Zero-Shot Multiclass Classification of Policy Documents with Large Language Models

Keywords: Large Language Models, Policy Agendas, Text Classification, Natural Language Processing, Public Policy

Extended Abstract

Classifying policy documents into policy issue topics has been a long-time effort in political science and communication disciplines. Human coding is still commonly used in policy agendas research, but because of the large number of documents to be classified and typically large size of each document, that kind of task usually requires many researchers to contribute to coding. Some researchers have tried supervised machine learning methods (Burscher, Vliegenthart De Vreese, 2015; Loftis and Mortensen, 2020, Sebők and Kacsuk, 2020), and they achieved promising accuracy performance. However, training a supervised algorithm still requires substantial human effort to prepare training data. We have large amounts of labelled data from the Comparative Agendas Project (CAP), but given the changing nature of the scope of some policy domains, those algorithms may need significant updates over time. In this work, we test the prediction performance of an alternative strategy. We use the GPT 3.5 model of the OpenAI, which is a pre-trained Large Language Model (LLM), to classify policy hearings and policy bills from the American Congress into CAP's 21 major policy issue topics, which include categories such as Macroeconomics, Transportation, Environment, Defense, Government Operations and Health. While the accuracy performance we obtain cannot beat some supervised machine learning strategies, it gets quite close to some supervised ML results which were obtained by training with a very large set of human coded examples.

We use a zero-shot classification strategy in which we show no labelled examples, but we only use detailed instructions from the CAP's master codebook in the prompt we send to the GPT 3.5 API. The instructions tell the algorithm the specific policy themes associated with each major topic. After the instructions, we show the title of a bill or a short description of a hearing and ask the algorithm to map the title or the description to one of the CAP major topic classes. We used two large human coded policy agendas datasets, which are the Congressional Bills dataset (Adler Wilkerson, 2011) and the Congressional Hearings dataset from the CAP. We randomly drew around ten thousands data points from each of the datasets to test GPT 3.5's multiclass classification skills in the context of policy document classification. The highest accuracy score we achived with the bills sample is around 67 percent, while the highest accuracy we achieved with the hearings dataset is around 64 percent. Class specific performances could be seen in Figure 1 for the bills dataset and in Figure 2 for the hearings dataset. One of the most recent examples of a supervised ML result with the bills dataset was conducted by Loftis and Mortensen (2018). The best accuracy performance they achieved was slightly higher than 70 percent, which they achieved by training a Naive Bayes classifier with more than 200K training examples. The accuracy performance we achieved was almost equivalent of the accuracy performance they achieved with around 50K training examples. Our result is promising given that we used no examples, but only instructions and that we used a pretrained algorithm without fine-tuning it.

Identifying the policy domain of a policy document can be challenging because the high dimensional nature of the issue topic space and small semantic distance between some dimensions. The accuracy performance we obtained may not be satisfactory for most policy agendas research projects yet, but our experiments with several different prompt scenarios indicate that there is some room to improve the accuracy performance through prompt engineering. Research also shows that, Large Language Models perform better when they are shown a few examples of each classes (Brown et al., 2020). Because the large number of classes and prompt token limits, we could not try a scenario with examples yet, but future versions of the GPT model promise to increase token limits. When we have greater token sizes, this performance might be improved further only by prompt engineering.

References

Adler, E. S., and Wilkerson, J. (2011). The Congressional bills project. http://www.congressionalbills.org.

Brown, T.B.; Mann, B; Ryder,; Subbiah, M; Kaplan, J; Dhariwal, P; Neelakantan, A; Shyam, P; Sastry, Askell, A.G. et al. Language models are few-shot learners. In NeurIPS, 2020. https://doi.org/10.48550/arXiv.2005.14165

Burscher, B; Vliegenthart, R De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? The ANNALS of the American Academy of Political and Social Science, 659(1), 122–131.

Hearings. The Policy Agendas Project at the University of Texas at Austin, (2017). www.comparativeagendas.n
Accessed February 28, 2023.
Loftis, M., Mortensen, P. B. (2020). Collaborating with the Machines: A hybrid method for classifying policy documents. Policy Studies Journal, 48(1), 184-206.
https://doi.org/10.1111/psj.12245

Sebők, M., Kacsuk, Z. (2020). The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach. Political Analysis, 29, 236–249. https://doi.org/10.1017/pan.2020.27

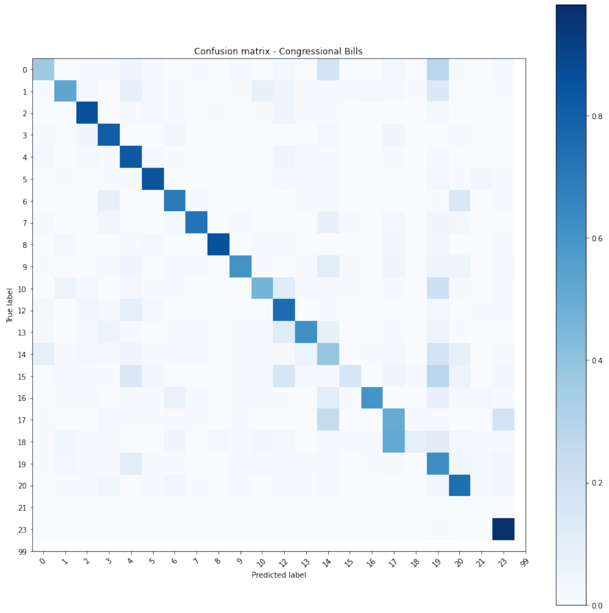

Figure 1: GPT3.5 Prediction Performance on Congressional Bills Data

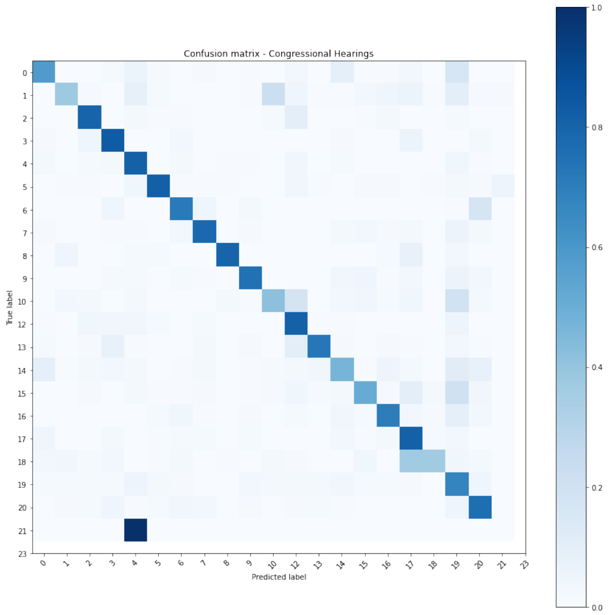

Figure 2: GPT3.5 Prediction Performance on Congressional Hearings Data