Using Artificial Intelligence to Predict Legislative Votes in the United States Congress

Anasse Bari
Courant Institute of Mathematical Sciences
Computer Science Department
New York University
New York City, USA
e-mail: abari@nyu.edu

William Brower
Courant Institute of Mathematical Sciences
Computer Science Department
New York University
New York City, USA
e-mail: wjb301@nyu.edu

Christopher Davidson
Courant Institute of Mathematical Sciences
Computer Science Department
New York University
New York City, USA
e-mail: cgd292@nyu.edu

2021 IEEE 6th International Conference on Big Data Analytics (ICBDA) | 978-1-6654-1513-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICBDA51983.2021.9403106

***Abstract*—We present in this study an experimental artificial intelligence tool to predict the likelihood of a legislative bill becoming a law. Using historical data of legislative bills, we designed an ensemble of predictive analytics algorithms that can predict whether or not a bill will pass both the Senate and the House of Representatives. Empirical results indicate that a bill's legislative vote could be predicted with an 80% accuracy using AI algorithms.**

Keywords — artificial intelligence, United States Congress, Senate, House of Representatives, legislation, machine learning, predictive analytics, ensemble algorithms.

I. Introduction

Artificial Intelligence (AI) is the art and science of designing intelligent algorithms to act as an extension of human intelligence. Predictive Analytics (a form of AI) is a technology that learns from experience (historical data) to help predict future events [1]. AI algorithms are driving decisions in several industries. In healthcare, AI algorithms have been deployed to help predict COVID-19 patients that would develop severe cases by creating an AI tool that can help medical doctors make better data driven decisions, including allocating hospital beds and ventilators [2]. In politics, AI is being applied to data from social media to help predict elections [3]. In real estate markets, AI algorithms detected novel hidden relationships that could help better predict housing markets [4,5]. In finance, AI tools are being used to mine alternative data sources including satellite imagery to better predict movements of financial markets [6,7].

In this study we research the applicability of AI in predicting votes for legislative bills in the United States Congress. The United States Congress' primary purpose is debating and enacting legislature in order to form the laws and rules which govern society. Considering the fact that their choices govern the lives of over hundreds of millions of Americans on a daily basis, developing effective legislature in a timely manner is pertinent for a productive Congress. In this research study, we aim to determine the predictive features of a bill that may ensure if a proposed piece of legislature will either pass or fail in the United States Congress.

The focus of our experiments is to design and deploy artificial intelligence algorithms that can identify characteristics of a piece of legislation, such as information surrounding the bill author and bill text among other attributes that might have a positive predictive power in determining whether or not a bill will pass.

II. The Legislative Process

The United States Congress serves as the legislative branch of the United States government and it consists of two chambers: The Senate and the House of Representatives. The upper chamber of Congress is the Senate. The United States Senate consists of two Senators representing each of the 50 states that make up the United States. Senators are elected to serve six-year-long terms, however not every Senator is elected at the same time. Instead, senatorial elections are staggered, with only one-third of the entire Senate being elected every two years. The lower chamber of Congress is the House of Representatives. Whereas only two Senators represent every state, members of the House of Representatives represent the individual districts that comprise each state based on population. The 435 members who comprise the House of Representatives are elected every two years, coinciding with the start of each new Congress. Together, Senators and Representatives comprise the 535 members of Congress. After being drafted into the proper format, a bill is presented to either chamber of Congress by at least one sponsor, who must be a member of either chamber of Congress. Additional representatives or senators may also act as cosponsors, providing additional endorsement for the proposed legislature. Upon being introduced in Congress, the proposed legislature can be sent to a congressional committee, a group of senators and representatives who specialize in a particular field relevant to the proposed bill's content. A congressional committee will debate the proposed legislature, suggesting additional amendments, substitutions, or even rejecting the measure entirely. Furthermore, the legislature can then in turn be passed to a subcommittee for further deliberation [Congress.gov]. Once it has received committee approval, the

978-0-7381-3167-2/21/$31.00 ©2021 IEEE

amended legislation is then returned to its respective congressional chamber of origin for further debate. Senators are granted nearly unlimited debate time to discuss proposed legislation. Alternatively, debate time is limited within the House of Representatives, in order to accommodate for the congressional chamber's larger size. After being debated in its chamber of origin, the proposed legislature is then voted on. If the measure receives the majority vote, then the bill pass its chamber of origin. Once a bill is approved in either the House of Representatives or the Senate, the bill is then sent to the other chamber of Congress for further consideration, wherein the process of committee, debate, sub-committee, and voting is repeated in its entirety. Ultimately, both chambers of Congress must agree upon the same version of a bill before it is sent to the President for final consideration and approval. In order for a bill to become enacted, the President must sign the proposed legislation. Likewise, in the event that a bill presented to the President remains unsigned for at least ten days while Congress remains in session, then the bill is automatically enacted as law without the President's signature. If, however Congress adjourns during this 10-day period, then the bill does not become a law in what's known as a "pocket veto" [Senate.gov]. Alternatively, the President can choose not to enact a law and veto it. The President may reject the legislature, returning it to its respective chamber of origin for additional amendments or consideration. Considering that the Presidential action stage of a bill's life cycle is complicated, for the purposes of our paper we will only be focusing on predicting whether or not a bill will pass Congress.

III. Related Work

In the recent AI literature, Predictive analytics has been applied to determine the outcome of proposed legislatures. For instance, in the work titled "Textual Predictors of Bill Survival in Congressional Committees", predictive modeling was utilized to predict if legislation would survive consideration by a congressional committee, focusing specifically on features related to a bill's respective sponsor and committee [8]. In addition, Yano et. al. augmented their dataset by focusing on textual bill features to predict several bill outcomes. In "Predicting Legislative Roll Calls from Text," Gerrish and Biel [9] use machine learning algorithms to link legislative sentiment to legislative text. Another machine learning algorithm by John J. Nay in [10] uses applied AI on bills spanning the 107th to 113th Congresses, applying word vectors to each sentence comprising a bill's text. Likewise, analyzing the predictive features of a bill's specific language to predict political support was the focus of the predictive modeled crafted by Kraft et. al. [11].

The experimental AI tool proposed in this work has three major components: legislative data ingestion and preparation, bill features engineering, and predictive algorithms design. In the following sections, we will outline each component in the tool and the experimental results.

IV. Data Understanding

A. Legislative Data Source

The primary data used to train our algorithms was collected using ProPublica's Congress API. Originally Built in 2009, this API "includes details about members, votes, bills, nominations and other aspects of congressional activity" [ProPublica.org]. We wrote data collection algorithms that call the API's Bill endpoint to build a relational database of legislative bills introduced in the most recent Congresses. While bill information is available from 1995 and onward, the algorithms we designed were trained to learn from a large dataset gathered from the three most recently completed Congresses: the 113th Congress (3 January 2013 - 2 January 2015), the 114th Congress (6 January 2015 - 3 January 2017), and the 115th Congress (3 January 2017 - 3 January 2019).

We extracted a number of different features for each bill that contain either core information about the bill or metadata for locating more information about the bill. Excluding all URIs and URLs that contained no predictive information about a bill, the features were:

- Bill ID - A concatenation of bill slug and the number of congress it was proposed in.
- Bill Type - This field indicates what chamber of congress the bill originated from.
- Number - The unique bill number assigned within each Congress.
- Title - The title of the bill given by the author.
- Sponsor Title - Describes which chamber of congress the bill's author is a member of
- Sponsor ID - A unique identification number assigned to each congressperson
- Sponsor Name - The given name and surname of the bill's author
- Sponsor State - The state or territory that the bill's author represents
- Sponsor Party - The political party of the bill's author, and can be either Republican, Democrat, or Independent.
- Introduced Date - Indicates when the bill was presented to Congress
- House Passage - Can be either true or false to indicate if the bill was passed in the House of Representatives
- Senate Passage - Can be either true or false to indicate if the bill was passed in the Senate
- Enacted - This field is null if the bill was not signed by the President or the date it was signed into law.
- Cosponsors - The number of cosponsors on the bill
- Cosponsors by Party - An array of numbers indicating the political party for each cosponsor.
- Committee - If the bill was first introduced in a specialized committee, this field indicates that committee.
- Summary - A brief summary of the bill that was proposed and written based on the original bill's text.

Ranging in length from a single sentence to multiple pages of text, bill summaries are official abstractions and outlines of proposed legislature that are provided by The Congressional Research Service (CRS) [Congress.gov]. The Congressional Research Service of The Library of Congress provides unbiased and objective analysis to members of either chamber of Congress, regardless of party affiliation. If a bill is reported by a committee or is enacted by one chamber of Congress, then the CRS analysts may expand upon their summary, detailing the enacted measure's effect upon preexisting programs and legislature. If the bill becomes a law, then the CRS provides a final bill summary. Table 1 demonstrates that only an average of three percent of all bills that were introduced in Congress in our dataset are passed by both the House of Representatives and the Senate. This closely aligns to the statistics provided by GovTrack [Govtrack.us]. About 3.96% of all legislation proposed by Republican sponsors (708 bills) were passed, 1.97% of all legislation proposed by Democratic sponsors (358 bills) were passed, and 1.00% of all legislation proposed by Independent sponsors (2 bills) were passed. The most popular times for passing legislation were in the three months preceding Congress' recess in August, as shown in Figure 1.

TABLE I. BILLS PASSED BY CONGRESS

Congress	**Bills**		
	Introduced	***Passed***	***Percentage***
113th	10,625	270	2.54%
114th	12,063	354	2.93%
115th	13,556	444	3.28%
Total	36,244	1,068	2.95%

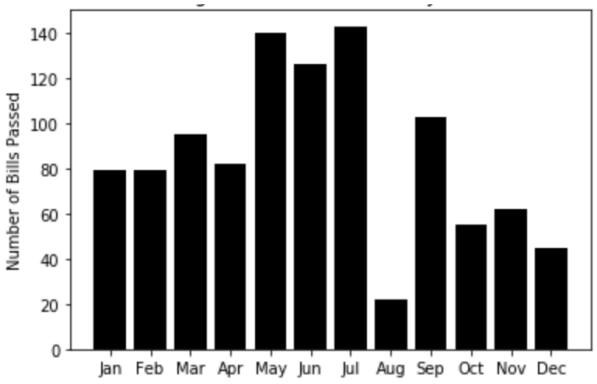

Fig. 1. Congressional Bills Passed by Month

V. DATA PREPARATION

The first step in preparing the data was to remove all features of the dataset that were either redundant, not available before the bill was voted on, or had no predictive qualities. Principle component analysis (PCA), feature correlation and variance analysis were conducted as a preliminary step to eliminate features with very low predictive power. Other features were also discarded in the analysis including: Bill Slug, Bill Number, Bill URI, Short Title, Sponsor URI, GPO PDF URI, Congress.gov URL, GovTrack URL, Active, Last Vote, Enacted, Vetoed, Committee Codes, Subcommittee Codes, Short Summary, Latest Major Action, and Latest Major Action Date.

The predictive label for each bill titled "Passed", was derived from the House Passage and Senate Passage features. The Passed label is a boolean type and was true if and only if both the House Passage and Senate Passage fields indicated that the bill passed, and false otherwise. The next step in the data preparation phase was to remove all entries with missing values for any of their features.

Our dataset had complete entries for all features except bill summaries. Out of the 36,244 bills retrieved, 3,635 had an empty bill summary. This left 32,609 bills in the dataset that could be used to train models. The dataset now contained 1,031 samples of bills that passed both the House of Representatives and the Senate and 31,578 sampled of bills that failed at least one chamber of Congress.

To address this data imbalance problem, we experimented with several sampling techniques including data augmentation, over-sampling and mini-batch sampling, similar techniques we used in our previous work in [13].

After filtering out all of the samples that had successfully passed, the remaining bills that had failed to pass one of the chambers of Congress were then randomly shuffled. From these negative samples, 1,031 entries were then randomly selected to create a balanced dataset. It was important to randomize the selection of negative samples to ensure that bills were sampled from all three of the Congresses in the dataset, as bills were listed sequentially by introduction date. The complete set to be utilized for training and evaluation contained 2,062 samples. The 2,062 samples were converted into a one-hot numeric data matrix. The last step in the data preparation process was to create the training and test datasets. We used several validation methods, and we report in this study the results from split validation. The data were split into two sets, with 70% of the samples (1,443 samples) in the training dataset and 30% of the samples (619 samples) in the test dataset.

VI. PREDICTIVE ALGORITHMS AND EXPERIMENTS

A. L2-regularized Logistic Regression

The first model chosen to predict whether a bill will pass or fail both chambers of Congress was L2-regularized logistic regression (2) [5].

$$min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\log(1+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i}) \qquad (2)$$

B. Support Vector Machine (SVM)

The second predictive model that was trained was a Support Vector Machine (SVM). Three kernels were used in training the SVM: the linear kernel (3), polynomial kernel (4),

and sigmoid kernel (5). Both the polynomial kernel and the sigmoid kernel classified all samples in the test dataset as false, while the linear SVM has a similar performance to the majority of the other models.

$$\mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1) \quad (3)$$

$$\mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d \quad (4)$$

$$\mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\alpha x^T y + c) \quad (5)$$

C. Decision Tree

We used the C4.5 algorithm to generate a decision tree.

We used the information gain measure to rank bill features with respect to the class label passed. The formula for information gain is represented in Eq. (2) Where f represents the feature variable, c represents the class label:

$$IG(f) = -\sum_{c \in \{p,n\}} P(c) log_2 P(c) + \sum_{j=1}^{m} P(v_j) \sum_{c \in \{p,n\}} P(c|v_j) log_2 P(c|v_j) \quad (2)$$

Pruning and sampling methods were applied to avoid overfilling of the classification model.

D. Neural Networks

The fourth model that was trained on the dataset was a Multi-Layer Perceptron (MLP) Classifier. This model utilized three layers of neurons: the input layer, a single hidden layer, and the output layer. The model was trained with a hidden layer of 100 neurons, and 200 iterations. The activation function that provided the best results was the rectified linear unit (ReLU) function (6). The model was also trained using the logistic sigmoid function (7) and hyperbolic tan function (8).

$$f(x) = \max(0, x) \quad (6)$$

$$f(x) = \frac{1}{1+e^{-x}} \quad (7)$$

$$f(x) = \tanh(x) \quad (8)$$

E. K-Nearest-Neighbors Algorithm

The fifth model that was trained was a K-Nearest-Neighbor (KNN) classifier. This classifier utilized the Euclidean distance between equally weighted samples and was trained from 1 to 40 neighbors. The model with the best accuracy, k = 9, was then chosen as the final KNN model as shown in Fig. 2.

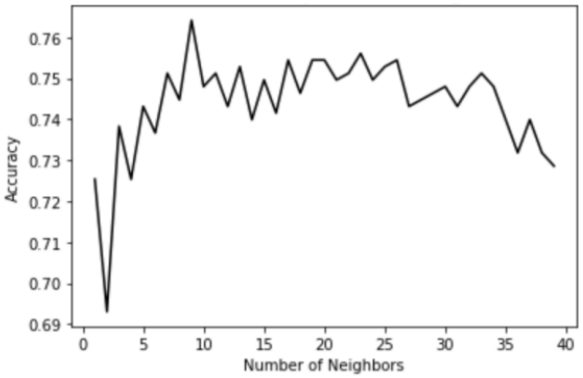

Fig. 2. K-NN Accuracy vs. Number of Neighbors (K)

F. Ensemble Methods

In Artificial Intelligence, ensemble methods aim at combining predictive decisions from multiple algorithms to improve the overall predictive accuracy [14]. Ensemble methods help minimize errors in individual machine learning algorithms due to bias and noise. We designed an ensemble model based on voting by majority rule from the best three performing models including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN). Then we rerun the experiment with cross and split validations, noticing an improvement in the accuracy.

Table 2 shows that the voting classifier based on an ensemble of three algorithms had the best performance of all the models in terms of accuracy and precision. The second-best model was the MLP neural network, based on recall.

When choosing the best performing model, we found it more important to choose a model that minimizes the number of false positives and maximizes precision, rather than the model that minimizes the number of false negatives. This tradeoff prevents bills from being voted on by Congress that could potentially fail to pass. By this metric, the model that performed the best was the MLP neural network.

TABLE II. MODEL RESULTS

Model	**Metrics**		
	Accuracy	***Precision***	***Recall***
L2-Logistic Regression	74.80%	91.99%	92.43%
Linear SVM	76.25%	93.08%	92.63%
Decision Tree	74.80%	91.77%	93.21%
MLP (ReLU)	79.16%	93.62%	93.89%
KNN, k = 9	76.41%	81.28%	77.50%
Ensemble Methods (Voting)	80.13%	94.34%	93.70%

VII. Conclusion

In this study we present an experimental AI tool that can be used to predict if a bill will pass the Senate and the House of Representatives to become a law. The tool consists of (1) a data collection process that utilizes ProPublica's Congress API to collect bills textual data and metadata that is stored in a relational database, (2) feature extraction from bill's text and feature selection algorithms and (3) ensemble methods based on three machine learning algorithms. The overall average accuracy of the ensemble algorithms was about 80% using split validation.

We noticed that in addition to the text of a bill's summary of its content, the identity features surrounding a bill, including sponsor and sponsor party, are predictive of a bill's likelihood of passing. These predictive features were mostly involving the sponsor's identity, the time of the year at which the bill is being proposed, the extracted features from bill text summary, and information about the co-sponsors of the bill. This appears to support the idea that identity politics are an important factor in deciding what bill ultimately ends on the President's desk.

As a future work, we plan to perform further validations with new bill data and extend the feature extraction to include full text of the bill as a replacement to the summary of the bills we are currently using. Another important addition to extend the tool is predicting each senator's or representative's vote on a piece of legislation, based on historical votes. In terms of the algorithms, we plan to extend the AI algorithms used in the tool to include biologically inspired algorithms [15, 16, 17] as well as research and evaluate their performance in the ensemble methods designed in this study.

Acknowledgements

This research study was partially supported by Amazon Machine Learning Research Award to Prof. Anasse Bari who leads the *Predictive Analytics and Artificial Intelligence Research Lab* at New York University.

References

[1] A. Bari, M. Chaouchi, and T. Jung, *Predictive Analytics For Dummies*. John Wiley & Sons, 2016.

[2] X. Jiang, M. Coffee, A. Bari, J. Wang, X. Jiang, J. Huang, J. Shi, et al. "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity." *CMC: Computers, Materials & Continua*, vol. 63, no. 1, pp. 537-551, 2020.

[3] A. Bari, "Big Data signaled winner days before Election Day." *The Hill*, Nov. 11, 2020. [Online]. Available: https://thehill.com/opinion/technology/526823-big-data-signaled-winner-days-before-election-day. [Accessed: Nov. 12, 2020.]

[4] R. M. Moraes, A. Bari, and J. Zhu. "Restaurant Health Inspections and Crime Statistics Predict the Real Estate Market in New York City." *International Workshop on Machine Learning, Optimization, and Big Data (LOD).* Springer, Cham, 2019.

[5] W. Watts, "How data scientists found a link between restaurant inspections and New York's wild real-estate market," *MarketWatch*, Feb. 4, 2020. [Online]. Available: https://www.marketwatch.com/story/how-data-scientists-found-a-link-between-restaurant-inspections-and-new-yorks-wild-real-estate-market-2020-02-03. [Accessed: Feb. 4, 2020].

[6] A. Bari, P. Peidaee, A. Khera, J. Zhu, and H. Chen, "Predicting Financial Markets Using The Wisdom of Crowds," *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, Suzhou, China, 2019, pp. 334-340, doi: 10.1109/ICBDA.2019.8713246.

[7] Z. Yang, Daniel Broby, "Sustainable Finance: AI Applications in Satellite Imagery and Data," *Centre For Financial Regulation and Innovation*, pp. 1-23, 2020.

[8] T. Yano, N. A. Smith, J.D. Wilkerson, "Textual Predictors of Bill Survival in Congressional Committees." *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'12).* Association for Computational Linguistics. pp. 793–802, 2012.

[9] S. Gerris, D.M. Blei, "Predicting legislative roll calls from text," *ICML'11: Proceedings of the 28th International Conference on Machine Learning*, pp. 489-496, Jun. 2011. [Online]. Available: https://doi.org/10.1371.journal.pone.0176999. [Accessed: Feb. 4, 2020].

[10] P. Kraft, Peter, H. Jain, and A.M. Rush, "An Embedding Model for Predicting Roll-Call Votes." *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2066-2070, Nov. 2016.

[11] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A Library for Large Linear Classification," *The Journal of Machine Learning Research,* Jun. 9, 2008

[12] J. Wang, R.M. de Moraes, and A. Bari. "A Predictive Analytics Framework to Anomaly Detection," *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*. Oxford, UK, 2020, pp. 104-108.

[13] A. Bari, G. Saatcioglu, "Emotion Artificial Intelligence Derived from Ensemble Learning." *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE).* New York, NY, USA, 2018, pp. 1763-1770.

[14] A. Bellaachia, A. Bari. "Flock by Leader: A Novel Machine Learning Biologically Inspired Clustering Algorithm," *International Conference in Swarm Intelligence.* Springer, Berlin, Heidelberg, 2012.

[15] A. Bellaachia, A. Bari, "A flocking based data mining algorithm for detecting outliers in cancer gene expression microarray data." *2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia, 2012, pp. 305-311.

[16] A. Bellaachia, A. Bari, "SFLOSCAN: A biologically-inspired data mining framework for community identification in dynamic social networks." 2011 IEEE Symposium on Swarm Intelligence, Paris, France, 2011, pp. 1-8.