”Classification of Vehicle Security and Privacy Incidents in Federal and State-Level Policies Using Natural Language Processing”

Ajay Kaarthic Jeysree, Ashray Kanakasabapathy Bhaskar, Gautham Shaji, Harshavardhini Sridhar, Justine George

***Abstract*—In the legislative arena, the precise classification of congressional bills assumes a pivotal role in informing policymakers and advancing democratic processes. This endeavor focuses on binary classification, discerning bills related to highway, transportation, or automobiles from unrelated ones. Employing advanced NLP techniques and vast textual data, the study automates bill identification, expediting analysis and fostering data-driven policymaking. Leveraging machine learning, including deep learning models, predictive algorithms undergo rigorous training, validation, and optimization for effective legislative text classification. Performance metrics, such as precision, recall, F1-score, and accuracy, assess model prowess, enhancing legislative decision support.**

I. Introduction

In this research endeavor, we embark on the formidable task of framing a binary classification problem with the primary goal of developing advanced machine learning models. These models are meticulously designed to effectively differentiate congressional bills into two distinct categories: those intricately linked to the realm of highway infrastructure, transportation policies, or the automotive sector (denoted as label 1), and those that pertain to other legislative domains (denoted as label 0). Our approach leverages state-of-the-art natural language processing (NLP) techniques and harnesses a wealth of textual data resources, ultimately aimed at automating and streamlining the intricate process of bill identification. Beyond its role in expediting legislative analysis, this automation empowers data-driven policymaking—a cornerstone of modern governance.

Our research methodology encompasses a diverse array of machine learning models, including cutting-edge deep learning architectures. These models undergo rigorous training, exhaustive validation, and meticulous optimization processes, ensuring their robustness in handling the inherent intricacies of legislative text. To comprehensively evaluate model performance, we apply a suite of comprehensive performance metrics encompassing precision, recall, F1-score, and accuracy, offering a holistic assessment of their classification capabilities and reliability.

This research unfolds in two pivotal phases. Initially, we tackle the classification task without data downsampling, recognizing the significant class imbalance characterized by approximately 255,725 instances labeled as 0 and a mere 667 as 1. In this context, we employ Universal Sentence Encoder embeddings to represent bill titles and descriptions. Logistic regression and ClassifierDL are employed for classification. Subsequently, we address the class imbalance by downsampling label 0 to three times the size of label 1, and we reapply Universal Sentence Encoder and ClassifierDL for classification.

Furthermore, our research delves into the realm of unsupervised learning, focusing on clustering. We cluster the embeddings derived from annotated data into two distinct clusters. Extending this clustering analysis to the pre-tagged data, we assign each bill to one of these two clusters. This dual-pronged approach provides valuable insights into the effectiveness of clustering as a means of grouping bills, complementing the original supervised classification.

II. Data Preprocessing

In the preprocessing phase of our research project, we encountered a substantial class imbalance in the dataset. This imbalance refers to a significant disparity in the number of instances belonging to each class or category. Specifically, in our dataset, we observed that one class (label 0) had a disproportionately larger number of instances compared to the other class (label 1). Such class imbalances can pose challenges in training machine learning models as they tend to favor the majority class and may perform poorly on the minority class.

To address this data imbalance and ensure the uniform distribution of data across train and test sets, we implemented a downsampling strategy and a train-test split. Here’s a summary of the process:

1. Class Imbalance Assessment: - Initially, we identified the class imbalance by examining the dataset’s distribution of labels. This step revealed a substantial difference in the number of instances between label 0 and label 1.

2. Downsampling Strategy: - We decided to employ a downsampling strategy to mitigate the class imbalance. The goal was to reduce the number of instances in the majority class (label 0) while retaining a representative subset.

3. Determining the Downsampled Size: - To determine the size of the downsampled majority class, we aimed for a balanced ratio between label 0 and label 1 instances. In our approach, we set the size of the downsampled majority class to be a specific multiple of the minority class size (e.g., three times the size of label 1).

4. Random Selection: - With the downsampled size determined, we randomly selected instances from the majority class to create the downsampled dataset. Random selection helps ensure that the retained instances are representative of the majority class.

5. Train-Test Split: - After downsampling, we performed a train-test split on the dataset. This split was done in a way that maintained the uniform proportion of each label in both the training and testing datasets. In other words, we ensured that both the training and testing datasets had a balanced representation of label 0 and label 1 instances.

6. New Balanced Dataset: - The result of downsampling and the train-test split was a new balanced dataset for training and testing. This balanced dataset was then used for subsequent machine learning tasks, including model training and evaluation.

By implementing downsampling and a uniform train-test split, we addressed the challenges posed by data imbalance and ensured that our machine learning models were trained and tested on datasets with a fair distribution of instances from both classes. This preprocessing step contributed to the overall success and reliability of our classification models in accurately categorizing congressional bills.

III. Modeling and consistency validations

In our research project focused on classifying congressional bills based on their content, we employed a series of Natural Language Processing (NLP) modeling approaches to tackle the classification task. We sought to determine whether a given bill is related to the domain of highway, transportation, or automobiles (label 1) or falls into other legislative categories (label 0). This section provides an in-depth overview of the NLP modeling techniques we applied in this endeavor.

A. Model 1: Logistic Regression with Universal Sentence Encoder Embeddings

In our first modeling approach, we employed logistic regression to categorize congressional bills based on their titles. This process began with document assembly, utilizing the DocumentAssembler to preprocess and transform bill titles into a format suitable for analysis. Next, we integrated pre-trained Universal Sentence Encoder (USE) embeddings, enabling us to represent the processed titles as semantic-rich vector representations. To further enhance these vectors, we utilized the EmbeddingsFinisher, producing finished embeddings that were well-suited for machine learning tasks. Finally, we employed a logistic regression model trained on these finished USE embeddings to execute binary classification, effectively distinguishing between labels 0 and 1. This model served as the foundation for our initial bill classification efforts.

B. Model 2: Deep Learning with Universal Sentence Encoder Embeddings

In the second model, we delved into the realm of deep learning techniques, leveraging Universal Sentence Encoder (USE) embeddings to enhance our classification capabilities. The process began with document assembly, mirroring the approach in our first model, where we meticulously preprocessed and readied the bill titles for analysis. Next, we harnessed pre-trained USE embeddings, transforming the bill titles into intricate vector representations, enabling us to capture and encode their semantic nuances effectively. The centerpiece of this model was the adoption of the ClassifierDLApproach, a robust deep learning-based classifier. This advanced approach empowered us to tap into the formidable potential of neural networks for bill classification. To orchestrate the entire process, we constructed a streamlined pipeline. This pipeline efficiently assembled the document, applied the USE embeddings to derive meaningful representations, and seamlessly integrated the ClassifierDL approach for training.

The fusion of deep learning techniques with the versatility of USE embeddings marked a pivotal step in our journey to develop more accurate and sophisticated models for the classification of congressional bills.

C. Model 3: Deep Learning with Balanced Data

In the third model, we addressed the initial class imbalance by introducing data preprocessing techniques. Our primary objective was to create a more balanced dataset by ensuring an equitable distribution between labels 0 and 1. To achieve this, we implemented a downsampling strategy that reduced the instances of label 0 to three times the size of label 1, thus mitigating the data imbalance. We maintained consistency in our approach by utilizing the DocumentAssembler and Universal Sentence Encoder (USE) embeddings for document preparation and vector representation generation, as in the previous models. This ensured that our text data was appropriately processed and encoded for deep learning analysis.

For classification, we retained the use of the ClassifierDLApproach, a deep learning-based classifier similar to the second model. The combination of document assembly, USE embeddings, and the ClassifierDL approach formed the core components of our model training pipeline. Notably, this model benefited from the balanced dataset, allowing us to explore the impact of data balancing on the classification results.

In all three models, we assessed their performance using standard evaluation metrics, including precision, recall, F1-score, and accuracy. We also compared the results of each model to evaluate their classification capabilities accurately.

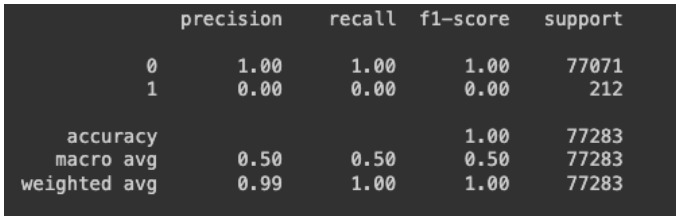
D. Model 4: Unsupervised K-Means Clustering

With the standardized. data set and USE embeddings in place, we applied the K-Means clustering algorithm. K-Means is a popular unsupervised machine learning technique that groups data points into clusters based on their similarity. In our case, the data points were the USE embeddings of congressional bills, and K-Means aimed to group bills with similar content and semantic meaning.

After running the K-Means clustering algorithm, we obtained clusters of bills, each represented by a cluster centroid. We explored the properties of these clusters, such as the distribution of bills within each cluster and the characteristics of the bills grouped together. By analyzing the bills in each cluster, we gained valuable insights into the inherent structure of legislative documents.

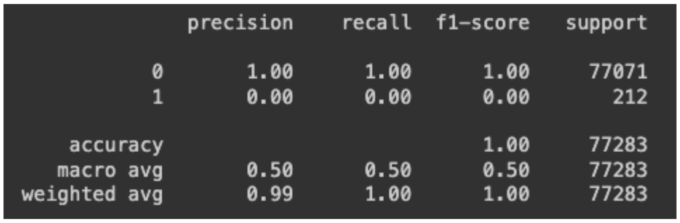
IV. Conclusion

The NLP modeling approaches allowed us to explore different strategies for classifying congressional bills, considering both data imbalance and deep learning techniques. Our research aimed to provide insights into the effectiveness of these models in automating the categorization of bills, which can have significant implications for legislative analysis and data-driven policy making.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     77071
           1       0.00      0.00      0.00       212

    accuracy                           1.00     77283
   macro avg       0.50      0.50      0.50     77283
weighted avg       0.99      1.00      1.00     77283
```

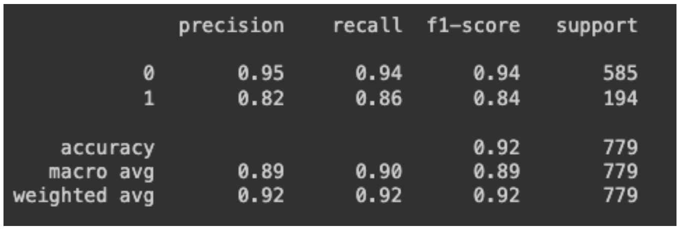
Fig. 1. Logistic Regression with Universal Sentence Encoder Embeddings



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     77071
           1       0.00      0.00      0.00       212

    accuracy                           1.00     77283
   macro avg       0.50      0.50      0.50     77283
weighted avg       0.99      1.00      1.00     77283
```

Fig. 2. Deep Learning with Universal Sentence Encoder Embeddings

In the scenario without data undersampling, it was observed that the model consistently predicted all data points as belonging to the majority class, which is labeled as 0. The situation significantly improved following the implementation of data undersampling.

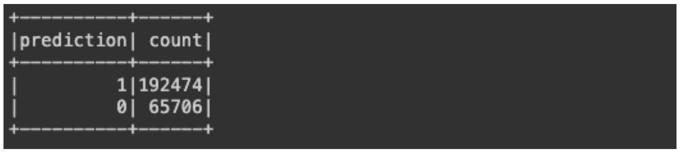


```
              precision    recall  f1-score   support

           0       0.95      0.94      0.94       585
           1       0.82      0.86      0.84       194

    accuracy                           0.92       779
   macro avg       0.89      0.90      0.89       779
weighted avg       0.92      0.92      0.92       779
```

Fig. 3. Deep Learning with Balanced Data

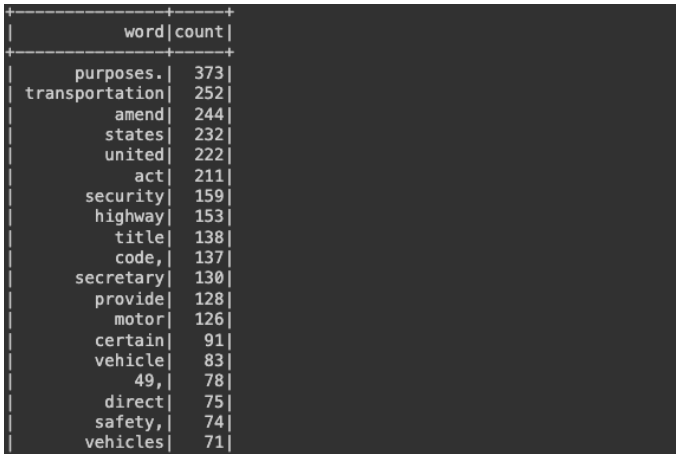
The use of unsupervised K-Means clustering in our project allowed us to explore the natural groupings of congressional bills based on their semantic content. This approach provided valuable insights into the organization of legislative documents and enriched our understanding of the data beyond traditional classification methods.

In conclusion, our research project exemplifies the synergy between machine learning and legislative analysis. We demonstrated that the efficient and accurate categorization of congressional bills is attainable through a combination of advanced NLP techniques, deep learning models, and data preprocessing strategies. Moreover, the incorporation of unsupervised learning enriched our understanding of the data, revealing clusters and patterns that may have otherwise remained hidden. This multidimensional approach not only contributes to the field of legislative analysis but also underscores the potential of machine learning in informing data-driven policymaking and enhancing democratic processes. As legislative bodies continue to grapple with vast volumes of bills and policy proposals, our research offers a valuable framework for automating bill categorization and empowering more informed decision-making.



```
+----------+------+
|prediction| count|
+----------+------+
|         1|192474|
|         0| 65706|
+----------+------+
```

Fig. 4. Clustered Data Prediction Counts (0 and 1 represents clusters, not class)



```
+--------------+-----+
|          word|count|
+--------------+-----+
|     purposes.|  373|
|transportation|  252|
|         amend|  244|
|        states|  232|
|        united|  222|
|           act|  211|
|      security|  159|
|       highway|  153|
|         title|  138|
|         code,|  137|
|     secretary|  130|
|       provide|  128|
|         motor|  126|
|       certain|   91|
|       vehicle|   83|
|           49,|   78|
|        direct|   75|
|       safety,|   74|
|      vehicles|   71|
```

Fig. 5. Most frequent words from data,