Improving Public Comment Review with Automation, Machine Learning, and Advanced Analytics

RTI SmartReview
Overview of Talk

Meet the RTI Team

Federal Rulemaking

How Automation, Natural Language Processing (NLP), and Machine Learning Can Improve Public Comment Review

User-Driven Features and Improvements

Keys to Successful Application of Advanced Data Science Methods for Rulemaking or Other Tasks
Meet the RTI Team

Kim Danforth, ScD
Senior Research Public Health Analyst, Health Care Quality and Outcomes Center

Rob Chew, MS
Senior Research Data Scientist, Center for Data Science

Anna Godwin, MS
Research Data Scientist, Center for Data Science

Ian Thomas
Research Data Scientist, Center for Data Science

Moderator: Cindy D’Annunzio, MBA
Managing Director, Federal Health Services
cdannunzio@rti.org
Federal Rulemaking

• Federal agencies obtain input about potential policy changes via a public comment period on proposed rules.

• Members of the public have the opportunity to submit written comments, which can lead to changes in the final rule.

• Medicare Physician Fee Schedule (PFS) rule is updated annually and uses the typical comment period of 60 days.
PFS Proposed Rule:

Challenges for Comment Review

- **Volume**
  - Large number of comments (recent years, average ~35,000)

- **“Needle in Haystack”**
  - Small proportion of relevant comments (recent years, <200)

- **Overlapping Content**
  - Multiple programs may have similar proposals (e.g., expansion of telehealth during pandemic)

- **Timing**
  - Large percentage of comments submitted during last week of comment submission

- **Short Turnaround to Final Rule**
  - The government needs relevant comments to be identified as quickly as possible to maximize the ability to consider public comments
## Goal for Comment Review System

<table>
<thead>
<tr>
<th>Prioritize</th>
<th>Group</th>
<th>Integrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prioritize likely relevant comments for review and rapid transfer to the government customer.</td>
<td>Group related letters/form letters together for efficient review.</td>
<td>Integrate with an existing Smartsheet® dashboard.</td>
</tr>
</tbody>
</table>
RTI SmartReview: RTI’s Comment Review Process

1. Receive Public Comments
2. Identify Relevant Comments
3. Categorize Relevant Comments
4. Draft Comment Response and Final Rule

Prioritized Review powered by automation and advanced data analytics.

Sequential Review of every comment.
Using Data Science to Improve Public Comment Review
Client-Centered Focus: Data Science Needs to Solve Key Problems

Data Science

- Artificial Intelligence (AI), Machine Learning
- Microsimulation, Agent-Based Modeling
- Automation of Routine, Manual Tasks
- Enhanced Surveillance Systems
- Hotspots and Scenario Planning
- Data Access and Dissemination
- NLP
- Data Visualization
Primary Goal:

Find relevant comments as quickly as possible

Understand client needs and key challenges that data science can solve.

Multidisciplinary teams are critical. Data scientists work closely with analysts throughout rulemaking.
Data pipeline automatically downloads and compiles comments with their attachments each day and performs keyword tagging, relevance modeling (once activated), and clustering.

Automated processes run daily:

- Comment/Attachment Text
- Comment Keyword Matches
- Comment Relevance Scores
- Comment Cluster Assignments

Data Pipeline: Machine Learning/Automated Processes

Comment Texts, Metadata, and Attachments

Convert to Readable Text as Needed

Comments Database

Keyword Tagging

Relevance Modeling

Clustering

Project Team

regulations.gov API

Convert to Readable Text as Needed
## Complementary Approaches Used

<table>
<thead>
<tr>
<th>Features</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated Keyword Tagging</td>
<td>• Can be used immediately; no training data required</td>
<td>• Requires fine tuning to reduce false positives if keywords are general</td>
</tr>
<tr>
<td></td>
<td>• Builds upon team's <em>a priori</em> knowledge of importance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Can be easily revised during rulemaking</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>• Can be used immediately; no data required for training</td>
<td>• Clusters can change as more comments become available</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ Solution: Allows project team to freeze clusters at desired date</td>
</tr>
<tr>
<td>Relevance Modeling</td>
<td>• Applies power of machine learning to identify likely relevant comments</td>
<td>• Requires data for training; not available for use immediately</td>
</tr>
<tr>
<td></td>
<td>• Triggers machine learning model automatically when there are sufficient</td>
<td></td>
</tr>
<tr>
<td></td>
<td>training data</td>
<td></td>
</tr>
</tbody>
</table>
Comment Volume Over Time

Number of Days into the Comment Period

Total Comments Received

2020 2021 2022

Final Days
Public Comment Review (PFS Calendar Year 2021)

- **31,259** Comments Submitted
- **1,106** Flagged Comments (96.5%)
- **180** Relevant Comments
- **168** Relevant Comments Flagged (93%)
- **2 days** After Comment Submission Closing
  → All Relevant Comments Provided to Agency
User-Driven Features and Improvements

Machine Learning
Pipeline Flow Interacts with User-Driven Inputs

Keyword Lists in Smartsheet
Keyword Tagging
Clustering
Relevance Modeling
Relevant Comment List in Smartsheet

Machine Learning Output
Text Preprocessing for Automated Pipeline

• Attachment text (e.g., doc, docx, and PDF files) is parsed using Apache Tika optical character recognition (OCR)
• Comment text is transformed into embeddings, which are produced using TensorFlow Hub
• Embeddings are used in NLP to provide a rich representation of the language used, accounting for synonyms
• Embeddings are used as inputs for the following downstream machine learning processes:
  - Clustering
  - Relevance modeling
Subject matter experts (SMEs) provide categories of interest and corresponding keywords.

**Feline Category**
- Cat
- Tabby
- Kitten

**Canine Category**
- Dog
- Hound
- Puppy

### Example SME Input

<table>
<thead>
<tr>
<th>Comment ID</th>
<th>Feline Category</th>
<th>Canine Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>2</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

### Example Output
Keyword Tagging

- Expanded to show contributing keywords that were flagged for each category

**Example Output**

<table>
<thead>
<tr>
<th>Comment ID</th>
<th>Feline Category</th>
<th>Corresponding Keywords</th>
<th>Canine Category</th>
<th>Corresponding Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>Cat, Tabby</td>
<td>FALSE</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TRUE</td>
<td>Kitten</td>
<td>TRUE</td>
<td>Puppy</td>
</tr>
<tr>
<td>3</td>
<td>FALSE</td>
<td></td>
<td>TRUE</td>
<td>Hound, Dog</td>
</tr>
</tbody>
</table>
Keyword Tagging

• Expanded to show contributing keywords that were flagged for each category
• Expanded to support both inclusion and exclusion keyword criteria

Example SME Input

Feline Category
• Include the following:
  • Cat
  • Tabby
  • Kitten

• Exclude the following:
  • Persian
  • American Shorthair

Canine Category
• Include the following:
  • Dog
  • Hound
  • Puppy

• Exclude the following:
  • German Shepherd
  • Belgian Malinois
Keyword Tagging

• Expanded to show contributing keywords that were flagged for each category
• Expanded to support both inclusion and exclusion keyword criteria
• Expanded for keyword adaptability: inclusion and exclusion keywords to be added, removed, or updated anytime by non-data scientists during the rule period
Keyword Tagging

• Expanded to show contributing keywords that were flagged for each category
• Expanded to support both inclusion and exclusion keyword criteria
• Expanded for keyword adaptability: inclusion and exclusion keywords to be added, removed, or updated anytime by non-data scientists during the rule period
• Expanded to support other types of tagging, such as tagging organization names
Keyword Tagging

- Expanded to show contributing keywords that were flagged for each category
- Expanded to support both inclusion and exclusion keyword criteria
- Expanded for keyword adaptability: inclusion and exclusion keywords to be added, removed, or updated anytime by non-data scientists during the rule period
- Expanded to support other types of tagging, such as tagging organization names

- Adapted keyword tagging to support different input data sources to assist with pipeline and ad-hoc runs

Database Source of Comment Text populated by Regulations.gov API

Local Source of PDF and HTML files for ad-hoc runs
Clustering

• Identifies similar documents and groups them together using embeddings

Example Output

<table>
<thead>
<tr>
<th>Comment ID</th>
<th>Cluster ID</th>
<th>Cluster Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Clustering

- Identifies similar documents and groups them together using embeddings
- Expanded to add Document Opening Text for quick identification of form letters or similar documents

**Example Output**

<table>
<thead>
<tr>
<th>Comment ID</th>
<th>Cluster ID</th>
<th>Cluster Probability</th>
<th>Document Opening Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.80</td>
<td>Feline and canine research is very…</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.50</td>
<td>Research on cats and dogs is critical…</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.95</td>
<td>To whom it may concern, I am writi…</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.95</td>
<td>To whom it may concern, I am writi…</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.75</td>
<td>I am a researcher with an organiza…</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.70</td>
<td>My name is Jane, I am a researcher…</td>
</tr>
</tbody>
</table>
Clustering

- Identifies similar documents and groups them together using embeddings
- Expanded to add Document Opening Text for quick identification of form letters or similar documents
- Expanded to add the capability to freeze clustering algorithm at a given date

Clustering algorithm developed with comments prior to freeze date

Clustering model applied to new comments

Freeze date
# Machine Learning Output

<table>
<thead>
<tr>
<th>Comment ID</th>
<th>Assigned to</th>
<th>Status</th>
<th>Final Outcome</th>
<th>Keyword Group A</th>
<th>Keyword Group B</th>
<th>Keyword Group C</th>
<th>Organization Keywords</th>
<th>Relevance &gt;=0.90</th>
<th>Relevance &gt;0.80 to &lt;0.90</th>
<th>Relevance Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reviewer 1</td>
<td>Complete</td>
<td>Not Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Reviewer 2</td>
<td>Complete</td>
<td>Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Reviewer 3</td>
<td>Complete</td>
<td>Relevant</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Reviewer 4</td>
<td>Complete</td>
<td>Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>0.85</td>
</tr>
<tr>
<td>5</td>
<td>Reviewer 1</td>
<td>Complete</td>
<td>Not Relevant</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Reviewer 1</td>
<td>Complete</td>
<td>Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Reviewer 3</td>
<td>Complete</td>
<td>Not Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>Reviewer 2</td>
<td>Complete</td>
<td>Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>Reviewer 2</td>
<td>Complete</td>
<td>Relevant</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comment ID</th>
<th>Assigned Cluster</th>
<th>Assigned Cluster Probability</th>
<th>Contributing Keywords</th>
<th>Comment Opening Text</th>
<th>Comment URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0.7</td>
<td></td>
<td>Lorem ipsum dolor sit</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>1</td>
<td></td>
<td>do eiusmod tempor in</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>0.73</td>
<td>wordA, wordB</td>
<td>labore nisi ut aliquip</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.75</td>
<td>wordX</td>
<td>aute irure dolor in</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.85</td>
<td>wordY</td>
<td>mollit anim id est lab</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0.9</td>
<td></td>
<td>mollit anim id est lab</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.85</td>
<td></td>
<td>qui officia deserunt</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.91</td>
<td></td>
<td>sint occaecat cupidata</td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.81</td>
<td></td>
<td></td>
<td><a href="https://www.regulations.gov/comment/">https://www.regulations.gov/comment/</a>...</td>
</tr>
</tbody>
</table>
User-Driven Features and Improvements

Operations
Machine Learning - Last Week of Comment Submission

- High volume of comments received in the last several days of public comment submission
  - 10,000+ comments submitted during this time

- Mid-day machine learning runs were needed to allow our prioritized reviews to keep up with public comment submissions

- Project team members would reach out to Center for Data Science to initiate mid-day machine learning runs
  - Required close coordination between team members
  - Potential bottleneck if data scientists were not available
User-Initiated Manual Runs of Advanced Data Science Pipeline
End User Can Easily Customize Settings

<table>
<thead>
<tr>
<th>Input</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>api_timeout</td>
<td>5</td>
</tr>
<tr>
<td>docket_id</td>
<td>CMS-2020-0088</td>
</tr>
<tr>
<td>embed_batch_size</td>
<td>100</td>
</tr>
<tr>
<td>end_date</td>
<td>2020-10-05</td>
</tr>
<tr>
<td>relevance_test_proportion</td>
<td>0.2</td>
</tr>
<tr>
<td>report_ids</td>
<td>['3599646636435332']</td>
</tr>
<tr>
<td>start_date</td>
<td>2020-08-01</td>
</tr>
</tbody>
</table>
Status Dashboard: Review Run State

(masked-muskox)

Last State Message:
24 September 2021 4:23 pm: Running workflow.

Latest State Message:
24 September 2021 4:23 pm: Running workflow.

Tasks:
- **download_document**: 26 Sep 2021 11:55, 10 hours, 0 mins, Success
- **query_documents_with_atta**: 26 Sep 2021 11:55, 26 Sep 2021 11:55, 2 days, 15 hours, 35 mins, Success
- **load_documents**: 26 Sep 2021 11:55, 26 Sep 2021 11:55, 1 second, Success
- **download_documents**: 24 Sep 2021 4:21, 26 Sep 2021 4:21, 2 days, 7 hours, Success
- **start_date**: 24 Sep 2021 4:21, 24 Sep 2021 4:21, 1 second, Success
- **download_relevance_data**: 24 Sep 2021 4:21, 24 Sep 2021 4:21, 5 seconds, Success
- **api_timeout**: 24 Sep 2021 4:21, 24 Sep 2021 4:21, 1 second, Success
- **POSTGRES_UP**: 24 Sep 2021 4:21, 24 Sep 2021 4:21, 1 second, Success
- **relevance_test_proportion**: 24 Sep 2021 4:21, 24 Sep 2021 4:21, 1 second, Success
Quality Control Dashboard: Identifying Issues with the Run
Benefits of Reviewer-Initiated Machine Learning Runs

• Project review team able to initiate machine learning runs at best times for them
• Allows prioritized reviews to stay current and to facilitate rapid identification of relevant comments
• “Push button” process makes it easy for project team to initiate and monitor runs
• Alerts and error logging provides feedback
  - Alerts project team if they need to reach out to Center for Data Science
  - Speeds up technical troubleshooting
Keys to Successful Application of Advanced Data Science Methods for Rulemaking or Other Tasks

• Several data science tools are available; selecting the most useful tool is key to developing efficient and effective solutions.

• Multidisciplinary teams and a firm understanding of client needs are critical.
  - For the PFS rule, the primary challenge is identifying relevant comments quickly from a pool of mostly non-relevant comments: Data science methods focused on identification.
  - For a smaller rule in which all or most comments are relevant, the primary goal may be automated classification of relevant comments into sub-topics.

• Effective collaboration is needed in the development stage of defining the problem and designing and testing solutions, as well as during implementation to respond rapidly to evolving challenges and needs.
Data Modernization and Advanced Data Science Methods

- Using all available data—including unstructured data, real-world data, and synthetic data
- Rapid deployment
- Real-time decision-making
- Data sharing and dissemination
- Automation
Questions?

Thank you!

RTI SmartReview Webinar Contact
cdannunzio@rti.org