Aligning Public Feedback To Requests For Comments On Regulations.gov

Manya Wadhwa, Silvio Amir, Mark Dredze
Department of Computer Science, Johns Hopkins University
Baltimore, Maryland, United States
mwadhwa1@jhu.edu, samir@jhu.edu, mdredze@cs.jhu.edu

Abstract

In an effort to democratize the regulatory process, the United States Federal government created regulations.gov, a portal through which federal agencies can share proposed regulations and solicit feedback from the public. A proposed regulation will contain several requests for feedback on specific topics, and the public can then submit comments in response. While this reduces barriers to soliciting feedback, it still leaves regulators with a challenge: how to produce a summary and incorporate feedback from the sometimes tens of thousands of submitted comments. We propose an information retrieval system by which comments are aligned to specific regulatory requests. We evaluate several measures of semantic similarity for matching comments to information requests. We evaluate our proposed system over a dataset containing several regulations proposed for electronic cigarettes, an issue that energized tens of thousands of comments in response[1].

Introduction

As part of the United States regulatory process, federal regulators solicit feedback from the public on proposed rules before they are finalized. Public participation in the regulatory process aids regulators in crafting rules that are responsive to public needs and concerns (Emery and Emery 2005; Shulman 2005). Therefore, increasing public participation in rulemaking, including through the use of digital technologies, has been a high priority for nearly two decades (Schlosberg, Zavestoski, and Shulman 2008). Launched in 2003, `regulations.gov` provides online access to Federal Register documents and enables federal agencies to solicit and receive public comments. US Federal eRulemaking websites have often drawn widespread attention in public feedback campaigns around major issues, such as in the summer of 2014 when the Federal Communications Commission (FCC) received 3.9 million comments regarding proposed net neutrality rules (Moxley 2016).

While web platforms improve the ability of regulators to solicit and collect feedback, organizing and understanding that feedback can be quite challenging. Consider docket FDA-2014-N-0189 regarding a series of new rules covering tobacco products. The proposed rule solicited feedback on several specific topics, such as definitions of premium cigars, and warning labels. In response to these specific information requests, 119,031 public comments were submitted covering the range of requested topics. How are regulators to sift through such a massive response to identify common responses and important issues that can then influence the regulation? Just as technology provides a means of collecting such information, it can provide a solution to analyzing the resulting comments.

We propose an approach that relies on natural language processing and information retrieval to organize and analyze public comments on federal regulations grounded by specific requests for information from regulatory agencies. We envision a multi-step process in which comments are aligned to information requests and then grouped by common theme. The regulator could then review comment groupings for each information request to develop an understanding of public opinions on these issues.

This paper takes the first step towards this goal by developing a system that can automatically align public comments to information request. We evaluate several text representations for computing similarity between information request and comment, including topic models (Blei, Ng, and Jordan 2003; Chang et al. 2009; Blei, Carin, and Dunson 2010), neural word embeddings based on the word2vec approach (Mikolov et al. 2013), and contextualized representations trained on language modeling objectives (Devlin et al. 2018; Peters et al. 2018; Howard and Ruder 2018).

As our use case, we consider regulations from regulations.gov on the topic of Electronic Nicotine Delivery Systems (ENDS), i.e. electronic cigarettes. Specifically, we consider a series of dockets created over the past several years as the Food and Drug Administration (FDA) sought to obtain feedback on regulations over this emerging market. Due to the popularity of ENDS products, these dockets were among the most commented across all of regulations.gov. Our evaluation demonstrates the promise of using these technologies to aid in the eRulemaking process, and creates new opportunities for research in this area.

[1] https://github.com/mdredze/regulations_gov_icwsm_2020

Feedback Request	Public Comment
Is it appropriate to include the $10 price point in differentiating premium cigars from other cigars? Please provide any data or information that supports the selection of a $10 price point or, if you believe a different price point is more appropriate, that supports the selection of that price point	Using price would place an unfair discriminatory economic burden on those adults who enjoy premium cigars by setting some arbitrary benchmark for what is premium. Consumer taste preferences are unique and different.
Do the words tobacco product in this proposed warning have the potential to cause confusion for consumers? If so, what are the product types where such a warning could potentially confuse consumers?	The words tobacco product could potentially cause confusion in regards to e-cigarettes because consumers do not consider these products to be tobacco products.
Are there other factors FDA should consider to further prevent or discourage people (especially infants and children) from inadvertently consuming or being exposed to liquid nicotine? If so, please explain. Examples of other factors may include: attractiveness of the product or packaging (e.g., appealing images, fragrance, flavors), resemblance of packaging to food and drink items (e.g., candy, fruit), color of the product (e.g., resemblance to beverages such as juice), resemblance of packaging to that of medications (e.g., eye drops).	The FDA should also consider other measures to protect children from the harms of liquid nicotine poisoning, including the use of graphics on warning labels, regulating the use of packaging and flavors attractive to children, and limiting the concentration and quantity of liquid nicotine per unit sold.

Table 1: Examples of feedback request (queries) from Docket FDA-2014-N-0189 along with relevant public comments (documents) returned by our information retrieval system.

Aligning Feedback to Information Requests

`regulations.gov` entries are organized around dockets, a collection of documents relevant to a specific rule making process, including a proposed rule, supplementary material, related notices and the final ruling. Each docket has a unique identifier, title, associated agency and other related information. Within the proposed regulation, the regulatory agency often enumerates specific requests for information. These can take the form of direct questions or open ended feedback requests, as well as questions that relate to other pieces of regulation or that require supporting evidence and data. A proposed rule can include between a few information requests or several dozen.

Any internet user can submit public comments during the open comment period; these are added to the docket for public viewing. Because anyone can participate the source and quality of the content varies significantly, ranging from well supported positions by domain experts (e.g. the businesses affected by the regulation) to lay opinions and off-topic remarks. A comment might include responses to information requests but also general opinions and suggestions about the regulation. Table 1 shows examples of feedback requests in our dataset along with public comments that respond to each request. While most regulations receive little to no feedback, some high profile topics (net neutrality, tobacco regulation) receive tens of thousands of comments. Sifting through and summarizing large volumes of unstructured text poses a significant challenge to the effective integration of feedback into the regulatory process.

We propose a system that organizes public responses by aligning them to specific information requests. Since a comment can address multiple information requests and discuss unrelated issues, the goal is to align each part of a response with an information request within the proposed regulation. For ease of exposition, we will refer to these *comment segments* (sentences in a comment) simply as *comments*. We seek an approach that can identify relevance despite differences in language use between formal regulations and the public. We formalize this as an information retrieval task. Given an information request (query), we return a (ranked) list of comments from the public feedback. Each comment can be judged as either relevant or irrelevant to the information request. When integrated into the regulatory process, a regulator can review each comment for a specific information request, and produce a summary.

We formalize the alignment task as follows. Given a set of information requests (queries) and a collection of public comments from a docket, we return for each query q a list of the top k comments $A = \{c_1, \ldots, c_k\}$ ranked according to their semantic similarity to q. We assume that relevant comments are *semantically similar* to information requests, as they will discuss the same topic. Therefore, we seek measures of semantic similarity to align feedback requests to comments.

We investigate how best to measure the semantic similarity between requests. We compare different text representation methods that map language into embeddings (vectors). We then evaluate each measure on an annotated aligned dataset created from `regulations.gov`. We assess which types of queries are easier to align and thus are more amenable to this kind of assisted analysis.

Feedback Alignment as Semantic Similarity

The vector space model (VSM) provides a general framework to measure similarity between sets of text data such as words, documents, or document collections (Salton, Wong, and Yang 1975; Turney and Pantel 2010). Each text string is represented in a high-dimensional vector space such that the metric similarity between vectors can be interpreted as the semantic similarity between the corresponding contents. A VSM has two main components: a feature function $\phi(d) \mapsto \mathbf{d} \in \mathbb{R}^n$ that maps objects d into n-dimensional vectors; and a metric similarity function to compare vectors (e.g. the euclidean distance or the cosine similarity). In this work we use cosine similarity, ranking comments for each query using cosine similarity between the two vectors.

We explore different choices for ϕ, which then yield different semantic similarity measures.

TF-IDF As a baseline, we consider a bag-of-words model (BOW), where each position in the vector corresponds to a word (type) count. We use a Term Frequency Inverse Document Frequency (TF-IDF) (Salton, Wong, and Yang 1975) weighting in our representation. Our representations

are sparse vocabulary-sized vectors $\mathbf{d} \in \mathbb{R}^v$ with TF-IDF weights.

LDA Topic models provide a topic based representation of text that links documents that discuss similar topics even with little lexical overlap. We create topics using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), which defines a generative model that characterizes documents as a mixture of latent topics, where each topic defines a probability distribution over the vocabulary. Our representations are vectors $\mathbf{d} \in \mathbb{R}^t$ representing the document topic distribution estimated by a trained LDA model, where each dimension represents the likelihood of a latent topic.

word2vec (W2V) NLP has a long history of creating vector representations (embeddings) to represent individual words. In recent years, embedding learning has focused on neural models. For example, the word2vec algorithm learns word embeddings from a large corpus of text with a self-supervised training objective, based on the idea that a word should be predictive of its lexical context (Mikolov et al. 2013). We use word2vec to create representations which are vectors $\mathbf{d} \in \mathbb{R}^e$ constructed by aggregating a pre-trained embedding representation $\mathbf{e}_w$ of each individual word w within the document $\mathbf{d} = \sum_{w \in d} \mathbf{e}_w$. We also consider representations based on word embeddings fine-tuned on our corpus.

BERT While word2vec learns fixed embeddings for each word type, more recent models consider contextualized representations of words and phrases (Devlin et al. 2018). In these models, the representation of a word varies based on its context. For example, while word2vec produces a single representation for every occurrence of the word type "bank", a contextualized representation model will produce a difference embedding for "bank" when it appears in the context "financial bank" versus "river bank." Multiple studies have shown that this leads to superior performance in a wide variety of NLP tasks (Perone, Silveira, and Paula 2018). These language models are typically pre-trained on a large corpus of text, and then applied to the domain of interest. In this work we use BERT to induce text representations. For each document, we first prepend the token $[CLS]$ and then construct a fix sized input matrix $\mathbf{D} = [\mathbf{e}_{[CLS]}, \mathbf{e}_{w_1}, \ldots, \mathbf{e}_{w_{|d|}}]$ by stacking the embeddings of each token, using padding to ensure that all the inputs have the same length. The document matrix is processed in a forward pass trough network to compute $\mathbf{Z} = [\mathbf{z}^d_{w_1}, \ldots, \mathbf{z}^d_{w_{|d|}}]$, where $\mathbf{z}^d_w$ is a representation of token w in the context of the document d, produced by the last hidden layer of the network. We consider two methods for constructing vector representations for an entire text document. **BERT-pool**, aggregates all the contextualized token representations $\mathbf{d} = \sum_i \mathbf{Z}_i \in \mathbb{R}^h$. **BERT-cls**, takes the contextualized embedding $\mathbf{z}^d_{[CLS]}$ and passes it through an additional dense layer to produce the final document representation $\mathbf{d} = g(\mathbf{z}^d_{[CLS]}) \in \mathbb{R}^h$ — this corresponds to the sequence encoding approach used to train BERT for the next sentence prediction task (Devlin et al. 2018).

Evaluation

We evaluated our approach over a set of regulations proposed in `regulations.gov` dockets concerning e-cigarette regulation. We selected this topic since it is one of the more popular topics and received tens of thousands of public comments. We used the official `regulations.gov` public API[2] to search for dockets with documents containing the keywords `cigarettes` or `tobacco`. We filtered the results to only include dockets where the contents contained at least one keyword related to electronic cigarettes (i.e., `electronic cigarette`, `e-cigarette`, `vape`, or `ENDS`), which resulted in 678 dockets. We extracted all the public comments associated to each docket and segmented the comments into individual sentences using the NLTK package[3]. Since the data is intended for use by the US goverment, we assumed all documents were in English.

Some comments include supporting documents or even the actual responses as attachments; we ignored these and kept those that had text in the response itself. We observed that some comments included direct quotes from the information request to either contextualize the response or address a specific issue. This overlap in content will naturally make these comments more similar to the queries, which can bias the system towards trivial alignments. To avoid this issue, we replaced the quoted text with a special token "[QUOTE]". Finally, we discarded duplicated and short comments (less than 5 words). We selected for annotation the five dockets with the most comments: 269,671 public comments, though 97% pertain to FDA-2014-N-0189.

We created a dataset of queries by manually identifying and extracting the requests for comments and feedback within each proposed regulation by searching for the keywords `feedback` and `comments`. Since a proposal can have multiple feedback requests, we tried to make queries as self-contained as possible while preserving enough context to be intelligible in isolation. We grouped multiple requests into a single request whenever it was necessary to preserve coherence. Table 1 shows examples of queries and aligned comments (see Appendix for more examples) . We found a total 77 queries across the five dockets ($avg = 15.4$; $std = 8.8$), with significant variation in terms of number of tokens ($avg = 67.6$; $std = 54.6$). To support our analysis, we categorized queries into the following types: 9 **closed questions** (i.e. that can be answered with yes or no), 8 **open questions** (i.e. that ask for general comments or opinions), 14 **compound questions** (i.e. that involve multiple questions), 8 **questions with external references** (i.e., that mention external regulations/documents/articles) and 18 **requests for evidence and data** . Table 2 summarizes the dataset derived from `regulations.gov` (see Appendix for examples).

Experimental Setup

Our document representation methods require a large corpus of raw text to derive word co-occurrence statistics or fit the

[2] https://regulationsgov.github.io/developers/

[3] http://www.nltk.org/

	Feedback Requests	Comments
FDA-2014-N-0189	14	262,408
FDA-2013-N-0521	13	1,769
FDA-2015-N-1514	30	4,081
FDA-2012-N-1148	6	748
FDA-2011-N-0467	3	665

Table 2: Summary of the `regulations.gov` dataset.

parameters of predictive models. We constructed such a corpus by sampling 100K documents consisting of both public comments and proposed regulations. We preprocessed the text by lower-casing, removing punctuation, infrequent terms (less than 5 occurrences) and stop-words, and then extracted a vocabulary of $v = 24,380$ unique tokens. We used this corpus to compute inverse document frequency statistics, fit the topic model and fine-tune the weights of pre-trained word embeddings. We trained an LDA topic model with $t = 100$ topics using the `lda`[4] python package with the default hyper-parameters. For the word embedding representations, we used the English pre-trained word2vec embeddings of dimensionality $e = 300$ trained on the Google News corpus[5]. We then fine-tuned these embeddings to our data for 5 epochs using the Skip-Gram implementation available in the `Gensim` (Řehůřek and Sojka 2010) package with the default hyper-parameter settings. For the BERT model we used a publicly available `PyTorch` implementation[6], which is a direct port of the original BERT-base release from Google (including the pre-trained parameters). The model consists of a deep feed-forward neural network with $L = 12$ layers of Transformer encoders (Vaswani et al. 2017) with 12 self-attention heads and hidden layer size $h = 768$, trained on the entire English Wikipedia. Since training BERT is computationally expensive, we only used the model with distributed pre-trained parameters without further fine-tuning.

Results

We compared the document representation strategies with respect to relevance ranking performance using two metrics: (1) mean reciprocal rank (MRR), which measures the quality of the overall rankings; and (2) Precision at $K = 5$ (Prec@K), which measures the proportion of top K ranked documents that are relevant to the query. We adopted a pool based evaluation methodology — we pooled the top K ranked documents by each method into a single evaluation set, and annotated these rankings. A human annotator (not an author) was asked to judge each alignment (i.e. query, comment pair) as *relevant*, *irrelevant*, or *need context* (when the information was not enough to make a confident decision). Since we did not train any methods for this task, this dataset was only used for evaluation purposes.

The main results are shown in the first two rows of Table 3. The best overall results are achieved with a word2vec representation with fine-tuned embeddings. We were surprised

[4]https://pypi.org/project/lda/

[5]https://code.google.com/archive/p/word2vec/

[6]https://github.com/huggingface/transformers

	TF-IDF	LDA	W2V	W2V (tuned)	BERT-cls	BERT-pool
Prec@K	0.42	0.33	0.48	**0.48**	0.16	0.28
MRR	0.52	0.51	0.56	**0.60**	0.31	0.47
avg. $\|q\|$	67.2	63.3	64.3	62.5	76.5	83.2
avg. $\|c\|$	21.1	26.0	38.1	39.0	36.1	43.6

Table 3: Results of the alignments produced by each method, in terms of precision@k and mean reciprocal rank (top 2 rows), and the average length of correctly aligned queries and comments (2 bottom rows)

that both BERT approaches perform rather poorly. BERT-pool does slightly better than BERT-cls, which was expected since the latter corresponds to a method tailored for next sentence prediction and needs to be fine-tuned for a new task. Our hypothesis is that the low performance is due to a mismatch between the pre-trained model and our task/data. BERT was designed for and trained on longer sequences than most of our comments, and on a different domain. To test whether length matters for BERT on this dataset, we collected the relevant alignments produced by each method and measured the length of the respective queries and comments. The bottom two rows of Table 3 show that correct alignments produced by BERT involve much longer comments and queries. When shorter documents are involved, BERT assigns similarities of more than 0.99 to a large number of alignments, suggesting that with such short sequences the document representations become very similar (most of the sequence is padding). In these cases, the relative orderings become much noisier and thus another set of top comments could just as easily have been picked. In contrast, TF-IDF tends to do much better on short comments that involve very explicit clues or highly discriminative phrases (e.g technical terms). Since the methods are complementary ensembles of methods could improve the overall performance of the system.

To understand which types of queries are easier to align with comments, we averaged the MRR of all the methods for each query type. We found that *open questions* and *requests for evidence and data* were the easiest to align with $MRR = 0.56$, followed by questions with *external references* with $MRR = 0.51$. The most difficult are *compound* ($MRR = 0.44$) and *closed* ($MRR = 0.35$) questions, which is not surprising since the former requires the system to keep track of multiple questions and it is unlikely that a single comment (sentence) addresses all of them; whereas the latter, requires the system find closed responses (i.e. yes/no) that address the question precisely, which would require a high degree of natural language understanding.

Conclusions

We present an information retrieval system to align requests for feedback in regulation proposals with public comments on the `regulations.gov` portal using varying methods of semantic similarity. We compared various semantic representations with respect to the quality of the alignments and found that fine-tuned word2vec representations performed the best. The output of our system is a (ranked) set of comments responding to each query. While this al-

lows a regulator to review comments per information request, it still presents a very large number of comments to read. Therefore, future efforts should be directed at methods to group and summarize the contents of aligned comments. This works presents only the first step towards improving public participation in the regulatory process with automated content analysis. To foster further research in this area, we have publicly released the annotated dataset collected for our experiments[1].

Acknowledgements

This research was funded by the Burroughs Wellcome Funds Innovation in Regulatory Science Award (1017617.01).

Appendix

Some examples of different types of queries that were manually identified and extracted from a proposed regulation posted on `regulations.gov`:

Open *Given the rapid growth of social media (e.g., Facebook, Twitter,YouTube, etc.), how can minors exposure to tobacco product advertising, promotion, and marketing through these types of media be restricted or minimized?*

Closed *Are there other tobacco product standards, regulatory, or other actions that FDA could implement that would more effectively reduce the harms caused by menthol cigarette smoking and better protect the public health than the tobacco product standards or regulatory actions discussed in the preceding questions?*

Compounded *Since the enactment of the PACT Act, have minors found alternative methods to purchase and/or acquire cigarettes or smokeless tobacco products by a means other than a face-to-face exchange? If so, what are they?*

External reference *What technologies, procedures or other methods are currently used by the tobacco industry (including, but not limited to, manufacturers, importers, distributors, and retailers) to restrict or minimize a minor's exposure to the forms of advertising, promotion, and marketing of tobacco products described in questions 11 and 12 of section II.B of this document?*

Requests for evidence and data *Are you aware of any existing evidence regarding whether warnings (text and any applicable color or graphic element) are effective for mitigating the risks of nicotine exposure? If so, please provide that evidence.*

References

Blei, D.; Carin, L.; and Dunson, D. 2010. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE signal processing magazine* 27(6):55.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, 288–296.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emery, F., and Emery, A. 2005. A modest proposal: Improve e-rulemaking by improving comments. *Administrative and Regulatory Law News* 31(1):8–9.

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moxley, L. 2016. E-rulemaking and democracy. *Admin. L. Rev.* 68:661.

Perone, C. S.; Silveira, R.; and Paula, T. S. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Salton, G.; Wong, A.; and Yang, C.-S. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18(11):613–620.

Schlosberg, D.; Zavestoski, S.; and Shulman, S. W. 2008. Democracy and e-rulemaking: Web-based technologies, participation, and the potential for deliberation. *Journal of Information Technology & Politics* 4(1):37–55.

Shulman, S. W. 2005. E-rulemaking: Issues in current research and practice. *International Journal of Public Administration* 28(7-8):621–641.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37:141–188.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.