There is a growing body of research addressing the task of using machine learning models to predict roll call votes made by legislators in a parliamentary context. Much of the existing literature has focused on the U.S. Congress, which provides an interesting opportunity to consider this task in an Irish context. This was highlighted by DCU researchers in a 2018 article for RTE, who noted with the availability of open data from the Oireachtas combined with recent developments in machine learning techniques such research would be particularly interesting to undertake because “Irish politics is evolving and […] we may gain greater insights into exactly how our politicians are operating” (Courtney et al, 2018). While there has not yet been a study predicting the results of votes in Dáil Éireann based on the features of the preceding debate, Sinnott and O’Reilly (2010) used an SVM classifier to study whether Irish parliamentary questions were more focused on national or local issues and a follow up study by Leheny (2018) used a Non-negative Matrix Factorisation algorithm to extract topics from parliamentary questions.

The study most closely comparable to my proposal is that of Abercrombie and Batista-Navarro (2020). This work introduced ParlVote, an annotated corpus of debate speeches from the UK parliament, with each entry associated with the speaker’s subsequent vote. They explored using this data to train several types of models to predict votes based on parliamentary speech. They used two techniques to transform the speech text into training features: a simple bag-of-words approach and a more complex approach using pre-trained BERT embeddings. They trained support vector machine (SVM) and artificial neural network (ANN) models to predict the vote based on the preceding speech. They further explored adding contextual features of whether the speaker was in government or opposition and the text of the motion being voted on. All their models outperformed a naive majority class classifier, and their best results combined the SVM model with the feature set including government/opposition context, giving up to 68% accuracy. They found the bag-of-words technique an effective way of transforming political speech into a training feature set.

A second highly relevant study was conducted by Budwhar (2018), who assembled a data set of transcribed speeches from the California legislature which was associated with votes on legislation. Budhwar tested six different sentiment analysis tools and compared three machine learning models (SVM, random forests, and TensorFlow) for their ability to predict vote outcomes. The feature sets used for training were a combination of structural features of the speech (number of interruptions, individual word count relative to overall debate word count, number of questions asked) and sentiment features (overall sentiment, number of negative/positive statements). The findings indicated that TextBlob was the best sentiment analysis tool for this data and of the models, TensorFlow achieved the highest accuracy at 83%.

There have been a range of studies which attempted to predict legislative votes using other methodologies, with earlier studies borrowing more heavily from political science and using versions of ideal point models/Latent Dirichlet Allocation (LDA). Gerrish and Blei (2011) predicted votes in both houses of U.S. Congress based on bill text. They trained two models, a text-regression/Lasso model which predicts based on text content, and an ideal point topic model (IPTM) which attempts to isolate the themes/topics of the bill. They found that the Lasso model had 88.1% accuracy and the IPTM model had 87% accuracy, which both outperformed the baseline model. Kraft et al (2016) used a different technique on the same prediction problem and data set. They used an average-of-embeddings approach to train a model for each legislator based on the legislator's historic voting record. They encoded the bill text as a vector of unique words to transform it into training features. Their results improved upon Gerrish and Blei, achieving 90.6% accuracy compared to an 84.5% baseline.

Recent studies have focused on further improving the accuracy of prediction. Kornilova et al (2018) found that inclusion of metadata features relating to the topic and ideology of the bill for predicting legislative votes augmented the approach of Kraft et al (2016), achieving 86.21% accuracy compared to a 68.31% baseline when using this approach in combination with a convolutional neural network. Patil et al (2019) addressed the problem of predicting future votes of legislators without prior voting records. They used news articles and a manually assembled knowledge base as training data. De Marchi et al (2020) considered additional metadata features including legislator constituency information, a score measuring legislator ideology, and the committee from which the bill came. They found that the inclusion of the committee improved performance, but the other factors were of lower relevance.

One potential difficulty identified in the literature is that analyzing political speech in comparison to text content usually used for sentiment analysis like Tweets or reviews, is that while "speakers must in theory address the proposed motions, the speeches can be long, cover diverse subject matters, include multiple targets of subjective language, and often feature irrelevant […] procedural language" (Abercrombie and Batista-Navarro, 2020, p. 5077). Another area that needs to be considered by researchers is that votes are often highly skewed towards Yes votes (Budhwar 2018) which needs to be considered when designing the models. Additionally, Kornilova et al (2018) observed that although there may be large numbers of bills introduced during a legislative session, only a small amount of them may be voted on, and future research should explore bootstrapping techniques to address this problem. There is agreement among the researchers that for a higher degree of accuracy, the data to be used in the model should be made more uniform. Some techniques for this include choosing to focus on bill as a whole and exclude votes on amendments (Gerrish and Blei, 2011; Kraft et al, 2016; de Marchi et al, 2020), eliminating votes other than Yes/No and procedural language (Abercrombie and Batista-Navarro, 2020), or focusing on only one house/session of parliament and removing legislators who often missed votes (Cain, Chua and Gampong, 2012).