1, indicating what degree of similarity any two comments have. This leads to the groupings shown below.

The user can see that in this example, there are 9 comment groupings that represent a total of 267 comments. The first row, for example, calls attention to one comment whose meaning is similar to 127 other comments that it represents.



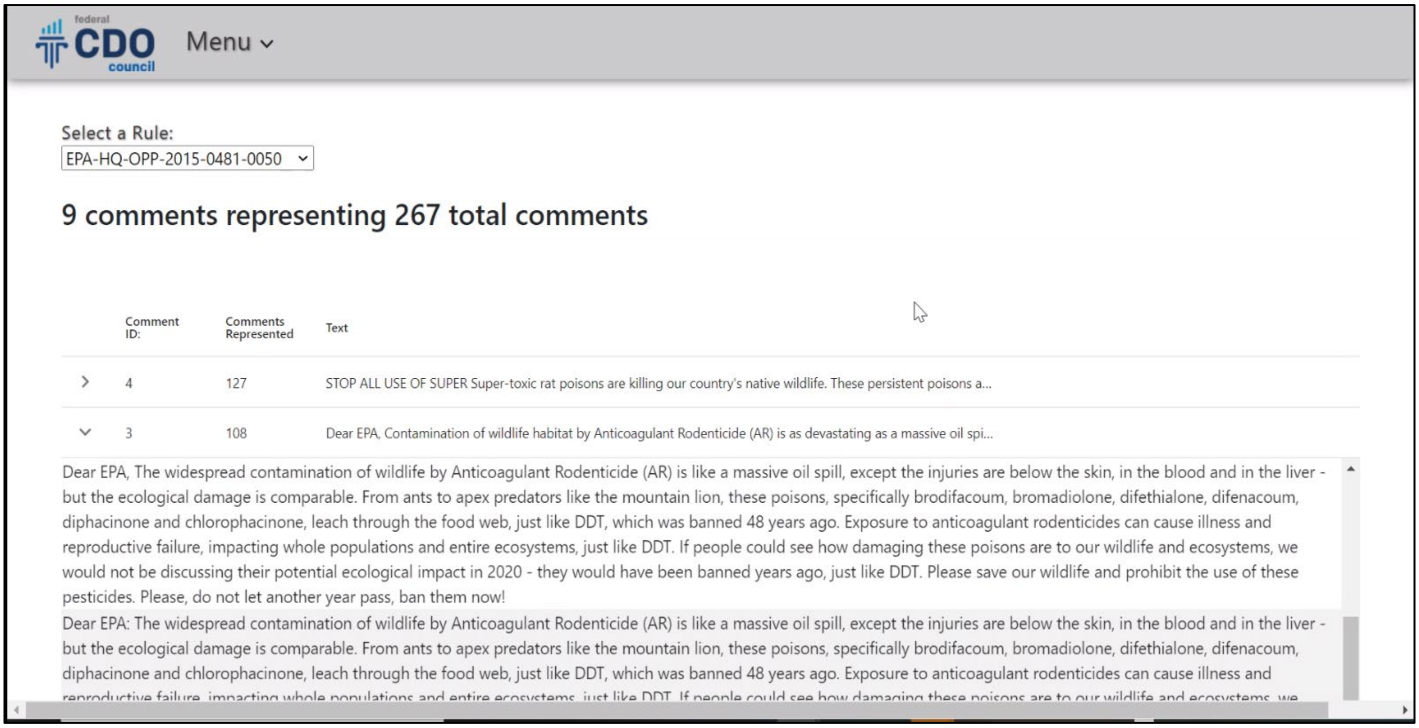
Comment ID:	Comments Represented	Text
4	127	STOP ALL USE OF SUPER Super-toxic rat poisons are killing our country's native wildlife. These persistent poisons a...
3	108	Dear EPA, Contamination of wildlife habitat by Anticoagulant Rodenticide (AR) is as devastating as a massive oil spi...
9	15	Rat poisons are overused, and they are used often incorrectly and irresponsibly. They kills valuable wildlife and do...
1	5	Ban super toxic rat poisons!
2	4	I am writing to plead with you to ban anticoagulant rodenticides for the environment, biodiversity, and safety. I fou...
5	2	Ban the use of anticoagulant rodenticides. They are killing mountain lions!

PILOT APPROACH AND OUTCOMES

EXPAND GROUPINGS TO VIEW INDIVIDUAL COMMENTS – DUPLICATES OR NEAR-DUPLICATES

The user can select the carat to expand the groupings and view all of the comments in that grouping. Doing this allows the user to quickly identify whether the comments are duplicates or near duplicates (e.g., form letters or mass mail campaigns) or similar in meaning.

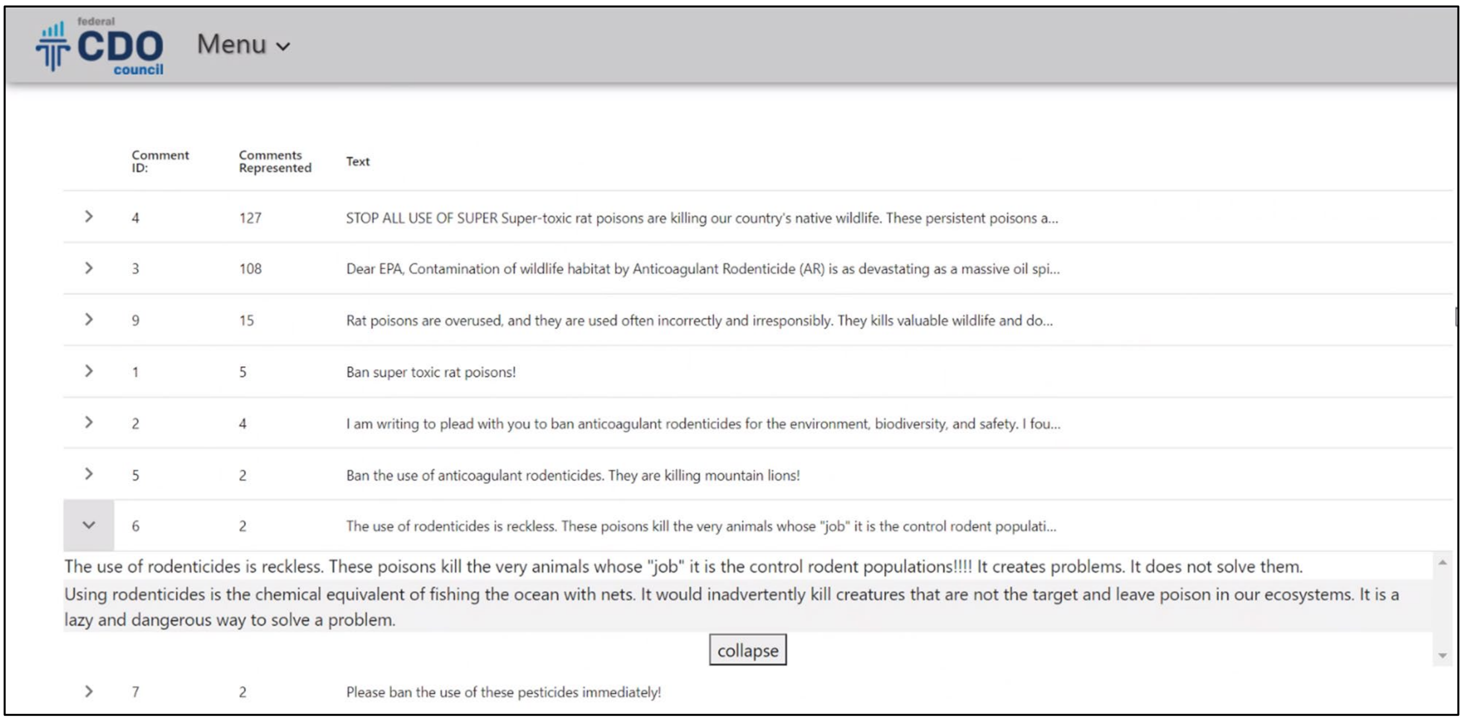
In this example, there are 108 comments that appear to be examples of a form letter. This demonstrates the tool's ability to identify exact duplicates or near-duplicates such as form letters.



PILOT APPROACH AND OUTCOMES

EXPAND GROUPINGS TO VIEW INDIVIDUAL COMMENTS – SEMANTIC TEXT SIMILARITY

Alternatively, this example is of a semantically similar grouping. Identifying not only text matching but also similarity in meaning of comments is a novel feature that, to the knowledge of our stakeholders and SMEs, had not been applied previously to comment analysis. After seeing a demo of the prototype, stakeholders identified this feature as adding value to their comment analysis process.



Comment ID:	Comments Represented	Text
4	127	STOP ALL USE OF SUPER Super-toxic rat poisons are killing our country's native wildlife. These persistent poisons a...
3	108	Dear EPA, Contamination of wildlife habitat by Anticoagulant Rodenticide (AR) is as devastating as a massive oil spi...
9	15	Rat poisons are overused, and they are used often incorrectly and irresponsibly. They kills valuable wildlife and do...
1	5	Ban super toxic rat poisons!
2	4	I am writing to plead with you to ban anticoagulant rodenticides for the environment, biodiversity, and safety. I fou...
5	2	Ban the use of anticoagulant rodenticides. They are killing mountain lions!
6	2	The use of rodenticides is reckless. These poisons kill the very animals whose "job" it is the control rodent populati...

The use of rodenticides is reckless. These poisons kill the very animals whose "job" it is the control rodent populations!!!! It creates problems. It does not solve them.

Using rodenticides is the chemical equivalent of fishing the ocean with nets. It would inadvertently kill creatures that are not the target and leave poison in our ecosystems. It is a lazy and dangerous way to solve a problem.

collapse

Comment ID:	Comments Represented	Text
7	2	Please ban the use of these pesticides immediately!

PROTOTYPE UI ENHANCEMENTS

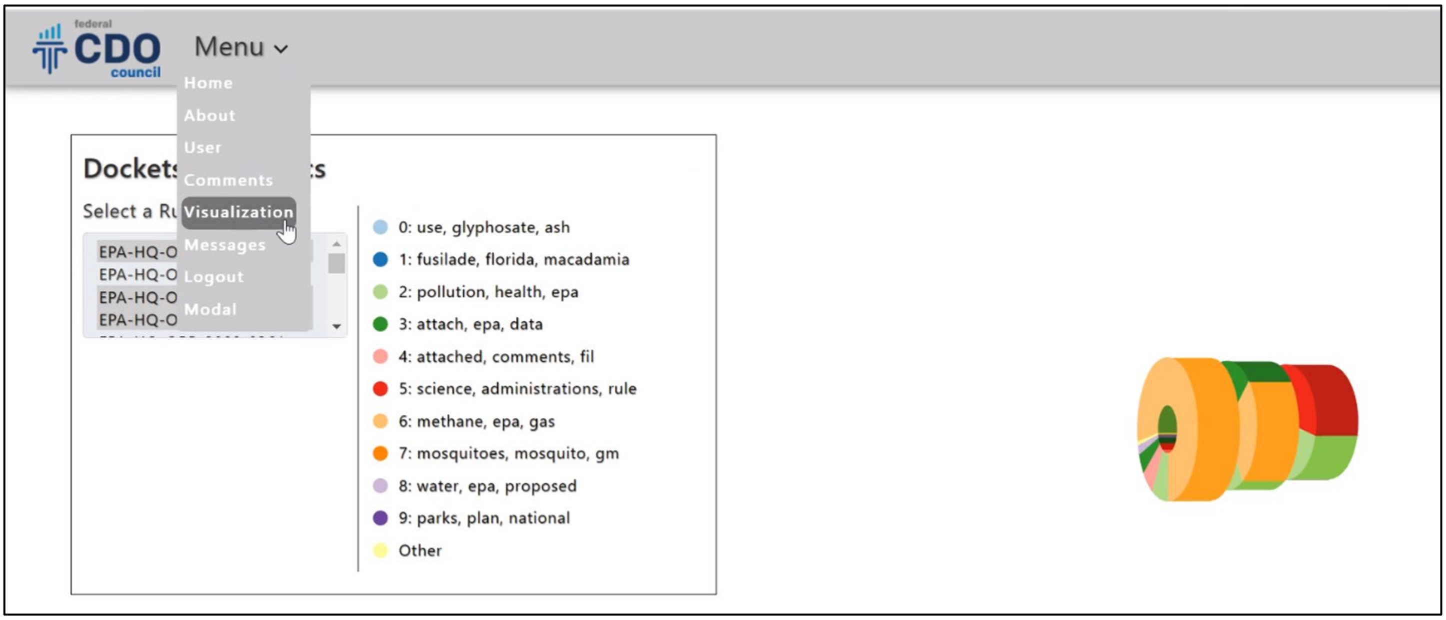
For the tool to be production-ready, we recommend the following enhancements:

- Display the cosine similarity metric of the comment to the “representative” comment or topic, in order to quickly determine and/or sort duplicates, near duplicates, and semantically similar groupings.
- Like prior deduplication tools, display deltas between nearly duplicate comments.
- Allow the user to tag groupings as the type of grouping, e.g., form letters/mass mail campaigns, in order to remove them from view or group/consolidate them.
- Add routing of semantically similar comments to subject matter experts (see the Topic Modeling Visualization for details on how this could be implemented).

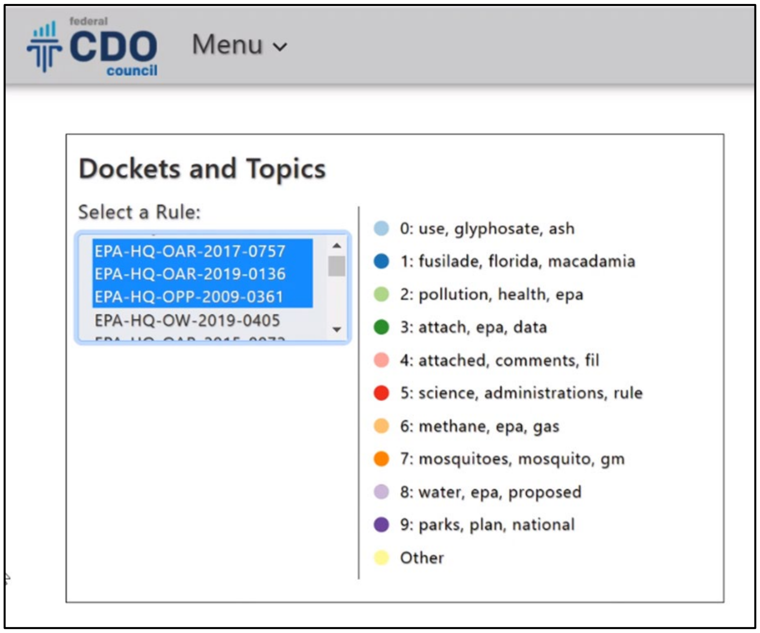
PILOT APPROACH AND OUTCOMES

TOPIC MODELING VISUALIZATION

After running the deduplication and semantic similarity feature, the comment processor can select the Visualization tab (to be renamed to Topic Model Visualization) in order to view and sort the topics identified within a regulation(s).



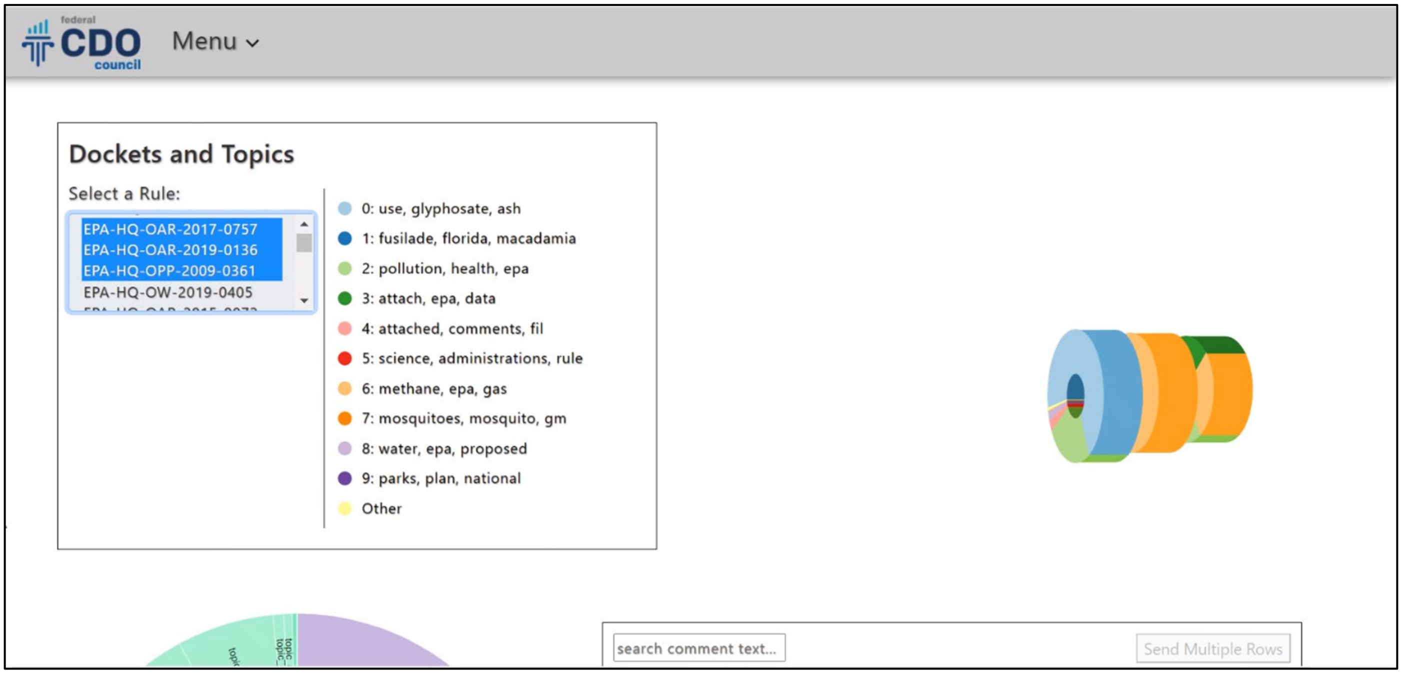
In the Select a Rule section, the user can select the regulation, or regulations, for which they would like to view topics. This serves several use cases: A docket manager may be focused on an individual regulation. A comment lead or someone working with multiple regulations can view all of the regulations in the comment review pipeline; they may be interested in viewing the overlap in topics between those regulations or assigning reviews to SMEs for multiple regulations on similar topics.



PILOT APPROACH AND OUTCOMES

VIEWING TOPIC AREAS PRODUCED BY THE MODEL

The colors, numbers, and keywords listed in the table to the left represent the top 10 topic areas that appear across all regulations pulled into the system.



This table is currently static, meaning that it reflects, for example, the top 10 topics found in all of EPA's current comments in the tool. Future enhancements would be to make this topic table dynamic, so that when the user selects individual regulations, the topic list changes based on the selected regulation(s) for comparison.

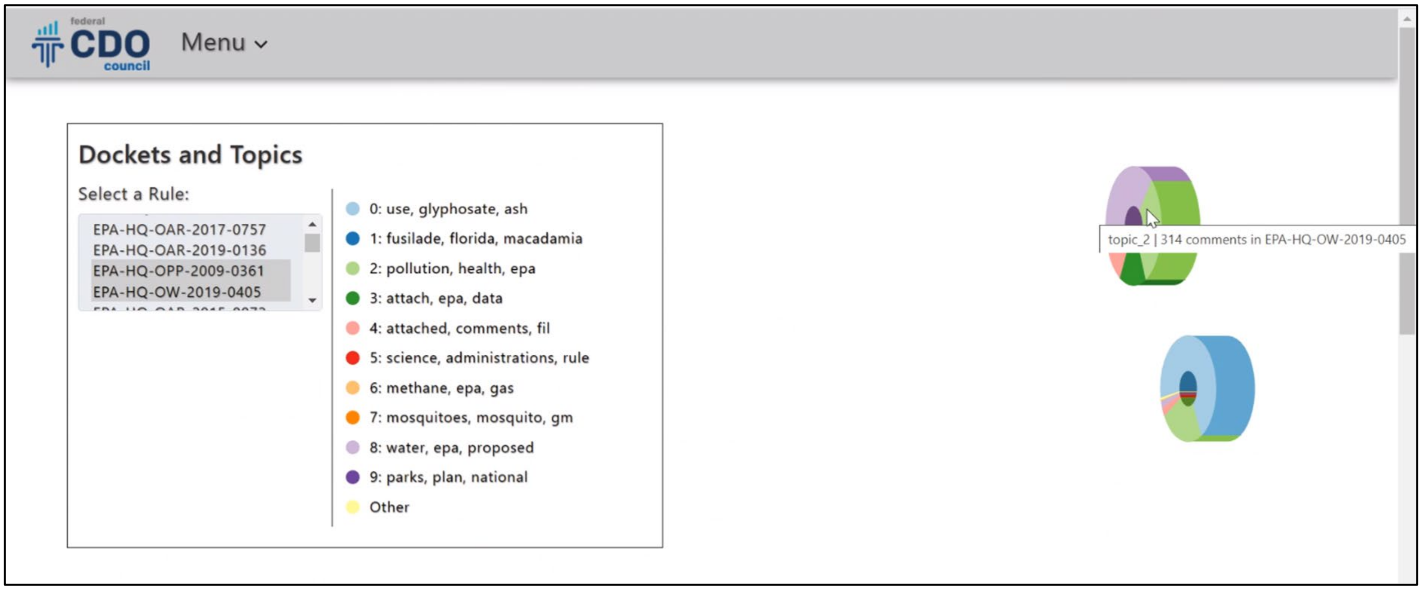
These topics are labeled using the top 3 keywords used in the comments. Note that these results are displayed using LDA, and the hLDA models developed in this pilot would involve more complex branching of topics within comments.

PILOT APPROACH AND OUTCOMES

TOPIC AREAS VISUALIZED FOR EACH REGULATION SELECTED

When you select each regulation, a new donut visualization appears to the right of the screen. This shows the topics, sliced into a 3D pie chart, to visualize the percentage of each topic area that appears in that rule.

Below the user can see that, in this example, there are 314 comments in the highlighted EPA regulation that relate to pollution, health, and the EPA.

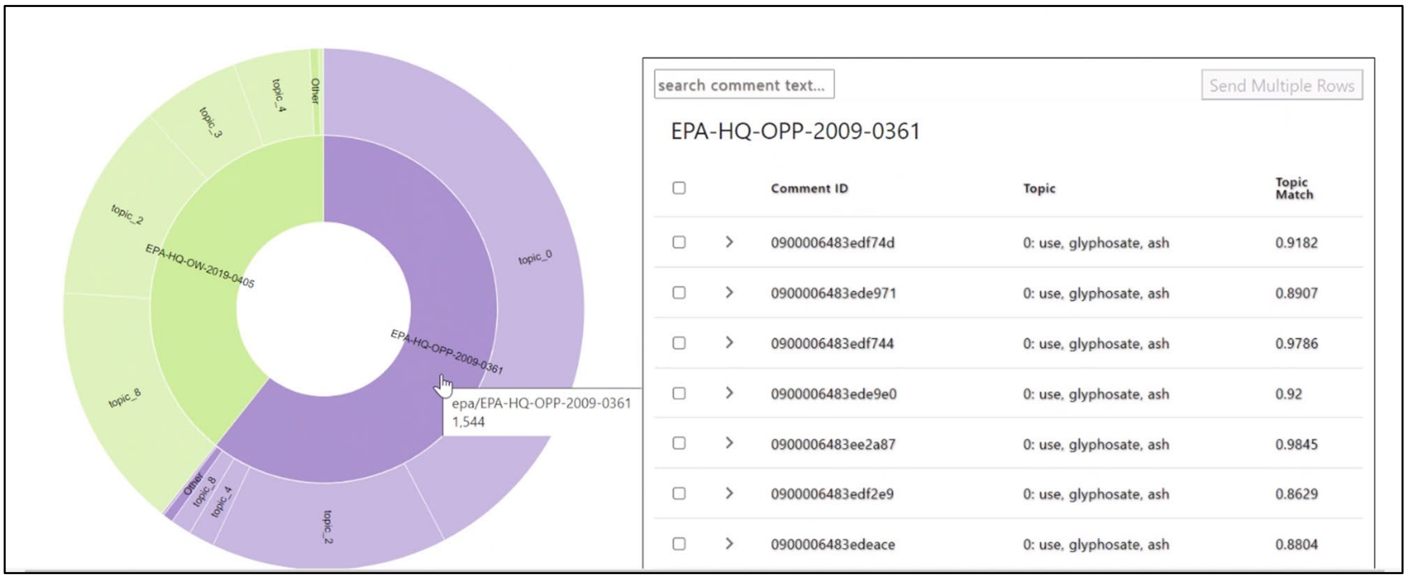


PILOT APPROACH AND OUTCOMES

VIEW INDIVIDUAL COMMENTS

Once the user has selected the regulation(s) they are interested in viewing results for, they can scroll to the bottom of the screen to view the following visualization.

On the left-hand side, a larger donut chart visualizes the 2 EPA regulations selected in the above table. This allows a comment lead to compare topics across multiple regulations if desired.



On the right-hand side, the user can view and drill down into individual comments (identified by their Comment ID, which is an FDMS-generated identifier), which are organized by and aligned to a topic area. A topic match score is also displayed, which indicates how closely that comment aligns to the topic area to which it is aligned. The user can sort based on the topic area or this topic match score.

As an example, the top comment in the table above has a 92% match to topic 0, so it most closely falls into that topic rather than being part of other topic areas. That said, although a comment may fall into one topic area over another, it may still have some overlap with other topic areas and therefore have a topic match score associated with other topic areas as well.

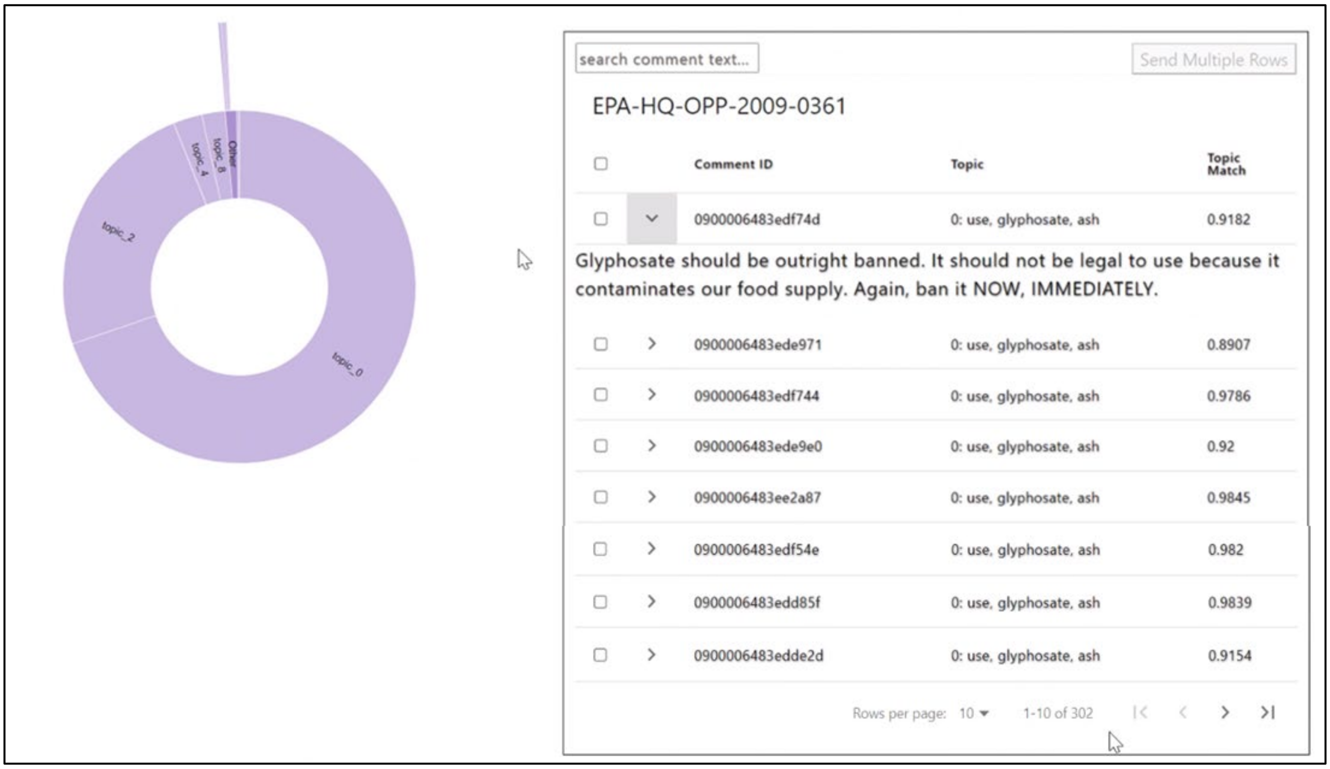
PROTOTYPE UI ENHANCEMENTS

An enhancement here would be to display the regulation titles rather than the Regulations.gov code associated with it, as well as the topic area keywords or labels in addition to the numeric labels of Topic 0, Topic 1, and so forth.

PILOT APPROACH AND OUTCOMES

VIEW TOTAL COMMENT COUNTS

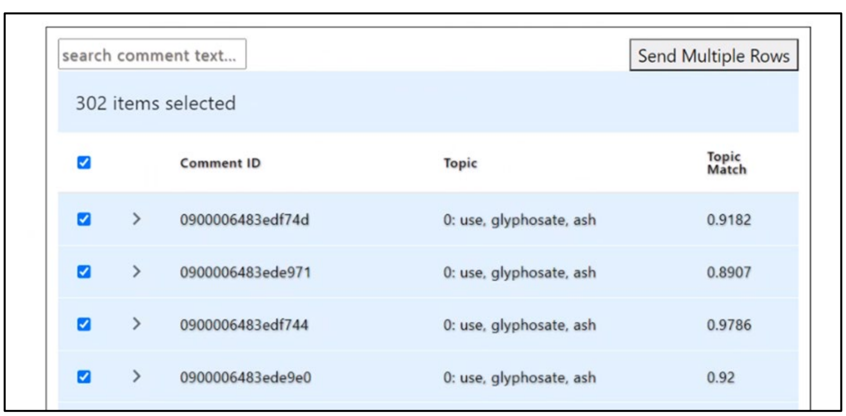
The user can view the comment content by clicking the dropdown carat for its row.



If the user scrolls to the bottom of the screen, they can also view how many total comments are on the selected regulation(s).

SEARCH KEYWORDS

Users can also search the comments for keywords, e.g., “chemical.” Stakeholders indicated the value of this feature. To search, click the box that selects all comments. Enter the term into the “search comment text…” box.



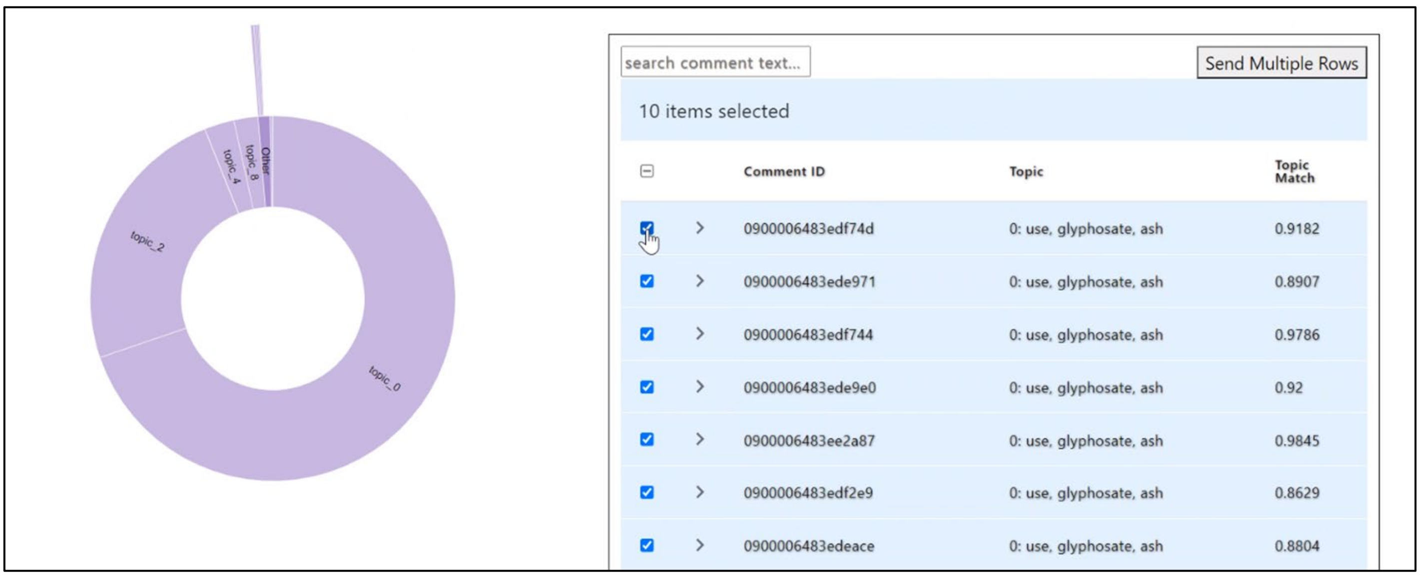
The table will update dynamically to show only the comments that include that keyword. Scroll to the bottom of the table to verify the number of comments that contain that word or phrase.

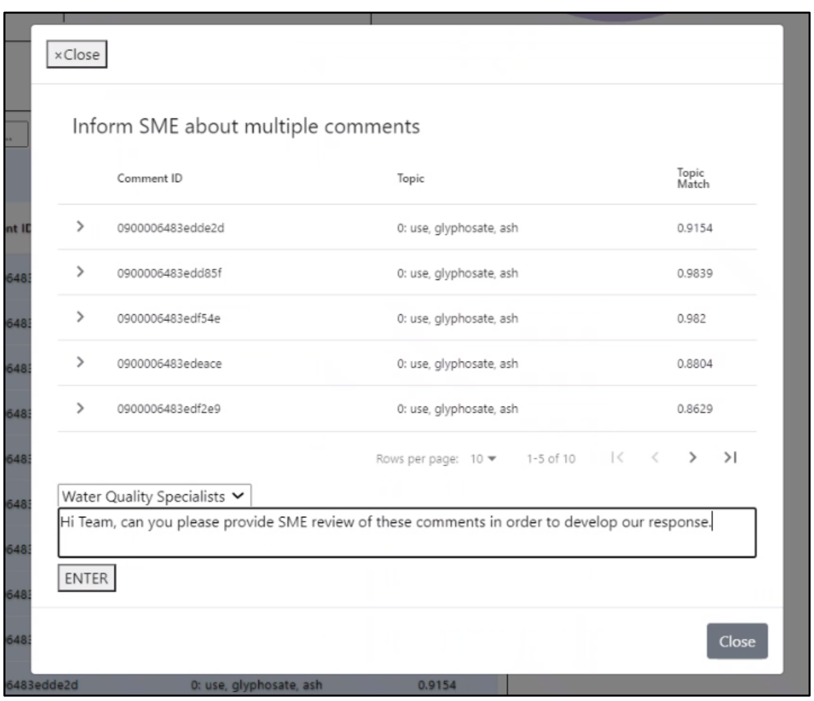
PILOT APPROACH AND OUTCOMES

ROUTE COMMENTS TO SUBJECT MATTER EXPERTS

Based on the topic area of the comments, the user can select individual comments (including multiple rows at a time) and send them to subject matter experts for further review and response. This allows the user to more quickly and easily identify comments to share with SMEs, ultimately reducing the overall time to sort and process comments.

After selecting the comments to share with the SMEs, click the Send Multiple Rows button.





INCLUDE A MESSAGE

A pop-up will appear, allowing the user to select the appropriate individual or group of SMEs to whom the comment(s) should be routed.

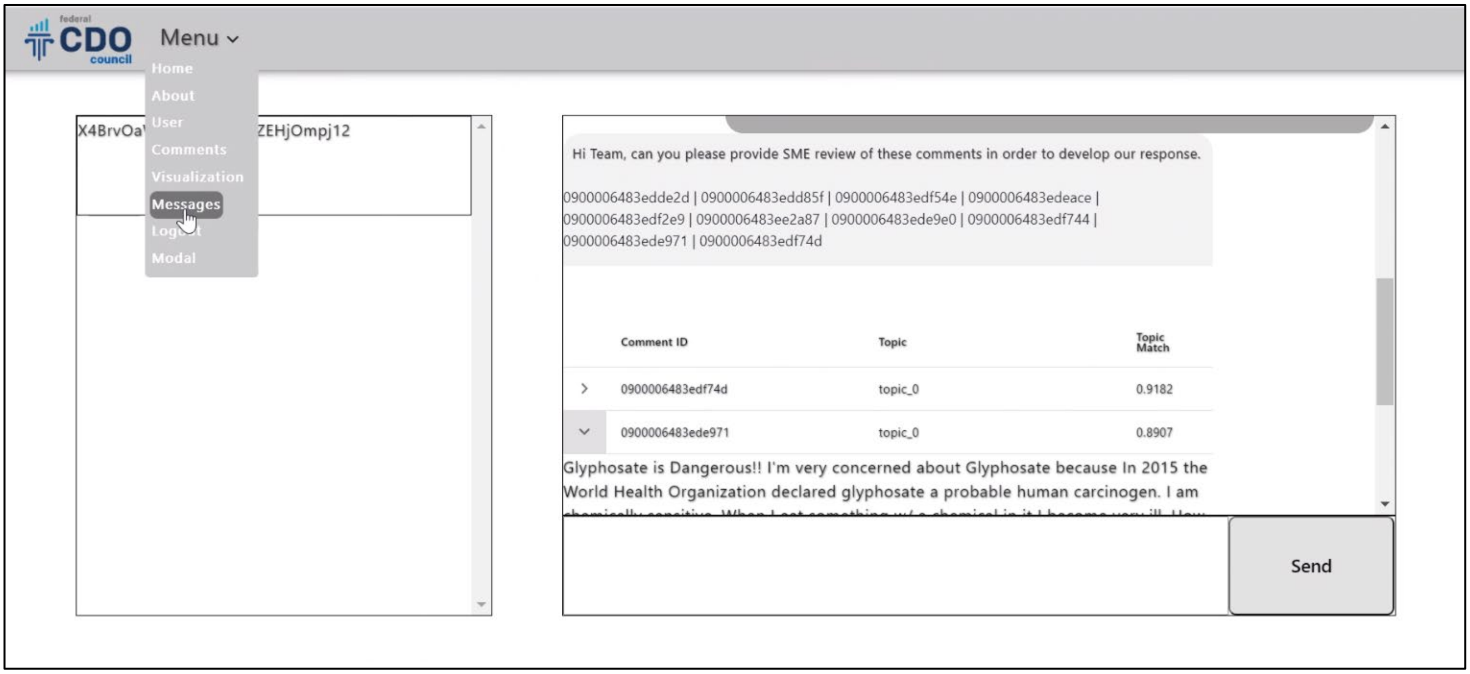
The user can also include a message with additional details on the request.

PILOT APPROACH AND OUTCOMES

SUBJECT MATTER EXPERT VIEW

When the SME logs into the tool, they can access their Messages dashboard to view comments that have been routed to them by comment processors for further review and response.

The SME can drill down into each comment shared with them by selecting the dropdown carat for the comment row. They can send messages back to the comment processor for clarification.



PROTOTYPE UI ENHANCEMENTS

Future enhancements of this feature would include distinct queues, so that once a comment processor routes a set of comments to a SME, those comments no longer appear in the comment processor's queue of comments awaiting review. Instead, they might appear on a dashboard under a status indicating they are with the SME and tracking their status. This would allow the comment team to more efficiently and effectively track, distribute, and respond to groups of comments.

Additionally, we suggest adding the ability to label groups of comments with more detail such as "form letter, mass mail campaign, unique comments, in opposition, scientific," or any other labels as users see fit. This also could lend to applications of interactive topic modeling, allowing users to have additional influence over the topic areas that ultimately result when the model has run.

RECOMMENDATIONS

RECOMMENDATIONS

OVERVIEW

This pilot was intended to prove the hypothesis that NLP tools can be created with high performance and low effort to streamline agencies' comment screening and analysis processes. This team was successful in proving this is possible.

While the tools we have developed have been customized from existing open-source datasets and models in the NLP landscape, they are now in a generalizable format that could be used by any federal government agency to improve their comment analysis process. We have not trained models on specific agency language because using agency-specific data would require a data labeling or annotating process. To supplement this, we identified language models that are pre-trained on a large amount of English text and in some cases trained on certain tasks like semantic text similarity and paraphrase identification. With these pre-trained models, we have found success in generalizable models that work to a degree of performance across agencies, regardless of technical content or comment type.

A generalizable, **base tool**:

- **Reduces upfront, duplicative development costs**
- Provides a standardized solution for agencies that **may not have considered developing their own NLP tool** previously
- **Creates time savings** (see Impact & Cost Savings section)

However, there are tradeoffs. While a generalizable tool may save time for staff up front, it also may **require staff to spend more time analyzing and verifying the results** of the tool. This is especially true when the tool is first implemented.

Additionally, an agency **may choose to spend additional time or resources to customize the tool** to meet an agency's mission or business needs. The agency must be able to evaluate the costs and benefits of each approach (see Impact & Cost Savings section).

Details on the generalizable toolset available at the conclusion of this pilot, as well as a step-by-step "toolkit" for customizing these tools to mission-specific business needs, can be found in the following sections.

RECOMMENDATIONS

CDO COUNCIL PILOT TOOLSET

At the conclusion of the pilot, the team crafted a “Read Me” package, which is a zip-file repository that includes the Python code for the deduplication/semantic similarity and topic models developed in this pilot; the code for the clickable prototype visualization; and instructions for how to use the files. These instructions can be found in the Customization Steps section.

This repository was shared with the CDO Council, GSA, and OMB and will be hosted on a private GSA GitHub. It can be referenced and linked on OMB Max.gov so that it is available to interested agencies. Upon requesting, and being granted, access to the GitHub, agencies could copy this code and follow the scaling steps in following slides to implement them in production.

GSA could identify agencies that are interested in exploring and evaluating the base model against their business needs and fine-tuning the base model to domain-specific knowledge. The intent would be to understand whether the amount of effort by each agency to evaluate the base model, and the time to customize it if required, outweighs the potential efficiency offered through the prediction of topics. For independent agencies interested in performing this validation effort, see the From Pilot to Production and Customization Steps, as well as details on weighing the cost of customizing versus the benefit gained using the base toolset in the Impact & Cost Savings section.

As part of any future development effort, all teams should stress a continued focus on user-centricity by including comment analysts – the end users and voice of the customer – in the human-centered design and iterative development process as the prototype is scaled to production.

RECOMMENDATIONS

PRODUCTION ENVIRONMENT

We suggest NLP tools developed for regulatory comment analysis be incorporated into, or inform modernization efforts of, FDMS. This would allow all Regulations.gov and FDMS users to access new deduplication, semantic similarity, and advanced topic modeling tools.

While many agencies use FDMS and Regulations.gov, some do not. GSA and the CDO Council can make code derived from this pilot available to any interested agency, including but not limited to the agencies that participated in this pilot, so that non-Regulations.gov agencies can leverage the generalizable toolset for their own comment analysis systems and business needs.

Future pilots should follow similarly, where GSA incorporates tools into FDMS but shares the code, so that all agencies government-wide can benefit from advances in NLP tools applied to comment analysis. GSA should also consider the opportunity to offer customization of these pilot toolsets as a shared service to agencies.

RECOMMENDATIONS

FROM PILOT TO PRODUCTION

The semantic similarity and hLDA models were fine-tuned and performed well using a sample of around 500 comments under each rule from various agencies. For the purpose of this pilot and prototype, this reduced the computational complexities described below.

For GSA or agencies that would like to implement these models in production, they can scale the models using the following steps.

	DEDUPLICATION & SEMANTIC SIMILARITY	hLDA TOPIC MODELING
COMPLEXITY	Comparing comments to each other grows quickly: if you have N comments, and you compare each comment to one another, it will correlate to N^2 comparisons. One ruling with even 1,473 comments would end up having 1,084,128 total comparisons. The ruling with the most comments in our pilot dataset has 54,737 comments, which would be 1,498,042,216 comparisons. Reducing the data size helped reduce computational complexity for the purposes of this pilot.	hLDA is extremely memory intensive. Our Floydhub CPU environments offer options of 8GB and 32GB. We encountered memory errors using 8GB. Depending on the number of comments an agency is looking to analyze, memory errors may be encountered at 32GB as well. Depending on an agency's comment count, we recommend using a server 32GB or higher.
1	Copy the de-duplication/semantic similarity Python code from the base CDO Council Pilot Toolkit "Read Me" package.	Copy the hLDA Python code from the base CDO Council Pilot Toolkit "Read Me" package.
2	If SBERT/BigBird does not scale, replace with Python packages FAISS or ANNOY which are designed for high throughput similarity search, but not as accurate.	To scale, select desired platform (e.g., Floydhub, Azure, AWS, etc.) with appropriate memory capacity. For <1,000 comments, use 8GB. For more, use 32GB or 60-64GB.
3	Run the models and assess results. If there are errors, return to step 2. If you would like to further refine the results, proceed to the customization steps.	

RECOMMENDATIONS

CUSTOMIZATION STEPS

In order to **customize the base toolset** provided through this pilot, agencies would need an internal data science team or contractor with knowledge of natural language processing and the selected models, and access to a GPU.

We suggest that agencies first implement the base toolset and evaluate the results and cost/benefit before attempting customization.

If agencies wish to tune the base **DEDUPLICATION & SEMANTIC SIMILARITY MODEL**, an agency's data science team would:

1. Manually identify examples of comment pairs that are semantically similar.
2. Label the data so that they are identified as semantically similar.
   Note: Labels should be created for two categories:
   a) Pairs of comments that are semantically similar (labeled "1," which is the cosine similarity metric indicating most semantically similar)
   b) Comments that are not semantically similar (labeled "0," to indicate they are not at all semantically similar).
      - Note: model performance will be improved if "not similar" comment pairs are at least talking about the same topic.
3. Using the CDO Council Pilot Toolkit developed in this pilot, copy the BigBird de-duplication/semantic similarity Python code from the "Read Me" package.
4. Import this code and existing BigBird model into their GPU or preferred platform.
5. Train the generalizable BigBird model on agency-specific documents with domain-specific terminology, using the labeled comment pairs in step 1.
   - To do this, encode the text with the BigBird model. This produces text embeddings that can be used as a feature vector (predictor variables) for a classification model.
   - The classification model would predict semantic similarity using the language model representation as input data.
6. Run the models and review and validate results.
7. Once there is a language model representation of comments and labels for comments, a classification model (e.g., logistic regression, random forest) can be created using the data scientist's preferred language.
8. Refine as needed.

RECOMMENDATIONS

If agencies wish to further tune the base **HLDA TOPIC MODEL**, they would build on the unsupervised topic modeling technique developed in this pilot to create a weakly supervised topic model. To do this, an agency's data science team would need to:

1. Work with comment processing SMEs to review a set of agency-specific comments and provide desired category labels for the topics found in these comments.
   a) These labels will help validate the results of the model.
   b) Labels should clearly indicate examples of topics.

2. Using the CDO Council Pilot Toolkit developed in this pilot, copy the hLDA Python code from the "Read Me" package.

3. Import this code into their GPU or preferred platform.

4. Train the generalizable hLDA model on agency-specific documents with domain-specific terminology, using the labeled comments in step 1.
   a) To do this, encode the text to produce text embeddings that can be used as input data (predictive variables)
   b) Once there is a language model representation of comments and labels for comments, a classification model (e.g., logistic regression, random forest) can be created using the data scientist's preferred language.
   c) Using the language model representation as input data, the classification model would predict comment topic labels.

5. Run the models and review and validate results.

6. Tune model parameters (e.g., for different topic areas) to influence performance as needed.
   a) Agencies may also consider creating separate models for each topic area, training on a more discrete set to improve algorithms for specific subject areas.
   b) Agencies can use the same code for each language model representation and classification model but different data and labels related to the topic area.

RECOMMENDATIONS

IMPACT & COST SAVINGS

Each agency will need to evaluate the **cost and benefit** of using the base toolset developed in this pilot versus customizing it to meet their needs. This can be thought of in terms of **cost savings.**

We recommend that agencies first use the generalizable toolset and evaluate its results before attempting customization so that they can more accurately estimate the improvement in results.

Agencies should consider the amount of time that they anticipate it would take their teams to complete the customization steps. This can then be compared with the potential and estimated time savings of the generalizable toolset provided .

To aid in this analysis, we solicited responses to an Impact Measurement Survey, in which respondents generally estimated the amount of time it would take them to deduplicate and categorize 100 comments. We also conducted a small-sample time study, in which several EPA comment analysts volunteered to track the work time they actively spent deduplicating, categorizing, summarizing, and responding to comments in real-time.

We averaged the results on the following page to estimate potential time savings.

NOTE: Additional time study responses are being collected, and the time and cost savings may be updated to reflect a more accurate understanding of potential efficiencies. Agencies should perform their own time studies internally as their comment types and comment lengths will vary and influence potential time savings based on agency-specific processes.

KEEP IN MIND

Whether agencies choose the generalizable toolset or customizable route, they will need to **spend time analyzing results**.

However, both options will **significantly expedite comment processing**.

RECOMMENDATIONS

We estimate that the generalizable
deduplication and semantic similarity tool
could save agencies that currently perform this process manually:

approx.
4.5 hours
per 100 comments

We estimate that the generalizable
hLDA topic modeling tool
could save agencies that currently perform this process manually:

approx.
8 hours
per 100 comments

Based on the number of comments an agency expects to receive, they can calculate the time they can expect to save by implementing the generalizable tool.

- For example, an agency that typically receives 1,000 comments may estimate time savings of 45 hours and 80 hours for deduplication and topic modeling, respectively.
- Agencies that receive millions of comments can continue to extrapolate these numbers.

Note that these estimates are averages and may vary based on each agency's comment types, comment lengths, and current processes. Additionally, scale will certainly play a role in this calculation. Agencies that receive more comments will reap more time and cost savings from the use of any tools (generalizable or customized) compared with agencies that process fewer comments. Agencies that receive more comments may want to opt for a customized solution because the time spent reviewing results of the generalizable tool may outweigh the time spent up front to customize the tool. On the other hand, agencies that typically do not receive that many comments may find that the time spent analyzing results of the generalizable tool is far less than the time they would need to spend up front to customize the tool. Other agencies may choose to forego the tool altogether because their comment process is so minimal.

Agencies with more technical domains may need to consider a customizable option. However, note that the pretrained models used in the pilot for semantic similarity and topic modeling produced high-performing results without tuning them specifically on an agency's vocabulary. This indicates a successful generalizable tool.

CONCLUSION

CONCLUSION

This pilot was successful in proving that recent advances in NLP could be used to develop generalizable, high-performing tools to aid any government agency's comment analysis process.

As a proof of concept, and leveraging state-of-the-art, recently released NLP models, the CDO Council pilot team developed tools that help comment reviewers identify and organize the topics and themes found within comments under a regulation, advancing existing tools that perform this function to provide more valuable topic categories when pre-screening comments. We also developed a tool that identifies and groups comments that are duplicates or near duplicates in addition to those that are semantically similar – innovating this feature in the comment analysis space. Finally, the team developed a clickable prototype that demonstrates how these tools can be integrated into a comment analyst's workflow.

Tools like these offer significant value by helping reviewers respond to comments more quickly and easily. They also offer new and better insights in the initial screening and classification of comments, and cost savings can be realized as agencies gain efficiencies and reduce up-front costs of comment analysis tool development. From a government-wide perspective, these tools could reduce duplicative development efforts at individual agencies. Services such as Regulations.gov/FDMS have created many cross-agency efficiencies in the rulemaking space, and this pilot built upon these efforts – with the goal of improving and standardizing comment analysis tools and processes across federal government by leveraging the latest advancements in NLP.

To benefit from the cutting-edge NLP tools developed in this pilot and explore government-wide implementation, we recommend GSA and interested agencies:

1. Leverage CDO Council base toolset developed in this pilot by:
    a) making the code accessible to any interested agencies,
    b) further investing in the model and prototype development, and
    c) using these results to inform FDMS modernization efforts.
2. Scale the CDO Council base toolset, using the steps described in this report, in order to make them production-ready.
3. Evaluate results of the CDO Council base toolset compared to agency mission and business needs and, if desired, customize the models by following the steps described in the Customization Toolkit in this report.

Continued cross-agency collaboration in efforts like these will be integral to realizing efficiencies, innovating, and advancing shared decision support government-wide.

APPENDIX

APPENDIX

DATA EXTRACTION CHALLENGES

The team first attempted to make API calls to Regulations.gov, but initial data extraction efforts presented some challenges:

- Calling the API is limited to 1,000 calls per hour. Any more than that and the user will receive an HTTP status code of 429 (too many requests).
- In addition, every API response is limited in the number of items it returns. Responses are limited to 250 items per page and 20 pages total for a total maximum of 5,000 items per API call.
    - An item is defined by the API in use. In the document API, the returned items are documents. Similarly, in the comment API, each item is information about the comment.
    - With the comment API, the items returned are information about the comment, but not the comment itself.
    - With the comment API, each comment on a document is considered an “item” – meaning, with each API call, the user is limited to 5,000 comments. This means there are some situations where even if you do a very specific API call, if there are more than 5,000 comment responses, there simply will be some comments you cannot get.
    - With agencies that typically see a lot of responses, using the API would mean risking missing comments that they are obligated to review. However, this may not be an issue given an agency only attempting to download comments for their own regulations, rather than across several agencies and regulations, like the pilot team did.
- The comment API returns information about the comments in bulk, but we currently can only get comment text by querying the comment API about each comment's "details" individually.
    - This combined with the 1,000-calls-per-hour limit makes obtaining comment data a time-consuming process.

As a result, the team instead used a bulk export provided by GSA as the primary dataset for this pilot. Note that the team encountered these challenges due to accessing such large numbers of comments for agencies government-wide, and this should not be an issue for individual agencies accessing comments. It is included to advise future cross-agency pilot efforts.

APPENDIX

EXPLORATORY DATA ANALYSIS (EDA) RESULTS

After surmounting the data extraction obstacles, the team performed several rounds of exploratory data analysis (EDA) to understand the dataset. EDA included running token count distribution, sentiment analysis, rapid automated keyword extraction (RAKE), tf-idf, Pandas profiling, attachment analysis in the form of determining the types of files, and counts of each, that appeared attached to comments, as well as testing PDF parsing. Some key findings of EDA are included here.

TOKEN COUNT DISTRIBUTION

What is a token?

"Tokens" are closely synonymous with "words." One token is differentiated from the next token by white space between them.

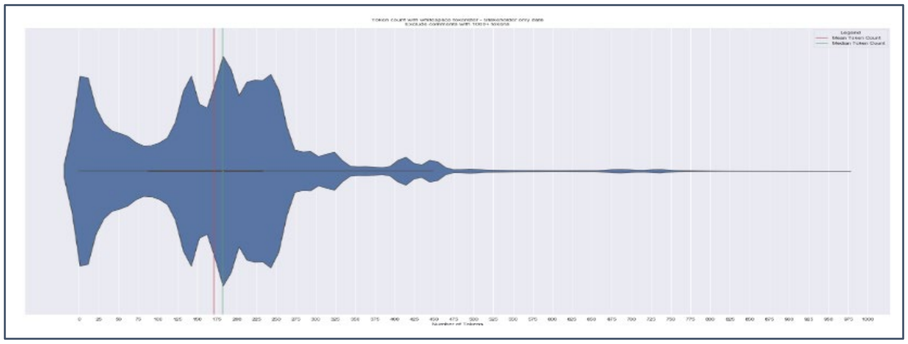
For example:

In the following line of text: "A man, a plan, a canal. Panama!" there are seven tokens, which are:

A	man,	a	plan,
a	canal.	Panama!	

What is token count distribution?

A method for visualizing the frequency of tokens ("words") found within a set of text, such as comments datasets.



What did we do for EDA?

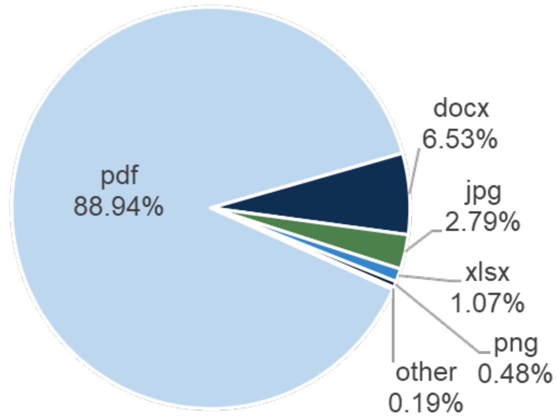
- Using comment data from Regulations.gov, we looked at all data and then specifically at stakeholder agency data to understand the mean and medium lengths of comments in terms of tokens.
- Outliers in the data included comments with over 1,000 tokens; the mean number of tokens was 170.60; and the median number of tokens was 182.
- This information informed the algorithm we selected, since some perform differently given the size of the data (shorter or longer text).

APPENDIX

RAPID AUTOMATED KEYWORD EXTRACTION (RAKE)

- RAKE and TD-IDF methods extract keywords and phrases from a dataset. These were used to provide first impressions of the data for 6 stakeholder agencies' comment datasets.

ATTACHMENT ANALYSIS



- 11% of the comments submitted to stakeholder agencies include attachments. Sometimes, an attachment accompanies a comment that simply says, "see attached." Other times, the attachments include additional evidence or references to complement a full comment. We know from stakeholders that attachments tend to indicate a "substantive" comment.

- Approximately 89% of the attachments sampled (18k of 21k attachments) were PDFs. The pie chart to the right shows the file types found in the stakeholder agency datasets.

- Our team was able to **successfully parse 95.5%** of the stakeholder agencies' attachments, encouraging the possibility of successfully analyzing their content as part of topic modeling efforts.

Why is PDF parsing difficult?

PDFs are historically difficult to parse due to the inconsistent ways that they are encoded. This causes trouble for PDF parsing tools that may be trained to analyze a PDF based on one way of encoding

APPENDIX

HOW WOULD FUTURE TEAMS ANALYZE ATTACHMENTS?

There are challenges associated with analyzing data with attachments.

To perform attachment analysis, there would be 3 automated systems:

1. **System to download attachments**: Minor lift and can be done with the Regulations.gov API. The API's 1,000-calls-per-hour limit would be the biggest obstacle here.
2. **System to determine if attachment is supplemental to the comment or replaces the comment**: This can be difficult to automate as you are making another NLP model to make this distinction.
3. **System to extract text, where applicable**: This system is mainly used for extracting text from PDFs, since in the case where the attachment replaces the comment, the attachment is typically a PDF. This is a minor lift if using open-source software such as Apache Tika to extract PDF text.

Additional details on this potential use case can be found in the Future Efforts – Additional Use Cases section in later slides.

APPENDIX

PANDAS PROFILING

- We ran the Pandas profiling tool on our Regulations.gov comment dataset to explore and better understand the data elements in it (such as Agency ID, Organization, Comment Category, etc.).
- There were many key takeaways (see below and next slide), which allowed us to better understand the data and the business context of that data.

What is Pandas profiling?

Pandas profiling analyzes a dataset (such as the comments on Regulations.gov) and creates a summary of the different data elements found in that dataset.

These data elements may come from the fields a commenter fills out when submitting a comment, or they may be extrapolated from those fields by FDMS or due to agency selections.

Context on the Pandas profiling outputs

- A new version of Regulations.gov was implemented in February 2021 (to standardize public submissions), which helps explain oddities or outliers in the outputs.
- Some results of the Pandas profile appear to extrapolate user-submitted information to generate categories. This could be categories that are manually inputted by reviewing agencies.
- It's helpful to think of "% missing/populated" in terms of the Regulations.gov comment submission form and how it branches.
    - E.g., if you select that you are an organization, you will not be asked to complete first and last name; therefore, the percent completion of first and last name will not be 100%, although they are required if you select individual.

PANDAS PROFILING – INTERESTING TAKEAWAYS

1. Agency ID, Document ID, Docket ID, and posted date are all 100% populated (i.e., every comment is associated with these fields).
    - This makes sense due to FDMS automatically associating the comment with the related agency, document, and docket.

APPENDIX

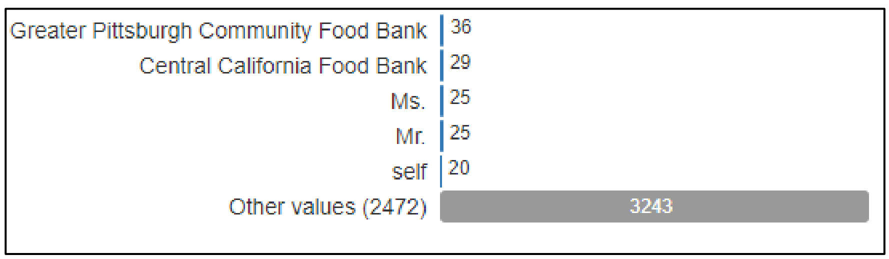
2 "Duplicate comment" is only completed 19.5% of the time, which makes sense since we assume the system is completing this field when identifying duplicates.

- The populated values are inconsistent. Most values are either 0s (no, not a duplicate) and 1s (yes, duplicate).
- 570 comments (out of 283k) have values above 1. This could represent running deduplication on top of already deduplicated comments.

3 First Name/Last Name is 63% populated. This could aid author identification.

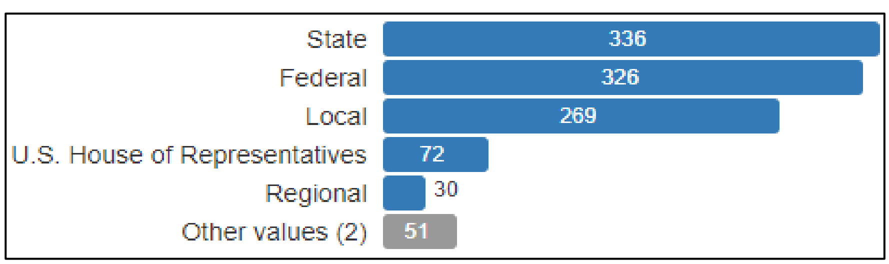
4 City is 32.4% populated. State/province is 31.9% populated. Could be beneficial for identifying issues that affect certain regions (e.g., environmental justice), but is missing too much to be very useful.

5 Organizations is 98.8% missing and includes values "Ms." and "Mr." appearing as organization names. This is a free-form field, completed when the commenter originally selects that they are an organization.



6 Gov agency type is 99.6% missing; this output is derived from a commenter selecting that they are an organization and that their organization type is one of the below options. There are very few of these selections in our sample.

- This comes from the Organization > Organization Type fields selected by users. It is only populated when:
  a) Organization is selected and
  b) anything other than 'company' or 'organization' is selected as Organization Type



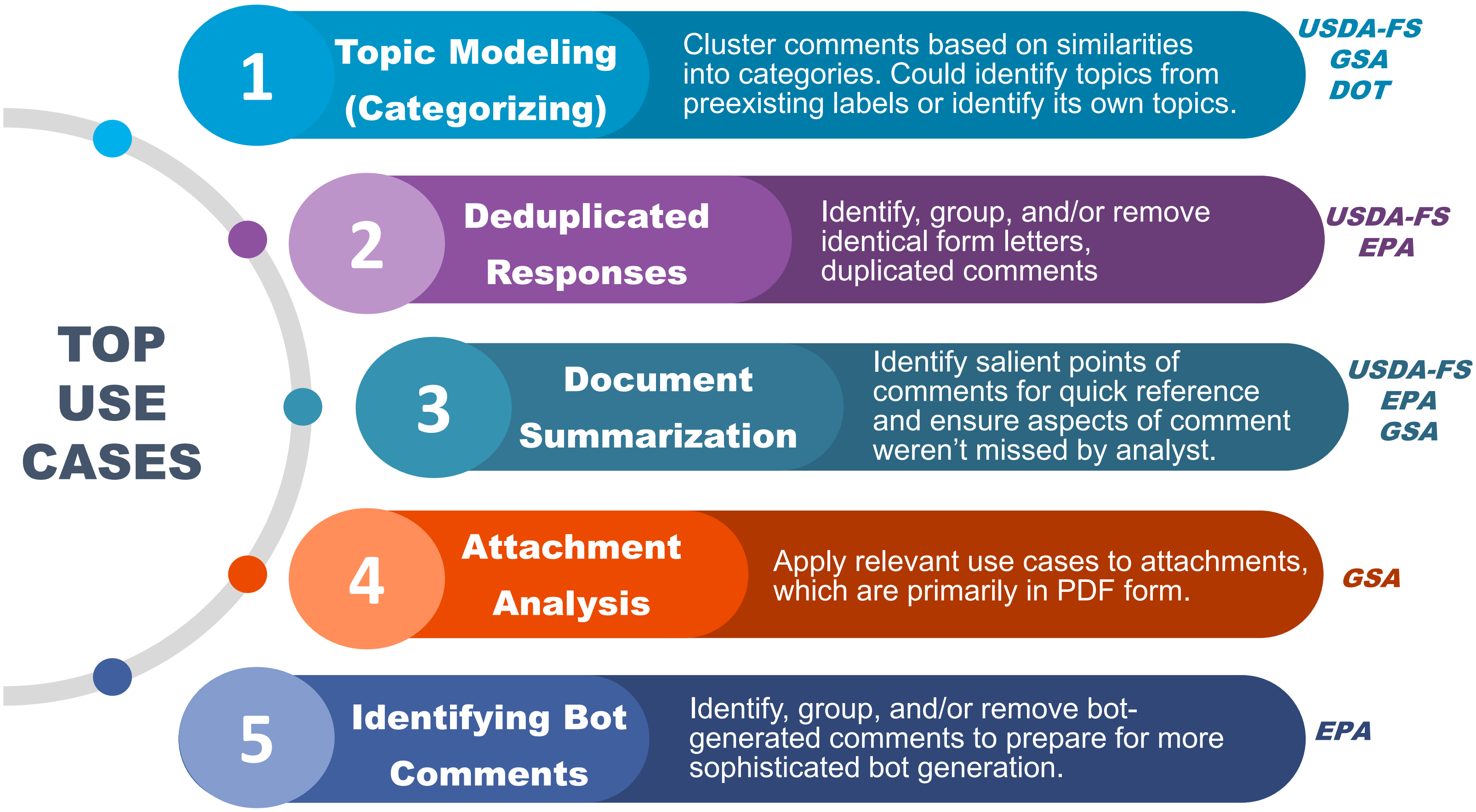
- Could be used if seeking perspectives from other gov. entities.

7 "Is withdrawn" indicates if a comment was withdrawn, which can be done on Regulations.gov but is rare (48 times out of 283k). Reason withdrawn is 99.9+% missing; only populated for withdrawn comments.

APPENDIX

HUMAN-CENTERED DESIGN THINKING OUTPUTS

We hosted a use case prioritization session with our stakeholder agencies. After discussing the proposed pilot use cases that were identified through survey responses discussing comment analysis pain points and potential solutions, we conducted a ranking activity, through which stakeholder agencies prioritized the following use cases.



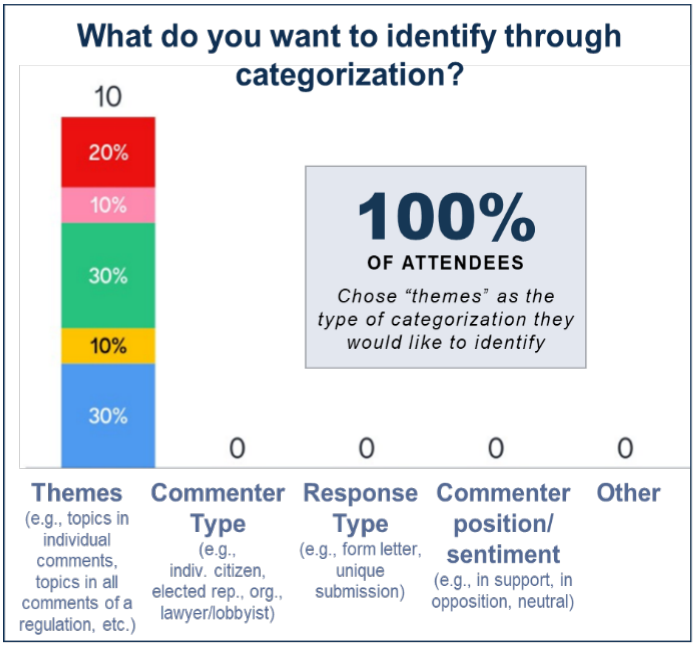
Note: Redaction was ranked #2 by USDA-FS. It was noted that Regulations.gov already has a redaction feature, so this use case may not be pursued. It is worth discussing this ranking with FS.

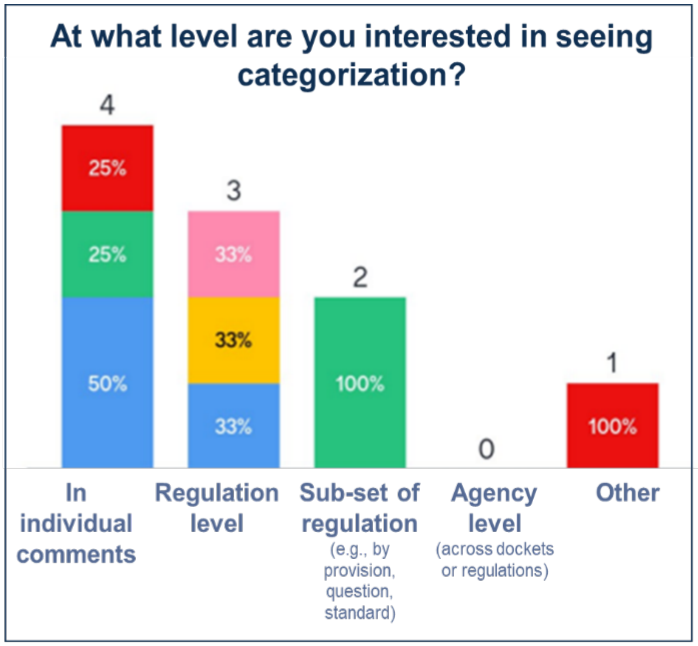
Due to the short nature of this three-month pilot project, the team focused efforts on two use cases – de-duplication and topic modeling – which led to the deduplication and semantic similarity model and the hLDA topic model that were developed in this pilot.

Future pilot efforts should consider the above use cases and build upon the EDA and other development results found in this pilot.

APPENDIX

The team conducted additional requirements gathering sessions with Comment Analysis stakeholders to further understand and refine the categorization use case. As seen in the figures below, 100% of stakeholders wanted to be able to categorize comments by themes or topics, and they were closely split between wanting to see trends in topics within individual comments and at the regulation level.





DEVELOPMENT TASKS: ADDITIONAL DETAILS

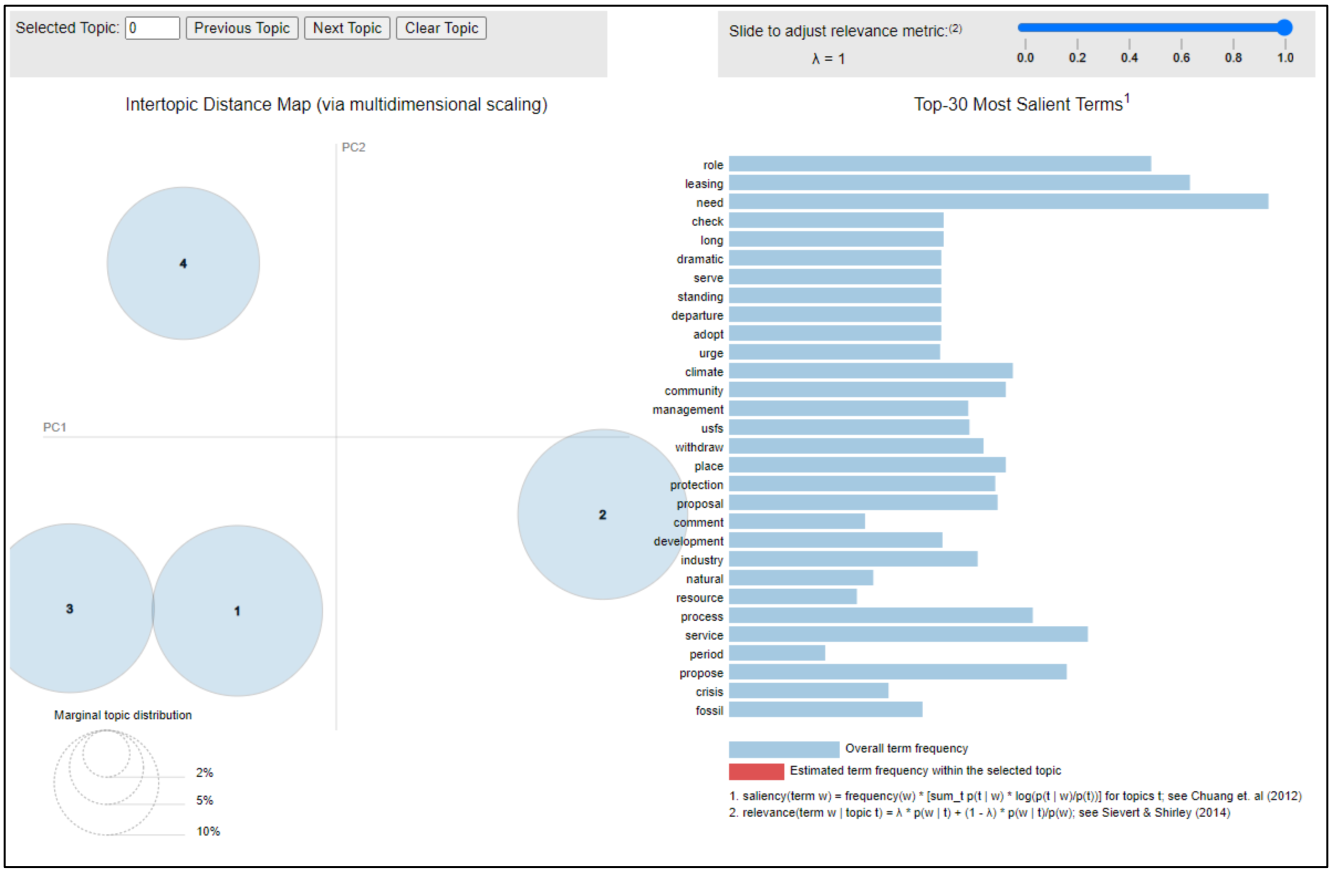
LDA TOPIC MODELING

Originally, the team explored Latent Dirichlet Allocation (LDA) topic modeling. The team briefly explored modeling topics across all regulations for the given time period under each stakeholder agency. Informed by additional requirements gathering sessions, the team narrowed in on the regulation level: organizing comments from one regulation at a time. The team also developed an approach for how topics could be modeled within individual comments, discussed in the Future Efforts section.

Stakeholders raised the concern that some regulations may only be focused on one topic, but others (e.g., omnibus regulations) may cover many very different topics. This led the team to hierarchical topic modeling, which allows for structural relationships between the topics. Stakeholders could identify the number of levels or topics the user is interested in viewing. Future versions of the tool could infer the best number of topics and levels to apply as default for a given regulation.

APPENDIX

When testing the LDA topic model, the team performed LDA on a random sample of comments from USDA Forest Service's Oil and Gas Resources rulemaking docket (Note: For comparison, the hLDA outputs of this rule can be found above).



The model includes a parameter for how many topics the user would like to emerge. The team tried 4, 5, 8, and 10, and found that 4 topics resulted in the most distinct groupings for this regulation.

While these results were interesting to see, stakeholders raised the concern that some regulations may only be focused on one topic, but others (e.g., omnibus regulations) may cover many very different topics. This led the team to hierarchical topic modeling (hLDA), which could solve this concern by allowing stakeholders to identify the number of levels or topics the user is interested in viewing. As discussed in earlier slides, hLDA allows us to identify the hierarchy of themes, producing a tree structure of parent/child relationships, which is ultimately more complex and provides more helpful results for the prototype shared in earlier slides.

APPENDIX

FUTURE EFFORTS – PILOT ENHANCEMENTS

Through the course of this pilot, we identified several enhancements that we recommend exploring.

WEAK SUPERVISION

- While a standardized toolset provides time and cost savings for agencies government-wide, agency staff want the ability to create custom categorization workflows.
- Stakeholders indicated interest in having some interaction with and control over the outputs of the topic models.

✓ **Weak supervision** is a technique that allows subject matter experts (SMEs) the ability to craft a classifier by creating hints.
✓ Weak supervision can be used on a variety of different domains and genres for a variety of different analyses.
✓ This classifier can use the hints supplied by the SMEs to label comment text and generate a classifier that can be applied to unseen comment text.
✓ This could be used in combination with unsupervised results, like the existing topic modeling prototype.
✓ Classifiers can be designed for an agency or even a specific regulation, as needed.

COMMENT-LEVEL TOPIC MODELING

- Stakeholders are interested in categorizing topics in individual comments, since commenters may address many different topics or areas of a regulation in one comment.
- Being able to identify topics in different areas of the comment allows them to quickly sort and distribute parts of the comment to SMEs for response, creating time savings in the overall time it takes to complete analysis.
- From a development perspective, multiple documents are needed in order to model topics found in the dataset.

✓ **The team devised an approach** in which comments are broken into individual paragraphs that were considered their own "document," allowing the paragraphs of an individual comment to be organized by topic.
✓ **Comment-level topic modeling** could be explored using this approach with the hLDA models developed in this pilot.

PROTOTYPE VALIDATION & DEVELOPMENT

- After demoing the final front-end prototype to stakeholders, comment analysts provided feedback and enhancement recommendations.
- We recommend the government invest in further development of the prototype for production-readiness and/or integration with existing tools.
- This includes end-to-end testing and validation of the topic modeling results, given additional time and scope.

✓ Show semantic similarity cosine metric for groupings
✓ Display deltas between nearly duplicate comments
✓ Add ability to route semantic similarity groups to SMEs, as in topic modeling viz
✓ Generate topic labels for semantically similar groupings; topic labels in addition to labeling "Topic 0"
✓ Ability to add additional labels/tags (e.g., tagging a group as form letters, in opposition, substantive, etc.)
✓ Ability to remove or consolidate semantically similar/duplicate comments tagged as form letters

INTERACTIVE TOPIC MODELING

- Interactive topic modeling allows comment processors to nudge and train models in real-time to fine-tune topic labels to fit agency-specific domains or business needs

✓ Leverage the topic models built in this pilot
✓ Allow user to update outputs or nudge model to match their current workflow and thinking

APPENDIX

FUTURE EFFORTS – ADDITIONAL USE CASES

We also identified additional use cases within regulatory comment analysis that could benefit from NLP tools. We recommend that the government explore the following use cases for development and implementation.

USE CASE	IMPACT
Redaction	By identifying comments containing profanity, obscenities, and threats, such comments can be filtered or redacted. This could also apply to PII and other sensitive information. This will block toxic or sensitive information and reduce the amount of manual oversight required, allowing reviewers to focus efforts on substantive comments. Note that FDMS already includes this feature, but future pilot efforts can consider this use case for non-participating agencies.
Attachment Analysis	The team took initial steps to explore this use case, detailed in the EDA section. Additional exploration of attachments could help reviewers more quickly analyze and understand attachments, which traditionally require more time-consuming analysis. Next steps might include creating the systems described above.
Document Summarization	At an individual level, this analysis tool breaks down text into aspects (attributes or components of a policy) and highlights the commenter's opinion or assertion in each. This helps reviewers to see critical points, especially in long responses.
Opinion Identification	Using text analysis techniques, we can identify the specific aspects of a rule or regulation that the public feels strongly toward. This provides reviewers insight into public opinion on an aggregate or individual comment level. This could use sentiment analysis and other techniques.
Identifying Fraudulent Comments	Identifies false information included in comments and/or commenters posing as someone they are not. This ensures comments are authentic and allows reviewers to focus efforts on valuable content.
Identifying Bot Comments / Artificial Comment Creation	Uses metadata to analyze comments and identify computer programs automatically submitting comments. This ensures comments are authentic and allows reviewers to focus efforts on valuable content. This could also incorporate natural language generation detection.
Identifying Commenter	These techniques help identify commenters based on prior known entries or by identifying likely attributes of individual commenters when previous entries are unknown.
Technical Merit Analysis	This use case was specifically requested by stakeholders. It involves assessing the accuracy and technical merit of a commenter's claim or suggestion (e.g., is the scientific evidence referenced accurate?).
Identifying Relevance to Regulation	This use case was specifically requested by stakeholders. It involves assessing the relevance of the comment to the regulation (e.g., is the commenter responding to the regulation, or is the content of their comment random, spam, or otherwise unrelated?).

APPENDIX

FUTURE EFFORTS – ADMINISTRATIVE PRIORITIES

As described in earlier slides, our team took initial steps to explore initiatives aligned with current administration priorities, such as the Executive Orders on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government and Tackling the Climate Crisis at Home and Abroad. We recommend the following phases and next steps to continue this work, building on the topic modeling efforts in this pilot.

EQUITY

- Explore how **supervised topic modeling** and **sentiment analysis** on comments surfaces claims of **inequity, bias,** and **disparate impact** due to a proposed regulation.

Outcomes:

- Improved understanding of regulations, programs, and processes that may cause disparate impact to constituents
- Supervised topic modeling will identify topics of inequity/bias/disparate impact in comments (e.g., a proposed regulation that may lead to disproportionate denial of farm loans to minority communities).
- Sentiment Analysis will show prevalence and intensity of public emotion as it relates to claims of inequity/bias/disparate impact across stakeholder agencies in public comments

✓ **Identify stakeholders with interest in and/or relevance to current administration priorities (HUD, Equal Opportunity Employment)**

✓ **Supervised topic modeling of themes re: claims of inequity/ bias/disparate impact**

✓ **Uses inputs, labels, crowd-sourced annotators, and potentially synthetically generated text to train model**

✓ **Sentiment analysis/emotion classification to identify intensity of racial equity claims in comments**

CLIMATE CHANGE

- Explore how **supervised topic modeling** and **sentiment analysis** on comments surfaces claims of climate change acceleration or other environmental impact due to a proposed regulation.
- Consider the environmental justice intersection with equity project.

Outcomes:

- Improved understanding of regulations, programs, and processes that may cause disparate impact to constituents or the environment
- Supervised topic modeling will identify topics of climate change or environmental impact in comments (e.g., environmental justice issues in impoverished or predominantly minority communities).
- Sentiment Analysis of public response

✓ **Identify stakeholders with interest in and/or relevance to current administration priorities (EPA, Department of the Interior, Department of Energy)**

✓ **Supervised topic modeling of themes re:** claims of climate change or environmental impact

✓ **Uses inputs, labels, crowd-sourced annotators, and potentially synthetically generated text to train model**

✓ **Sentiment analysis/emotion classification to identify intensity of claims in comments**

CDO COUNCIL SPECIAL PROJECTS

Produced by the CDO COUNCIL

Questions?

Please contact Ted Kaouk or Chris Alvares.