1 Introduction

In the 2018-2019 sessions nearly 200,000 bills were introduced in state legislatures in the United States.

2 Related Work

- Legislative Data
- Text Classification with Metadata

2.1 Quantitative Political Science

Quantitative Political Scientists have been analyzing legisltive data and it's behavior for years. To date most studies in the US have focused on Federal legislation given the breadth of high quality datasets available. The methods roughly track the advances in natural laguage processing over the years.

Gerrish and Blei (2011)[1] construct a topic model given bill text and individual legislator's votes and use supervised methods to explore how topics relate to voting paterns.

Nay (2017) [2] Takes Federal legislation and predicts enactment using word vectors, gradient boosting, and ensemble stacking. A dozen features are constructed out of bill sponsors, committee data, and subject. Various combinations of text, metadata, and text plus metadata are analysed

2.2 Text Classification with metadata

Zhang etal. (2021)[3]

3 Data and Models

3.1 Data

We were provided data from a private company GovHawk that collects state and federal legislative data in realtime and sells subscriptions to parties interested in the legislative process. The data encompasses the 2018-2019 legislative session for upper and lower house in every state (nebraska has one). Substantial amount of data on individual votes and party identification of congresspeople was provided but was not complete for every state. The data was processed into the following features for use in modeling.

Bill Text - The raw text provided by each legislature is in legal markup and contains a large amount of extra content not related to specifics of the bill. Headers, footers, and boilerplate language were removed. References to other legislation, section/subsection language, and all numbers and extraneous symbols were removed. The text was made into lower case. The remaining text is a good faith effort at the natural language of the bill. See appendix for an example (TODO).

Revision Number - Bills go through a natural revision process and the current number of revisions is available at prediction time and thus has relevant information. Inclusion of all revisions of a bill would contribute to overfitting of a model. Therefore, for each bill we select one revision at random and include it in our dataset. This reduces the number of observations from over 600,000 to around 200,000.

Partisan Lean - The liberal/conservate membership of each legislature was calculated and coded as a continuous variable from 0.0 to 1.0 where 0.0 is an all conservative legislature, 0.5 is an even split, and 1.0 is an all liberal legislature. The intuition for this feature is that an NLP model should upweight the passage of ”liberal” legislation in a liberal legislature and downweight it in a conservative one. To the extent possible this was derived from available metadata, where not the values were filled in from wikipedia.

Session-Chamber - A categorical variable encoding the legislature, session pair from the dataset. Some legislature’s sessions comprise multiple years and bills carry over from one to the next. Texas meets every two years. In all there are 133 categories that are one-hot encoded.

Passage - A two class (0/1) variable indicating whether the legislation passed the specific legislature.

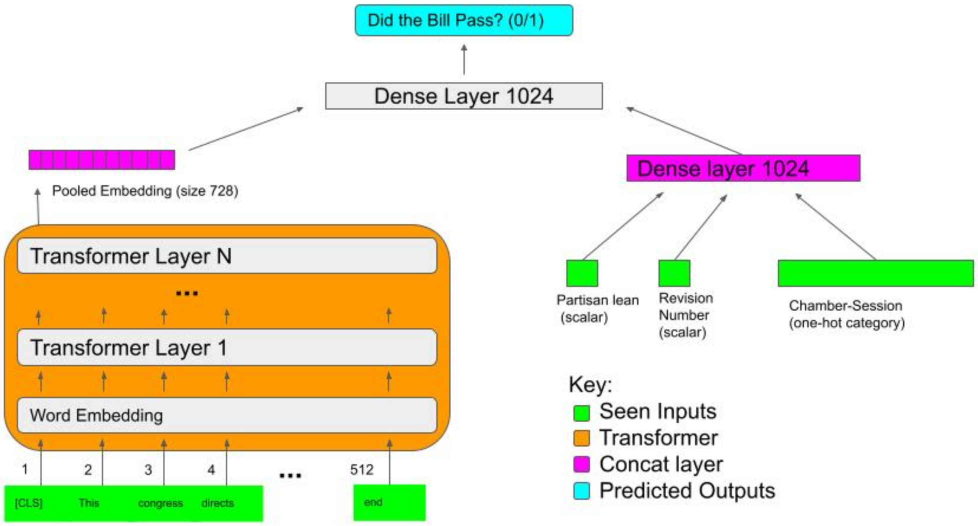
Signed - A two class (0/1) variable indicating whether the legislation passed by both legislatures and the executive.

3.2 Baseline Models

Gradient Boosting Decision Trees (GBDTs) are powerful models for both classification and regression tasks. Often, for tasks where the representation does not need to be learned (neural networks), they are best in class. Catboost is a member of the GBDT family introduced in late 2018 that produces two innovations that make it best in class for data with categorical variables. It has two innovations, Ordered Target Statistics and Ordered Boosting

3.3 Bag of Words Models

3.4 Transformers



- Transformers
- Proposed architecture
- DistilBert Bert – Encoder only models

4 Results

	AUC	Precision	Recall
Metadata Only			
Logistic Regression	0.869	0.73	0.37
Catboost	0.885	0.73	0.45
Dense Neural Network	0.886	0.75	0.44
BoW Text Models			
Logistic Regression with LDA	0.880	0.73	0.42
CatBoost with LDA	0.914	0.77	0.56
Transformer with Metadata combinations			
Only Metadata	0.886	0.75	0.44
Text Only			
Text, Revision Number, Legislature			
Text and Partisan Lean			
Text and All Metadata			
Transformer with Catboost			
DistilBert 128	**0.921**	**0.79**	**0.58**
DistilBert 512	0.916	0.77	0.57
Longformer 4096			
Different Transformers (all metadata)			
DistilBert (128 tokens) (in question)			

[1] Dense Neural Network is precisely the same inputs and dense layer as used inside the transformer-based models and is included as to measure the gain from transformers.

4.1 Partisanship in Legislation via Ablation Analysis

- Partisan Lean and quantifying partisanship in a bill.
- Revision number and predicting left-out revisions.

5 Summary

6 Bibliography Todo

- **papers to digest**
- Categorical Metadata Representation for Customized Text Classification

- Large Scale Legal Text Classification Using Transformer Models

References

[1] Sean M Gerrish and David M Blei, *Predicting legislative roll calls from text*, Proceedings of the 28th International Conference on Machine Learning, ICML 2011, 2011.

[2] John J Nay, *Predicting and understanding law-making with word vectors and an ensemble model*, PloS one **12** (2017), no. 5, e0176999.

[3] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han, *Match: Metadata-aware text classification in a large hierarchy*, arXiv preprint arXiv:2102.07349 (2021).