Multi-Factor Congressional Vote Prediction

Hamid Karimi*, Tyler Derr*, Aaron Brookhouse, and Jiliang Tang
Data Science and Engineering Lab
Michigan State University, East Lansing, MI, USA
{karimiha, derrtyle, brookho8, tangjili}@msu.edu

***Abstract*—In recent times we have seen a trend of having the ideologies of the two dominant political parties in the U.S. growing further and further apart. Simultaneously we have entered the age of big data raising enormous interest in computational approaches to solve problems in many domains such as political elections. However, an overlooked problem lies in predicting what happens once our elected officials take office, more specifically, predicting the congressional votes, which are perhaps the most influential decisions being made in the U.S. This, nevertheless, is far from a trivial task, since the congressional system is highly complex and heavily influenced by both ideological and social factors. Thus, dedicated efforts are required to first effectively identify and represent these factors, then furthermore capture the interactions between them. To this end, we proposed a robust end-to-end framework Multi-Factor Congressional Vote Prediction (MFCVP) that defines and encodes features from indicative ideological factors while also extracting novel social features. This allows for a principled expressive representation of the complex system, which ultimately leads to MFCVP making accurate vote predictions. Experimental results on a dataset from the U.S. House of Representatives shows the superiority of MFCVP to several representatives approaches when predicting votes for individual representatives and also the overall outcome of the bill voted on. Finally, we perform a factor analysis to understand the effectiveness and interplay between the different factors.**

I. Introduction

Recently there has been an enormous interest in computational approaches to solve political science related problems, especially in relation to political elections and congressional voting. With the seemingly ever-growing tension between the two dominant political parties in the U.S. [1], congressional representatives are receiving immense social pressure towards blindly following their political party and associated leaders. However, due to the nature of some representatives refusing to give up their beliefs and ethical grounds, they sometimes vote against their party or cast no vote; thus resulting in a highly complex system.

Although knowing the voting behaviors in the congressional system are undoubtedly complicated, we must remain diligent towards the goal of being able to predict them. If we can construct better vote prediction models, we could utilize this information to better inform the public of the real intentions of those running for re-election on upcoming critical issues. Similarly, congressional leadership could utilize these models for specifically targeting potential swing voters. Thus, although the path to overcome the associated challenges might at first not be clear, we must continue the efforts towards the development of such a framework.

As previously mentioned, congressional voting is a complicated process and influenced by various factors. We recognize and identify two sets of effective factors. The first set being ideological factors, which are well recognized to play an important role in the U.S. congress [2], [3] and come from both the congressional representatives as well as the ideology of the bills, whose values and beliefs are woven deep into the content of the bills. The second set of influential factors are social factors and are in relation to 1) the party affiliations of representatives, and 2) how their past voting recording intertwines with other representatives. In relation to the first social factor, it is well known that representatives in the U.S. Congress are polarized [1], [4]–[7]; and thus likely to follow their political affiliation when casting their votes (although not all the time). As for the second social factor, we propose the voting records to be modeled as a signed bipartite social network (i.e., contains both positive and negative connections) between the representative and the bills [8], which opens the door to extracting a plethora of novel predictive features.

In this paper, we press onward to embrace the opportunities and challenges of congressional vote prediction. To achieve this, we propose an end-to-end framework **M**ulti-**F**actor **C**ongressional **V**ote **P**rediction (MFCVP). The proposed framework MFCVP first utilizes Wikipedia (https://www.wikipedia.org/) pages of the representatives to learn an embedding that encodes ideological information associated with each representative. As for bills, we use their texts to directly learn an embedding that encodes their semantic ideological information. Next, we utilize signed network analysis to first construct a bipartite voting network between the representatives and the bills, followed by harnessing powerful signed social theories to construct novel features. Finally, all the extracted features coming from multiple factors are combined to be utilized for vote prediction (details will be presented later). Our main contributions are as follows:

- We construct a principled solution to capture the different aspects of congressional voting behaviors by extracting

*Equal contribution and co-first authors.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08/$15.00
http://dx.doi.org/0.1145/3341161.3342884

TABLE I: Notations.

Notations	Descriptions
$\mathcal{R}$	The set of representatives.
$\mathcal{B}$	The set of past roll-call votes and their bills.
$\mathcal{V}$	The set of past votes $\mathcal{R}$ gave on $\mathcal{B}$.
$\tilde{\mathcal{B}}$	The set of future roll-call votes and their bills.
$\tilde{\mathcal{V}}$	The future votes we seek to predict.
t_i	The representative r_i when they are voting.
s_j	The sponsor of bill b_j.
c_j	The set of cosponsors for the bill b_j.
$v_{ij}(\tilde{v}_{ij})$	The vote associated with voter t_i on bill b_j ($\tilde{b}_j$).

information from multiple important factors associated with the political voting system and constructing novel features to gain better vote predictions.

- Extensive experiments are conducted to show the effectiveness of MFCVP for predicting individual representative votes and the overall outcome of the roll-call vote for new incoming bills and furthermore we perform an analysis of the impact each factor has in our framework.

The rest of the paper is organized as follows. In Section II, we formalize the problem of predicting individual congressional representative votes and define needed notations. Section III introduces our end-to-end multi-factor framework followed by experiments in Section IV. Related work is then discussed in Section V and finally we conclude the paper in Section VI.

II. Problem Statement

To introduce the problem, we first denote the set of n representatives as $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$. We let $\mathcal{B} = [b_1, b_2, \ldots, b_m]$ denote the sequence of bills associated with the past m roll-call votes for which we know the voting outcomes. These voting outcomes are denoted in the set $\mathcal{V} = \{v_{ij} | r_i \textit{ voted on bill } b_j\}$ and $v_{ij} \in \{+, -, o\}$, which denotes a "yea", "nay", or "present"/"no vote", respectively. Furthermore, we have the sequence of $\tilde{m}$ future roll-call bills denoted as $\tilde{\mathcal{B}} = [\tilde{b}_1, \tilde{b}_2, \ldots \tilde{b}_{\tilde{m}}]$. The sequence $\tilde{\mathcal{B}}$ has corresponding votes $\tilde{\mathcal{V}} = \{\tilde{v}_{ij} | r_i \textit{ will vote on bill } \tilde{b}_j\}$ which we seek to predict . Lastly, we denote any additional contextual feature or those extracted from the past votes as the set $\mathcal{X}$. Note that these notations and others used throughout the paper can be found in Table I.

With the above definition we can formally define the congressional vote prediction problem as follows:

Given a set of congressional representatives $\mathcal{R}$, a sequence of past roll-call votes on the bills $\mathcal{B}$ having associated votes $\mathcal{V}$, features $\mathcal{X}$, and a future sequence of the upcoming roll-call votes on the bills $\tilde{\mathcal{B}}$, we seek to learn a model F as follows:

$$F : \{\mathcal{R}, \mathcal{B}, \mathcal{V}, \mathcal{X}, \tilde{\mathcal{B}}\} \rightarrow \tilde{\mathcal{V}} \tag{1}$$

III. The Proposed Framework

For congressional vote prediction, we must overcome the challenges of how to represent the underlying factors influencing the voting system and how to handle this added complexity introduced by incorporating multiple factors. To address these we propose the end-to-end framework Multi-Factor Congressional Vote Prediction demonstrated in Figure 1. In this section, we first explain how different factors are represented through both learning embeddings and constructing novel hand-crafted features. Thereafter, we discuss how the representations of different factors are combined and used for the vote classification.

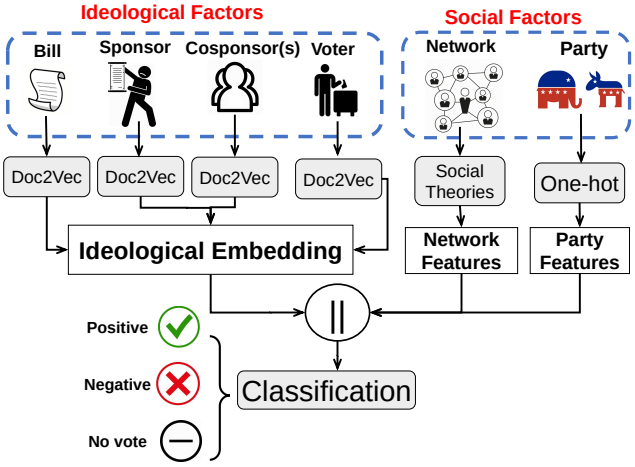


Fig. 1: The proposed Multi-Factor Congressional Vote Prediction framework (MFCVP)

A. Ideology Factors

The first set of factors is ideology factors. It is without a doubt that representatives' ideology and ideological information reflected in a bill are influencing how a voter will vote on a bill. To effectively and comprehensively represent ideology factors in our framework, we recognize and propose the use of two other entities (besides a bill and a voting representative) which are associated with ideology factors, namely the sponsor and possible cosponsor(s). These two entities are essentially representatives who construct and promote a bill. Hence, we seek to learn representations about the beliefs and values of the voters, sponsors, and cosponsors, along with those that are present in the bills.

To represent the representatives, many previous works focused on ideal point models [2], [9], [10]. Nevertheless, ideal point methods require many assumptions about voter behaviors which are inherently highly complex, so instead, it seems more natural and reasonable to extract a vector representation from the raw data [3], [5] (e.g., Wikipedia pages that are collectively written about the representative from the large online community). Furthermore, extracting vector representations are practically more feasible than attempting to compute ideal points [2] which are also open for biases in their human construction. Given the more recently developed deep models for extracting meaningful representations for text documents, we propose to utilize doc2vec [11] as an efficient embedding method to represent the ideological factors. Doc2vec has shown significant improvement in many approaches [12], [13].

We use Wikipedia pages to learn a representation for each of the congressional representatives using doc2vec as illustrated in Figure 1. We combine all textual information about a representative from their Wikipedia profile page as a single docu-

ment. Then, we train a doc2vec model which learns a compact embedding about each entire document (i.e., a representative's Wikipedia page) encoding the semantic information about a representative including their political ideology. Due to the fact that voters, sponsors, and cosponsors are all representatives, we utilize the same representation obtained through the learned embeddings of our trained doc2vec model (i.e., the doc2vec model fed with Wikipedia pages of representatives) as the ideology factors for the representative in all three roles. We should emphasize that in our experiments we utilize historical Wikipedia pages to ensure there is no data leakage.

The usefulness of Wikipedia is that this ideological perspective is less susceptible to biases or falsehoods since it is maintained by a large community. However, other data sources could be used to obtain the ideological representation, such as the generated content of voters on social media (e.g., their tweets on Twitter (http://www.twitter.com) or their campaign financial information as to which organizations are supporting them. We leave connecting other sources of data about the congressional representatives as one future work. Finally, we let E_{t_i}, E_{s_j}, and E_{c_j} denote embeddings of the voter $t_i \in \mathcal{R}$, the sponsor $s_j \in \mathcal{R}$ sponsoring the bill $b_j \in \mathcal{B} \cup \tilde{\mathcal{B}}$, and the cosponsor $s_j \in \mathcal{R}$ cosponsoring the bill $b_j \in \mathcal{B} \cup \tilde{\mathcal{B}}$ for the votes $v_{ij} \in \mathcal{V} \cup \tilde{\mathcal{V}}$.

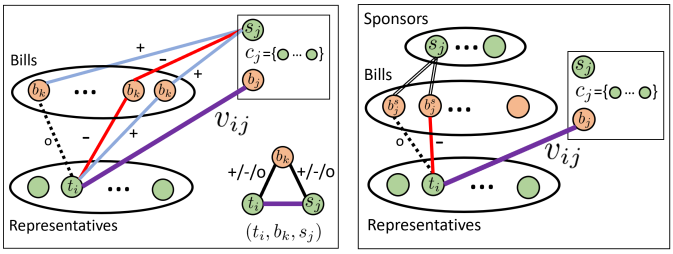
The textual content of the bill offers very essential information. In fact, the text of a bill reflects both the conscious and sometimes even subconsciously instilled ideologies of the sponsor and cosponsors who prepared it. Therefore, it is of great importance to effectively represent the semantic information about a bill in a compact and efficient way. To achieve this, similar to our embeddings for the representatives, we utilize a doc2vec model to represent the bills, where each bill's textual data (after some preprocessing) is considered as a document. Let E_{b_j} be the learned embedding of the bill $b_j \in \mathcal{B} \cup \tilde{\mathcal{B}}$. Note that we train the bill doc2vec model on $\mathcal{B}$.

We can now succinctly represent the set of embedded ideological features that we will utilize when considering the relation between a voter t_i and a bill b_j (along with their sponsor and cosponsor(s), s_j and c_j, respectively) as $\mathcal{E}_{ij} = \{E_{t_i}, E_{b_j}, E_{s_j}, E_{c_j}\}$.

B. Social Factors

Having discussed the ideological factors that get incorporated into MFCVP, here we discuss the more novel social factors (with an emphasis on the network features) that have been commonly overlooked by previous methodologies and analyses in relation to the predictions and understanding of congressional votes. We propose to categorize these social factors into two main groups as follows: 1) political party affiliation features, and 2) features coming from the network constructed from the past voting records. Next we discuss these two feature categories.

1) Party Features: The inspiration of these features for our proposed framework comes from the fact that sometimes there is an influence coming from voters of a political party to cast their votes aligned with the party's interest.



(a) $N^B_{t_i s_j}$ (b) $N^S_{t_i s_j}$

Fig. 2: Illustrations of the signed network features.

Given a single vote v_{ij} made by voter t_i on bill b_j that was sponsored by s_j and cosponsored by the set of representatives c_j, we construct the corresponding features P_{t_i}, P_{s_j}, and P_{c_j} to represent their party affiliations, respectively. More specifically, P_{t_i} and P_{s_j} are one-hot vectors indicating the affiliated party of the voter and sponsor, respectively. Then for the set of cosponsors c_j we obtain the distribution of the cosponsors across the party affiliations. Note that if there are no cosponsors, we simply use a vector of zeros for P_{c_j}. These three features are represented in the set of features $\mathcal{P}_{ij} = \{P_{t_i}, P_{s_j}, P_{c_j}\}$.

2) Network Features: Typical network representations that are used for congressional voting records are the two one-mode networks coming from a bipartite network which ultimately separates and/or condenses the "yea" and "nay" votes [14]. However, this is inherently destined to lose drastic amounts of vital information that could have perhaps been extracted if using network analysis techniques that incorporate the "yea" and "nay" votes simultaneously. Therefore, we propose a more advanced representation - signed bipartite network.

Let $\mathcal{G} = \{\{\mathcal{R} \cup \mathcal{B}\}, \mathcal{V}\}$ denote the signed bipartite network that is constructed using the set $\{\mathcal{R} \cup \mathcal{B}\}$ of $n + m$ nodes (i.e., the representatives and bills), and set of links (i.e., votes $\mathcal{V}$) between them where we treat "yea", "nay", and "no vote" as a positive, negative, and non-existent link in the signed network. Now, given that we have modeled the voting history in the form of a signed network, we can utilize signed social theories to extract insightful features. More specifically, we utilize balance theory, which colloquially can be summarized as "a friend of a friend is a friend" while "an enemy of a friend is an enemy" [15], [16].

The first set of features we construct when considering the relationship between a voter t_i and a bill b_j can be seen in Figure 2a. We can observe that we want to extract information on how the voter t_i and the sponsor s_j have interacted together on other bills b_k to gain information on how t_i might vote on the current bill b_j. We note that there can be 9 possible situations when considering the triplet (t_i, b_k, s_j), since both t_i and s_j can have either a positive, negative, or no link to the other bills $b_k \in \mathcal{B} \backslash \{b_j\}$. We utilize this information to construct a feature vector $N^B_{t_i s_j}$ that represents the distribution over the nine aforementioned possibilities where the triangles involving an even number of negative links are adhering to balance theory. The distribution over the number of balanced and unbalanced triangles along with the number of open

structures (i.e., those involving at least one "no link") should provide great insight for our model to discover the patterns related to this fundamental social theory. Signed triangle distributions have also recently been used in benchmarking generative signed network models [17], since they hold such rich information about a signed network.

We note that these features are similar to the ones utilized in the seminal work [18] that focused on building a supervised model to predict the missing sign between t_i an s_j, but here we use s_j as a proxy for their introduced bill b_j. This relates to balance theory because the signed social theory would suggest that if t_i has voted equally to s_j (i.e., $v_{ik} = v_{jk}$), then it is likely that t_i should think positively towards b_j. Similarly, we construct a feature vector $N^B_{t_i c_j}$ where instead of using s_j, we obtain the average over the cosponsors in the set c_j.

We furthermore extract the second type of feature from our constructed signed network. In the first network feature (described before), we sought to discover how the overall distribution of balance between the votes from the voter t_i and the current sponsors and cosponsors (i.e, s_j and c_j) towards the rest of the bills b_k. However, unlike the first features, here we want to directly observe how t_i has interacted on the bills sponsored by s_j or sponsored by someone in c_j (i.e., a more personalized set of social features), which is related to the polarity of their interactions in the signed network [19]. In Figure 2b we show an illustration for how we construct the feature vector $N^S_{t_i s_j}$ having length 3. Given the fact that we want to extract information about how t_i might vote on b_j, we observe the distribution over the three possible votes (i.e., positive, negative, or no link in terms of the signed network) that t_i has given to all other bills b^s_j that were also sponsored by s_j. Similarly, we construct the feature vector $N^S_{t_i c_j}$, but rather than observing the vote distribution over the set of bills b^s_j, instead, we average over b^c_j, which denotes the set of bills sponsored by the cosponsors c_j (who has cosponsored b_j).

Finally, we construct the full set of network features $\mathcal{N}_{ij} = \{N^B_{t_i s_j}, N^B_{t_i c_j}, N^S_{t_i s_j}, N^S_{t_i c_j}\}$, where $|\mathcal{N}_{ij}| = 24$. Note that these network features are in fact general and if given additional context (e.g., the connections between the voters, sponsors, and cosponsors on Twitter), we could easily extend these ideas to obtain a larger social context between the representatives; we leave this as future work along with the use of advanced signed network embeddings [20].

C. Classification

Now that we have discussed all the features coming from multiple factors, we next discuss how we can utilize them together for training a model for congressional vote prediction. We note that our framework is flexible in that the choice of the classifier is not fixed and can be chosen based on the desired outcome. One choice is to utilize a random forest [21] since it is typically an easy off-the-shelf model to train and also have the added benefit of being interpretable. More specifically, feature importance can be calculated from this model that can give insight into which features are more important for the correct classification of the votes (this will be shown in Section IV-E). Another choice could be made to utilize the power of deep learning [22] for obtaining perhaps better performance in prediction, but losing the ease of interpretation (although we note that interpreting deep neural networks is a current hot topic field in itself). In this work, we utilize both random forest and a deep neural network as classifiers.

TABLE II: Dataset Statistics.

113th House of Representatives	Total Dataset	Train (80%)	Dev. (10%)	Test (10%)
# roll-call votes	499	400	49	50
# total "Yea" votes	137,926	110,882	12,407	14,637
# total "Nay" votes	68,487	54,874	7,790	5,823
# total "Present"/No votes	8,929	6,934	902	1,093

IV. Experiments

To evaluate the performance of the proposed framework MFCVP, we conduct a set of experiments for predicting individual representative votes and the overall outcome of the roll-call vote for a set of new incoming bills when giving a training set of historical information. Through the conducted experiments, we seek to answer the following research questions:

Q1. How does the proposed framework perform on congressional vote prediction?

Q2. How different factors contribute to the congressional vote prediction?

Next, we describe the dataset followed by experimental setting. Then, we describe the baselines methods and comparison results. We conclude this section by presenting experiments and discussions on factor analysis.

A. Data

For our experiments, we have focused on the 113th U.S. Congress House of Representatives. We collected the roll-call vote data along with the sponsor, cosponsor, and party affiliation from the Govtrack database (https://www.govtrack.us). After obtaining this dataset, we filtered out the roll-call votes not associated with a bill, joint resolution, concurrent resolution, or a simple resolution; for example, roll-call votes related to amendments are not included in our dataset. We obtained ideological embeddings for each of the bills based on the bill's text, which we obtained from the Library of Congress (https://www.congress.gov). Ultimately, we split the dataset chronologically into three sets i.e., a train set, a dev set, and a test set as shown in Table II. The training set is constructed with roughly the first 80% of the roll-call votes and all happened before March 5, 2014. Thus, as we mentioned before, to ensure no data leakage, we searched the historical Wikipedia profile pages for each of the representatives to find the date closest to but before March 5, 2014; this data was then collected and used to obtain our ideological embeddings.

B. Experimental Settings

First, we obtain the results for the prediction of individual representative votes. Next, we utilize these individual vote

predictions to get the aggregated prediction as to whether the roll-call vote will pass or fail (which is the overall outcome of the roll-call vote). Since our MFCVP framework is flexible in utilizing different classifiers, we utilize a random forest and a deep neural network. For random forest we utilize the scikit-learn library (https://scikit-learn.org) and we used the PyTorch library (https://pytorch.org) for our neural network implementation. We denote these two as variants of our framework as MFCVP_RF, and MFCVP_NN, respectively. For the random forest, we use the library default settings. For the deep neural network model (see Figure 1), we employ a multi-layer fully connected network with Leaky ReLU (Rectified Linear Unit) [23] as the non-linear activation function. Hyperparameters are set by the grid search via evaluating the framework on the dev set. Using the grid search, the number of layers is set to 5 with 100 hidden units and no regularization is utilized. We utilize ADAM [23] as the optimization algorithm whose learning rate starts from 0.01 and is adjusted dynamically every 100 optimization steps with the decay rate of 0.9. Each simulation is run 2000 steps with the batch size of 100 votes at each step. The embedding size of doc2vec model is set to 50. We repeat each simulation five times and report the average F1 score and accuracy in regards to the test set. Our code and data are available at https://github.com/DSE-MSU/MFCVP.

C. Baselines

To show the effectiveness of our proposed framework MFCVP, we present a set of baseline congressional vote prediction methods and discuss why we have selected these baselines from a political standpoint.

Random Guess: This method performs a random guess when presented with a vote $\tilde{v}_{ij} \in \tilde{\mathcal{V}}$ to predict for voter t_i on a future bill $\tilde{b}_j$. The random guess is based on the class distribution of "yea", "nay", and "no vote" from the set of past votes $\mathcal{V}$. This method is selected to just give context into how difficult this problem is as compared to the most naïve approach.

Personalized Random Guess: Extending the Random Guess method, here rather than a global class distribution, we extract a personalized class distribution for each voter. In other words, to guess the vote $\tilde{v}_{ij} \in \tilde{\mathcal{V}}$ we extract the class distribution from the set $\{v_{ik} | \forall b_k \in \mathcal{B}$ and $v_{ik} \in \mathcal{V}\}$. This method is used to test if indeed individual voters have their own unique patterns in terms of their vote distribution (e.g., one representative might abstain and not vote significantly more often than another).

Party Voter: This method forces all the representatives to vote aligned with the political parties. More specifically, for predicting a vote $\tilde{v}_{ij} \in \tilde{\mathcal{V}}$ if the voter t_i has the same party affiliation of the sponsor $\tilde{s}_j$ of bill $\tilde{b}_j$, then we predict "yea" and otherwise we predict "nay".

Sponsor Biased Voter: Given a vote $\tilde{v}_{ij} \in \tilde{\mathcal{V}}$ to be predicted, first the sponsor $\tilde{s}_j$ is obtained from $\tilde{b}_j$, and then we obtain the set of all past votes $\{v_{ik} | v_{ik} \in \mathcal{V}$ and $\tilde{s}_j$ is the sponsor of $b_k \in \mathcal{B}\}$. This represents the votes that voter t_i has given on past bills b_k that were also sponsored by $\tilde{s}_j$ and we choose the highest vote type over the class distribution. The Sponsor Biased Voter does not necessarily adhere to the political affiliation when voting, but they base their vote on their past experiences with the sponsor of the current bill. In other words, if they have liked (i.e., voted "yea") the past bills of this sponsor, then they will again vote "yea", having similar reasoning for voting "nay" or "no vote" on a bill.

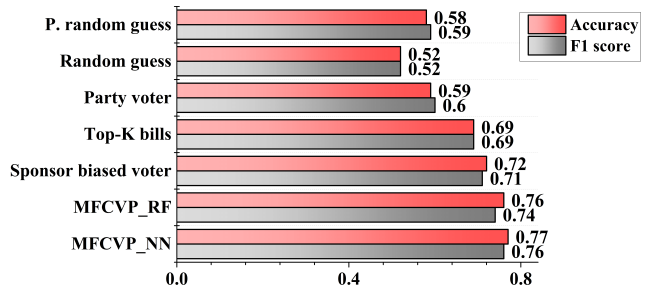


Fig. 3: Performance evaluation of the proposed MFCVP framework on predicting individual representative votes when compared against the baseline methods.

Top-K Bills: When seeking to predict the vote $\tilde{v}_{ij} \in \tilde{\mathcal{V}}$ this method first obtains the ideological bill embedding E_{b_j} and then finds the closest K bills $b_k \in \mathcal{B}_k$ based on their embeddings E_{b_k}. Top-K Bills method solely bases their vote on the ideological factors of the proposed bills text. That is to say, predicting the votes using the Top-K Bills method ignores all direct or indirect party affiliations and allows the voter to cast their vote only based on their ideologies. To select hyperparameter K, we varied the value of K in the set $\{1, 3, 5, 8, 10, 20, 30\}$ while predicting on the dev set; which resulted in $K = 8$ being the best performing value. We utilized the Euclidean distance for determining the closest K bills based on their embeddings.

D. Comparison Results

To answer the research *Q1*, we compare the proposed framework MFCVP with the representative baselines for both the local individual representative vote level and also for the global overall roll-call vote. Similar to MFCVP variants, we repeat the Random Guess, Personalized Random Guess methods 5 times and report the average F1 score and accuracy (since they are non-deterministic methods).

1) Individual Representative Vote Predictions: The results are shown in Figure 3. Based on the results presented in this figure, we make the following observations:

- Among the baselines methods, sponsor voter approach outperforms the others. This shows the fact that the historical relations between a voter and sponsor have a significant impact on determining the vote status of a voter for an upcoming bill. Further, as described before, our proposed framework, unlike sponsor voter method, incorporates these relations in a sophisticated way by extracting more principled features from the constructed signed network.

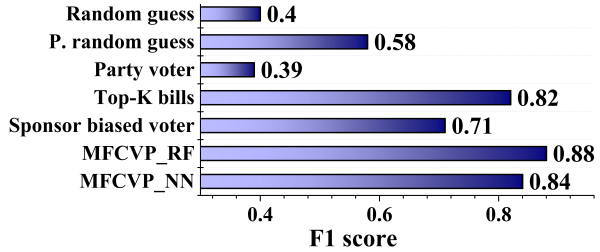


Fig. 4: Performance evaluation of the proposed MFCVP framework on predicting the overall roll-call vote outcome when compared against the baseline methods.

- Comparing Top-K bills method with party voter, we can note that the content of a bill is more important than blindly voting based on a bill's sponsor party. In fact, the low performance of party voter method supports the argument that despite the polarized voting behavior of the U.S. Congress, some representatives adhere to their prior beliefs and ideology instead of merely always voting with or against a proposed bill based on the sponsor's political affiliation.
- Personalized random guess outperforms the random guess. This is not surprising, as personalized random guess incorporates, not effectively though, the prior history of how a representative voted on past bills.
- The variants of the proposed framework MFCVP outperform all baselines methods and in some cases very significantly. This framework, in a comprehensive and sophisticated manner, incorporates various influencing political factors associated with congressional voting. Although MFCVP_NN achieves slightly better performance than MFCVP_RF, we opt to use the random forest for the rest of experiments since it provides with more interpretable insights into the proposed factors.

Therefore, from a local congressional vote perspective, this shows that MFCVP can be utilized as a reliable congressional vote prediction framework. Next, we investigate the global predictions as to whether MFCVP can accurately detect when a proposed bill will pass or fail.

2) Overall Roll-call Vote Predictions: Here, we utilize the predictions from the local level (i.e., the individual representative vote predictions) to obtain the overall global roll-call vote outcome of whether the bill will pass or fail. The results are shown in Figure 4. Based on the results presented in this figure, we make the following observations:

- The first observation is that although the personalized random guess performed worse than the party voter for determining individual representative votes, here it significantly outperforms the party voter method. This means that for individual representative votes the better predictor is based on their political party. However, when aggregating all representative votes to the prediction of whether the bill with pass/fail, using the representatives previous voting patterns is better than just considering their party.
- Next we observe a similar swap in that the Top-K bills is now outperforming the Sponsor Voter model. This is interesting since it implies that the overall pass/fail decisions for roll-call votes are happening more likely due to the correlation the voted upon bill has with past similar bills as compared to the relationship all the voters have with the sponsor of the bill. More specifically, this indicates two phenomena: 1) the representatives are quite stable in their ideologies; and 2) when averaged out, the prediction of whether a bill will pass or fail is better predicted through the representatives history according to the bill content as compared to their relation to the sponsor of the proposed bill.
- Although the MFCVP_ NN is better able to predict the local individual representative votes better, it is likely to have slightly overfit the training data (since the models were trained on the local voting patterns) and thus cannot generalize as well when aggregated to the global level. However, when pairing the random forest model with our MFCVP framework (i.e., MFCVP_ RF), we see it enjoys a better generalization over the neural network variant.

Therefore, based on the results for both the local individual representative predictions as well as the global pass or fail aggregated predictions for the proposed bills, it is clear that our MFCVP framework is a superior and effective methodology for predicting the congressional votes.

E. Political Factor Analysis

The research question *Q2* is concerned with the contribution of political factors for congressional vote prediction. To answer this question, we conduct some experiments for the local individual representative congressional vote prediction. We focus our attention on using the random forest (i.e., MFCVP_RF) as it provides us with feature importance values in an explainable manner.

First, we compute the importance of the three essential factors in our framework i.e., ideological, network and party factors (the latter two are social factors) using the Gini importance [24]. Figure 5 shows the importance of these three factors where *Embeddings* indicate the contribution of ideological factors. Based on this figure, we make the following observations:

- Embeddings (i.e., ideological factors) are the most important features in individual vote prediction. This shows that ideological factors play a central role in determining a vote cast on a bill and many representatives adhere to their ethics and beliefs.
- Quite interestingly, network features turn out to be very important. This indicates 1) the interactions and connections among U.S. House of Congress representatives have a significant bearing on the voter's voting behavior, and 2) any political vote prediction should not merely focus on ideological factors and avoid overlooking the role of social networks established among representatives and their historical votes, since it is quite effective in vote prediction.

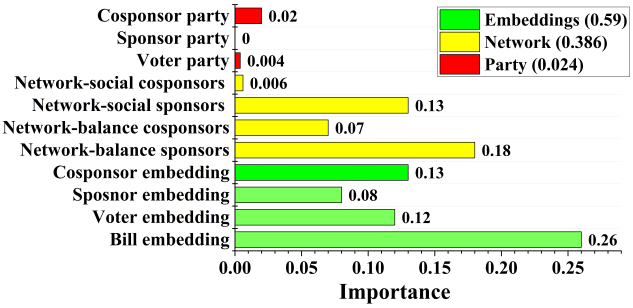


Fig. 5: Feature analysis when aggregating per feature category using the feature importance values from MFCVP_RF.

- In line with the performance of the Party Voter baseline (see Figure 3), party features have an inconsequential effect on individual vote prediction. This is politically reassuring as representatives do not submissively follow the inclination of the political party of a bill's sponsor. In the future, we will follow up more on this line of research to investigate if such a phenomenon persists at other points in time throughout history in the U.S. House and Senate or even in other country's political systems.

Now, we narrow down the feature analysis illustrated in Figure 5 to investigate contributing features in each of the three overall factors in more detail. We make the following observations according to the results shown in this figure:

- Among the ideological factors, the bill embedding has the highest contribution. This seems reasonable since, after all, it is a bill that is being voted on.
- Interestingly and somehow surprisingly, the embeddings associated with cosponsors are more effective than those of sponsors. It is known that over half of bills being introduced into U.S. Congress are cosponsored [25]. Therefore, based on this fact, it allows our model to categorize whether a given bill has received no cosponsors, or when aggregated across all cosponsors the average representative embedding can provide insight into whether it has received bipartisan support, or only from a single party. This is due to the fact that the embeddings of the representatives are designed such that they hold their ideology and thus likely easily separable in the embedded space for our model.
- Almost the entire contribution of the party features (though very insignificant compared to other factors) stems from the cosponsors party of bills. Similar to ideological factors, this indicates that cosponsors play an important role whose even party affiliation should be taken into account, but as expected, the learned embeddings about the representative's ideology are significantly more important than just knowing the political party they are associated with.
- However, for the network features, we observe the opposite as compared to the embedding and party features in that here the sponsors have a stronger signal. This is likely because aggregating over all the votes a representative has given to a bill proposed by any of the current cosponsors (which tends to be a very large set of votes) results in a more noisy signal as compared to the party or embeddings features (which is just on the order of the number of cosponsors) that can retain most of the information.
- When comparing between the two network features, we observe that the features related to balance theory were more insightful than the social ones that looked at how a voters behavior was with past proposed bills by the same (co)sponsor. One reasoning for this is that the balance theory based features are more principled and looking at a more global view, as compared to the local social features that only look at bills previously proposed by that (co)sponsor. Also, this is likely due to the fact our balance related features are based on pseudo-triangles we extracted from our constructed signed bipartite network (that we note naturally does not contain triangles) and are related to the features extracted in [18] where they were observed to be well suited for predicting whether the sign of a missing link would be positive or negative.

V. Related Work

Our work has focused on the problem of predicting future congressional votes using our end-to-end multi-factor framework MFCVP. Some of the early work on congressional roll-call vote analysis focused on using Bayesian statistical methods [2]. Later in [5], a thorough investigation showed that ideal point models were lacking as compared to directly utilizing the bills to provide a natural vector representation. More recently some work has begun to incorporate and link other metadata into the analysis and predictions of congressional votes [3], [26], [27]. In [28] they focused their attention on using Twitter to analyze how representatives interact on social media and how this correlated with their voting habits. There recently has been a few works that have also modeled the roll-call vote history as a signed network where they wanted to either investigate the correlation clustering problem for the Brazilian Chamber of Deputies [29] or European Parliament [30], focus on the analysis of communities [6], or analyzing their structural balance [8]. Other related work was to specifically predict votes on the North American Free Trade Agreement (NAFTA) [31], whether or not proposed bills will ever even make it to a roll-call vote [32], and attempting to understand the coalitions over time and the stability of the government when parties split [7]. Finally, similar to our approach, hybrid models wherein simultaneously multiple factors are taken into account are increasingly being employed in different machine learning areas [33]–[36].

VI. Conclusion

In this paper, we identified multiple broad and comprehensive factors for predicting congressional votes, namely ideological and social factors. We introduced our end-to-end framework MFCVP that represents the ideology factors of the representatives and bills using doc2vec models fed with their Wikipedia profile pages and bill texts, respectively. For the social factors, MFCVP constructs a signed bipartite

network from the representatives' historical voting behaviors to extract principled features utilizing social balance theory. Lastly, MFCVP takes into account the party affiliations as a final social factor. We observed, after conducting extensive experiments, that our MFCVP framework is able to achieve superior performance at both the local individual representative vote prediction as well as at the global roll-call vote prediction when compared to several representative baselines.

There exist several future directions to extend our MFCVP framework. First, the ideological factors can be extended to incorporate more social media sources (e.g., Twitter or Facebook). Second, we will focus our attention on obtaining more advanced network features such as global signed node similarities [37] and investigating the applicability of other signed social theories, such as status theory [18], to gain better prediction accuracy.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF) under grant numbers IIS-1714741, IIS-1715940, IIS-1845081 and CNS-1815636, and a grant from Criteo Faculty Research Award. We would like to thank our reviewers and members of the Data Science and Engineering (DSE)[1] lab at Michigan State University for their constructive comments.

REFERENCES

[1] Z. P. Neal, "A sign of the times? weak and strong polarization in the us congress, 1973–2016," *Social Networks*, 2018.

[2] S. Jackman, "Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking," *Political Analysis*, vol. 9, no. 3, pp. 227–241, 2001.

[3] P. Kraft, H. Jain, and A. M. Rush, "An embedding model for predicting roll-call votes," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2066–2070.

[4] N. McCarty, K. T. Poole, and H. Rosenthal, *Polarized America: The dance of ideology and unequal riches*, 2016.

[5] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 2, pp. 355–370, 2004.

[6] M. A. Porter, P. J. Mucha, M. E. Newman, and C. M. Warmbrand, "A network analysis of committees in the us house of representatives," *Proceedings of the National Academy of Sciences*, vol. 102, no. 20, pp. 7057–7062, 2005.

[7] C. Dal Maso, G. Pompa, M. Puliga, G. Riotta, and A. Chessa, "Voting behavior, coalitions and government strength through a complex network analysis," *PloS one*, vol. 9, no. 12, p. e116046, 2014.

[8] T. Derr and J. Tang, "Congressional vote analysis using signed networks," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 1501–1502.

[9] S. Gerrish and D. M. Blei, "Predicting legislative roll calls from text," in *International Conference on Machine Learning*, 2011, pp. 489–496.

[10] I. S. Kim, J. Londregan, and M. Ratkovic, "Voting, speechmaking, and the dimensions of conflict in the us senate," in *Annual Meeting of the Midwest Political Science Association*, 2014.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[12] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[13] H. Karimi, C. VanDam, L. Ye, and J. Tang, "End-to-end compromised account detection," in *International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2018, pp. 314–321.

[14] C. Andris, D. Lee, M. J. Hamilton, M. Martino, C. E. Gunning, and J. A. Selden, "The rise of partisanship and super-cooperators in the us house of representatives," *PloS one*, vol. 10, no. 4, p. e0123507, 2015.

[15] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory." *Psychological review*, vol. 63, no. 5, p. 277, 1956.

[16] F. Heider, "Attitudes and cognitive organization," *The Journal of psychology*, vol. 21, no. 1, pp. 107–112, 1946.

[17] T. Derr, C. Aggarwal, and J. Tang, "Signed network modeling based on structural balance theory," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 557–566.

[18] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *International Conference on World Wide Web*. ACM, 2010, pp. 641–650.

[19] T. Derr, Z. Wang, and J. Tang, "Opinions power opinions: Joint link and interaction polarity predictions in signed networks," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 363–366.

[20] T. Derr, Y. Ma, and J. Tang, "Signed graph convolutional networks," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 929–934.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*, ser. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[25] R. K. Wilson and C. D. Young, "Cosponsorship in the us congress," *Legislative Studies Quarterly*, pp. 25–43, 1997.

[26] S. Smith, J. Y. Baek, Z. Kang, D. Song, L. El Ghaoui, and M. Frank, "Predicting congressional votes based on campaign finance data," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 1. IEEE, 2012, pp. 640–645.

[27] A. Kornilova, D. Argyle, and V. Eidelman, "Party matters: Enhancing legislative embeddings with author attributes for vote prediction," *arXiv preprint arXiv:1805.08182*, 2018.

[28] T.-Q. Peng, M. Liu, Y. Wu, and S. Liu, "Follower-followee network, communication networks, and vote agreement of the us members of congress," *Communication Research*, vol. 43, no. 7, pp. 996–1024, 2016.

[29] M. Levorato and Y. Frota, "Brazilian congress structural balance analysis," *Journal of Interdisciplinary Methodologies and Issues in Science*, 2017.

[30] N. Arinik, R. Figueiredo, and V. Labatut, "Signed graph analysis for the interpretation of voting behavior," *arXiv preprint arXiv:1712.10157*, 2017.

[31] I.-B. Kang and K. Greene, "A political economic analysis of congressional voting patterns on nafta," *Public Choice*, vol. 98, no. 3-4, pp. 385–397, 1999.

[32] T. Yano, N. A. Smith, and J. D. Wilkerson, "Textual predictors of bill survival in congressional committees," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 2012, pp. 793–802.

[33] H. Karimi, P. Roy, S. Saba-Sadiya, and J. Tang, "Multi-source multi-class fake news detection," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1546–1557.

[34] H. Karimi, J. Tang, and Y. Li, "Toward end-to-end deception detection in videos," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1278–1283.

[35] H. Karimi, C. VanDam, L. Ye, and J. Tang, "End-to-end compromised account detection," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 314–321.

[36] C. VanDam, P.-N. Tan, J. Tang, and H. Karimi, "Cadet: A multi-view learning framework for compromised account detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 471–478.

[37] T. Derr, C. Wang, S. Wang, and J. Tang, "Relevance measurements in online signed social networks," in *Proceedings of the 14th International Workshop on Mining and Learning with Graphs (MLG)*, 2018.

[1] http://dse.cse.msu.edu/