### Text summarization of congressional legislation

Connor Cabrey1 and Dalia Habiby1

1*School of Public Affairs, American University, Washington, DC, 20016*

#### Abstract

This project aims to address the issue of accessibility concerning Congressional legislation. Using data from the 117th Congress Second Session, we conducted a text summarization task through an abstractive approach, using a pre-trained spaCy pipeline. and methods to effectively summarize bills. These tasks will enable us to create an interactive conversational system, through which users may inquire about certain policy areas and they will receive summaries of a few related bills. This will ensure that users can fully understand bills of interest in efficient time.

**1. Introduction**

Many bills are over thousands of pages long and contain verbose language that many people do not have the time or bandwidth to decipher, which can lead to slow decision-making and fundamental misunderstandings of the content [1]. Furthermore, the summaries offered on Congress.gov are slow to appear and tend to be vague or very brief.

Creating accurate, automatic summarizations of these bills would not only be beneficial for the general public, but also for government officials who work in busy environments and need to make well-informed decisions about whether to vote in favor of or against a bill. Additionally, citizens will be more aware of what their representatives are voting for, as these summaries will be objective descriptions of the bills, rather than explanations filtered through the media that may have a partisan spin.

**2. Literature review**

There is a substantial collection of literature on text summarization for many professional areas, especially in the legal field. Manual drafting of case summaries is a long and grueling process that is slow and expensive. The process relies on a group of “specialized staffs whose task is to summarize cases” [3]. The number of legal documents produced every year seems to be comparable to the number of legislative documents our Congress produces as well, and methods used in this research may lead to similar results.

A study published in the Expert Systems With Applications international journal presents an unsupervised method for extractive multi-document summarization using a centroid approach and sentence embedding representation [6]. The authors utilize a sentence importance score that is a composite of three other scores to create summaries of the Multi-News dataset. After evaluating with cosine similarity and Recall-Oriented Understudy for Gisting Evaluation (ROUGE), they concluded that their approach is comparable to the current cannon of text summarization methods, including supervised deep learning [6].

Automatic summarization of Congressional bills is a relatively new field, and thus there is much room for improvement. In 2019, two researchers from FiscalNote Research created the first corpus for this task, titled “BillSum” [4]. In this study, they applied a “Document Context Model” (DOC) using TF-IDF, a “Summary Language Model” (SUM) using a BERT model, and an ensemble of both models to predict a label they created indicating whether or not it belongs in the summary.

We intend to build upon the aforementioned methods by implementing a

**3. DATA**

The data we are using for this project comes from the govinfo.gov web API [2]. We used this API to download only House Resolution (HR) Bills submitted to the floor for the 117th Congress Second Session, as the total number of HRs is slightly overwhelming for an experimental analysis such as this one.

We used 4674 individual HRs for our data set. Since the files were originally in XML format, the first step was extracting the text elements using the Element Tree module of the xml package in python. This included parsing each file and extracting the root element, which identifies the heading of the file for identification purposes. Then, we located all elements that included text relevant to the bill and extracted them into a data frame. Each observation in the data frame consists of an identifying row number, the raw bill text, and an initialized column to hold the generated summaries.

**3.1. Data pre-processing pipeline**

Next, we tokenized each document into sentences and applied lowercasing, stop word removal, and punctuation removal. We decided against using stemming or lemmatization to preserve the grammatical accuracy of the summaries. Due to memory limitations, we also capped bill length at 100,000 characters.

**4. METHODS**

For text summarization tasks, there are two main approaches that can be used. The first approach, generally the simpler and less advanced approach, is “Extractive.” This method uses simple and traditional algorithms that observe the frequency of words within the text and looks for sentences that contain those high frequency words. These sentences are then used in the final summary. The more complex and advanced method is an “Abstractive” approach. This method attempts to make smaller sentences with the same semantic meaning as the original sentences. This approach generally uses deep learning transformers like the BERT method, and this is the approach we implemented.

Using a pre-trained spaCy model called “en\_core\_web\_sm” which includes vocabulary sematic processes

This method applies a vectorization to calculate an importance score of each token, or sentence, in a bill. We then use these scores to extract the most important sentences and combine them to create a summary.

Capped the summaries at 1,000 words

In order to evaluate the success of this task, we calculated readability scores

Cosine similarity between the bill and the summary

LexRank- eigenvectory centrality

From sumy package, using LexRankSummarizer: graph based algorithm for text summarization that identifies most important sentences based on page rank algorithm from Google. Has its own rules about what a sentence is and vectorizes sentences from those rules.

7. Results

6. Discussion

**6.1. Conclusions**

There is statistically significant evidence that the proportion of exam failure among the Black or African American participants is greater than the proportion of failure among the White participants. Among the Black population, keeping all other variables constant, Pell Grant recipients have 0.94 times the odds of passing that non-Pell Grant recipients have, those with a GPA above 3.0 have 1.066 times the odds of the GPA below 3.0 group of passing, conditional accepts have 0.77 times the odds of regular accepts of passing, and the non-binary group has 1.24 times the odds of females of passing.

Furthermore, keeping all other variables constant, of the Black participants in DC, Pell Grant recipients have 1.059 the odds of never passing that non-Pell Grant recipients have, conditional accepts have 1.181 times the odds of never passing that regular accepts have, and males have 1.115 times the odds of never passing that females have. Of the Black participants in DFW, those with a GPA above 3.0 have 0.873 times the odds of never passing that those with a GPA below 3.0 have, and conditional accepts have 1.203 times the odds of never passing that regular accepts have. Of the Black participants in Baltimore, conditional accepts have 1.206 times the odds of never passing that regular accepts have.

As of now, there are mixed results about whether Mometrix use is statistically significant, and whether it has a positive or negative relationship with the odds of passing a content exam.

**6.2. Future research**

Given the results of this project, we are now aware of the vulnerable populations when it comes to passing licensure exams and ensuring that they can become educators. To build upon this research, future exploration and analysis should focus on the implementation of new policies to help these groups, and whether they are effective or not. In terms of the Mometrix analysis, investigation into a floor effect may indicate that a minimum amount of time using Mometrix products is required for improved likelihood of passing. A randomized experiment using Mometrix products would illuminate more about its effectiveness than the minimal data currently available.

For such a study to be statistically sound, there must be randomized treatment and control groups with a pre-test and a post-test. Additionally, as many other variables as possible must be held constant, such as time spent studying, study tools used, and courses taken. The ethical implications should be considered carefully, since some participants would be receiving more resources than others. These research endeavors will illuminate more about the populations that need more intervention to make it to the classroom while license exams are still required.

7. Closing REMARKS

**7.1. Limitations**

The limitations of this analysis include the extremely large available corpus, our processing power, and the subjective nature of text summarization. There are thousands of Congressional bills that have been written, and thus limiting our data to the second session of the 117th Congress excludes the vast majority of available data. Furthermore, due to limited processing power, we had to cap the length of a bill, which constrained our data. Additionally, text summarization is a nuanced task, and what one person thinks is a good summary may not be what another person thinks, and evaluation methods cannot capture every aspect of a summary.

**8. References**

1. Bennani, H., & Neuenkirch, M. (2022). Too complex to digest? federal tax bills and their processing in us financial markets.

2. Discover u.s. government information. (n.d.). <https://www.govinfo.gov/>

3. Kanapala, A., Pal, S., & Pamula, R. (2017). Text summarization from legal documents: A survey. *Artificial Intelligence Review.*

4. Kornilova, A., & Eidelman, V. (2019). Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523.*

5. Key to versions of printed legislation. U.S. Senate: Key to Versions of Printed Legislation. (n.d.). https://www.senate.gov/legislative/KeytoVersionsofPrintedLegislation.htm

6. Salima Lamsiyah, Abdelkader El Mahdaouy, Bernard Espinasse, Saïd El Alaoui Ouatik, An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings, *Expert Systems with Applications*, Volume 167, 2021, 114152, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2020.114152>