Multi-label Model for Legislators' Questioning Classification
2023 Applied Deep Learning Final Project
Group 7

Yu-Hung Sun
Graduate Institute of Public Affairs
National Taiwan University
r10343017@ntu.edu.tw

Peng-Ting Kuo
Department of Political Science
National Taiwan University
r10322029@ntu.edu.tw

Abstract

The concept of text-as-data has grown increasingly vital in the field of political science, yet the process of transforming text to data is still largely done by human. To address this gap, this paper aims to provide a pre-trained model for legislators' questioning classification tasks. We demonstrate a multi-label classification model using Taiwan legislators' transcripts and human-labeled data to achieve this task. Applying the Longformer architecture, our model reaches up to 0.8389 of Hamming score. The result reaches almost 90% of human labeling. We release our code and demo interface.[1] [2]

1 Introduction

The concept of text-as-data has grown increasingly vital in the field of political science. It provides researchers with a robust means to analyze the positions and behaviors of political figures, supported by compelling empirical evidence (Akirav, 2011; Haselmayer et al., 2022). One of the primary approaches involves analyzing the content of texts and assigning labels according to the purpose of the research (Maricut-Akbik, 2021). Nonetheless, this method typically depends on human labeling, which is a process that is expensive and time-consuming.

While numerous studies have delved into the realm of automated labeling, the predominant focus of these investigations has been within the ambit of Western political science scholars, who primarily analyze English-language corpora (Rudkowsky et al., 2018; Haselmayer and Jenny, 2017). In contrast, research dedicated to Chinese language corpora, particularly in the political science domain, remains scant (Shao, 2019). In an effort to bridge this research gap, our study endeavors to develop an automated labeling model. This model harnesses the capabilities of a pre-trained language model, and applies it to a curated dataset comprising interpellations transcripts from Taiwanese legislators, with a specific emphasis on councilors representing Kaohsiung City.

2 Related Work

The idea of treating text as data has a long history in the study of political science and even in the field of social science. The evolution of algorithms, from the earliest bag-of-words models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non negative Matrix Factorization (NMF) (Lee and Seung, 2000), to word2vec models based on word embedding, and finally to language models built on the Transformer architecture, has played a significant role in advancing the application of natural language processing (NLP) in the field of political science.

Normally, political scientists use NLP methods to help them explain theories or phenomena by solving a specific NLP task (Cardie and Wilkerson, 2008). This might include detecting political behaviors of a nation or political figure by using topic modeling, or explaining political behaviors of legislators and voters by using sentiment analysis (Ceron et al., 2015; Dorle and Pise, 2018; Hasel-

[1] Our code is available at: https://github.com/deankuo/ADL2023_Final_Project

[2] Our demo interface is available at: https://huggingface.co/spaces/dean22029/multi-lable_classification

mayer and Jenny, 2017). However, until recently, political scientists still heavily rely on bag-of-word methods, such as topic modeling using LDA, or sentiment analysis using pre-defined sentiment dictionaries (Shao, 2019; Haselmayer et al., 2022). This can be a problem as it primarily relies on calculating the frequency of specific words with small corpus, and it may not precisely capture the context and nuances within the text.

To overcome the limitations of the previous bag-of-words method, researchers began adopting the word2vec model based on word embedding, first introduced by Google's Mikolov and others in 2013 (Mikolov et al., 2013). In contrast to the bag-of-words model that views text as an unordered collection of words, ignoring word order and context, the concept of word embedding involves representing each word as a vector with a fixed dimension. This vector captures the semantic information of the word itself. By utilizing pre-trained word embedding models on large-scale text datasets, researchers can fine-tune existing models with their own training data, enabling the model to achieve higher accuracy with fewer samples. Additionally, if new vocabulary appears in predicting other data, not present in the training data, the word embedding model may already have semantic information for this new vocabulary. Therefore, the model can make predictions using words with similar semantics (Haselmayer and Jenny, 2017; Haselmayer et al., 2022; Rudkowsky et al., 2018)

Despite the notable successes achieved by the word embedding model in natural language processing, it is not without its limitations. One significant challenge arises from its reliance on clear tokenization, akin to the bag-of-words approach. This task becomes particularly arduous when dealing with the Chinese language, which, unlike Indo-European languages such as English, lacks natural spaces within sentences. Consequently, Chinese tokenization necessitates precise word segmentation, a process that is inherently complex due to the absence of clear delimiters between words. Although there are several segmentation tools available, such as jieba and CKIP Tagger, they do not always guarantee accurate tokenization (Li et al., 2020). Inaccurate tokenization can significantly impair the performance of models that rely on bag-of-words or word embedding techniques. This is because improper segmentation can lead to the misinterpretation of the context and semantics of words, thereby undermining the model's ability to accurately process and analyze the text. Additionally, the complexity of Chinese characters, which may represent multiple meanings, further complicates the tokenization process, posing a unique challenge for developing effective and reliable word embedding models for Chinese language corpora.

Vaswani et al. pioneered a revolutionary neural network architecture, termed the Transformer, which marked a significant advancement in the field of natural language processing (Vaswani et al., 2017). This architecture distinguishes itself through its innovative use of a self-attention mechanism, enabling it to attain an enhanced contextual understanding of input data.

The Transformer relying solely on attention mechanisms to draw global dependencies between input and output. This design allows for more nuanced and context-aware interpretations of text, a feature that was somewhat limited in previous models. Building on this foundation, Google developed Bidirectional Encoder Representations from Transformers (BERT), a pre-trained model that leverages the Transformer's architecture to achieve groundbreaking results in a wide range of language processing tasks. BERT's key innovation lies in its bidirectional training, which enables a more profound understanding of language context and nuances by analyzing text sequences in both directions (Devlin et al., 2018).

By fine-tuning BERT, researchers have been able to achieve superior performance with fewer training samples compared to traditional word embedding models. BERT's effectiveness stems from its deep bidirectional nature, allowing it to grasp the subtle contextual meanings of words in a sentence, which is crucial for tasks like sentiment analysis, question answering, and language inference. It also entailed with a plethora of its variant models like

RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019) and SpanBERT (Joshi et al., 2020).

Following the success of BERT, the RoBERTa model introduces several key changes, including extended training duration with larger data batches, removal of the next sentence prediction (NSP) objective, training with longer sequences, and dynamically altering the masking pattern of the training data. (Liu et al., 2019). Notably, RoBERTa demonstrates significant improvements over BERT in various benchmarks and once a state-of-the-art model, the advancements made by RoBERTa demonstrated the potential of transformer-based models in capturing complex patterns and nuances in corpus such as political speeches.

However, pre-trained models such as BERT and RoBERTa are subject to certain limitations, one of which is their token constraint of 512 tokens. This limit restricts the ability of these models to process texts exceeding 512 tokens in length. Unfortunately, texts in political science often encompass long and extensive documents, presenting challenges due to this token length limitation. In our dataset, many contexts extend well beyond the 512-token threshold.

To address this limitation, researchers have turned to alternative models capable of handling longer sequences. Among these is the Longformer, introduced by Beltagy in 2020, which incorporates a blend of global and local attention using dilated sliding window mechanisms (Beltagy et al., 2020). This design enables the Longformer to process lengthy sequences efficiently, reducing the time complexity from $O(n^2)$ to $O(n)$. As a result, the Longformer emerges as an apt model for political science researchers who work with extensive textual data, as it facilitates a comprehensive capture of context and dependencies over broader textual spans [3]

Continuing with the exploration of advanced language models, the Generative Pre-trained Transformer (GPT) developed by OpenAI represents a paradigm shift in natural language processing. Unlike BERT and RoBERTa, GPT adopts an auto-regressive language modeling approach, generating coherent and contextually relevant text based on input prompts. Introduced in 2018, GPT has undergone multiple iterations, with GPT-4 being one of the most sophisticated versions.

GPT's distinctive feature lies in its ability to consider context and generate text in a sequential manner. The model's architecture is designed to predict the next word in a sequence given the context of preceding words. This auto-regressive nature enables GPT to capture long-range dependencies and intricate patterns in textual data. While initially recognized for its language generation capabilities, GPT has proven versatile in various natural language understanding tasks.

One major draw back of GPT-3 and other GPT series are the sizes of the models. With parameters of 175 billion, it is nearly impossible for individuals to fine-tune the model. To address this issue, large language models with smaller parameter but similar strength was developed, most notably were the LLaMA and LLaMA-2 developed by Meta (Touvron et al., 2023). Using the LLaMA series models, researchers are able to fine-tune large language model based on its needs, for example, building a model that is more suitable to the cultural context of a certain region or country, thus we use the Taiwan LLaMA for zero-shot and few-shot prompt engineering Lin and Chen, 2023b.

As the field continues to evolve, LLMs has shown their abilities of evaluation (Lin and Chen, 2023a), and GPT and LLaMA represent avenues for exploration in political science research. Researchers may benefit from considering the strengths and limitations of each model to choose the most suitable approach based on the specific requirements of their political science analysis.

3 Dataset

3.1 Data Source

The foundation of this study lies in a human-annotated dataset comprising oral questions posed

[3] Although there are models that are also able to deal with longer sequence like BigBird (Zaheer et al., 2020) and Transformer-XL (Dai et al., 2019), but we will not discuss these models due to the restriction of pages.

by Kaohsiung City councilors. The source of data comes from the transcript of all 66 Kaohsiung City councilors during their question periods from 2018 to 2022. The coding methodology employed method from Maricut-Akbik's (2021), which is a comprehensive analysis of parliamentary questioning within European parliaments (Maricut-Akbik, 2021). In alignment with this coding framework, each oral question is systematically categorized into one of four main categories: requesting policy information, seeking policy justification and explanation, advocating for a change in policy, or applying condemnation or sanctions. It's important to note that these categories are not mutually exclusive, allowing for a more nuanced understanding of the diverse nature of questions raised by legislators, making this question as a multi-label task. The original human-labeled dataset was completed by three master's students and a PhD student, with overall inter-coder reliability of 0.91, which means that the dataset is highly reliable.

3.2 Coding Rules

Within the established coding rules, the category of "Requesting policy information" encompasses instances where a question seeks objective policy details from the executive branch. This involves inquiries regarding specific budget figures or the solicitation of written reports.

> 「請問工務局長，*108*年度道路刨鋪，每平方公尺的單價為何*?*」

> 「局長，這個後續書面回覆我。」

Moreover, "Requesting policy justification or explanation" is a distinct category that pertains to questions seeking clarification and insights into policy actions. This involves inquiring about the subjective views or opinions of officials within the executive branch.

> 「你認為取消印花稅對高雄市政府財政有多大的影響*?*」

> 「防治流感很重要，為什麼這個打一年就停止*?*」

Additionally, the category "Requesting policy change" addresses instances where questions propose specific actions or make requests to alter existing policies.

> 「可不可以具體要求把危急個案也放進這個救護辦法*?*」

> 「本席建議，用台糖的土地去打造青年住宅」

Lastly, "Condemnation or sanctions" refers to questions that involve criticism or the threat of legal investigations against the executive branch or officials.

> 「我們今天非常嚴厲的譴責，韓國瑜落跑市長」

> 「這是對的嗎*?*對得起高雄市民嗎*?*政風處處長，要不要調查一下*?*」

In summary, each document (in this context, referring to the transcript) will be assigned anywhere from zero to multiple labels, contingent upon the behavior of the encoders. Consequently, it is possible for each document to be associated with two or more labels.

3.3 Data Structure

The dataset we use contains 10050 sets of questions or parts of questions from the councilors. Inside of the dataset, 3373 of texts includes requesting for policy information, 4907 of texts includes requesting for policy justification, 3996 of texts includes requesting for policy change or actions, 874 of texts includes condemnation or sanctions.It is also worth noting that 3100 of texts do not have any label, these texts or questions are either policy overview introduced by the councilors, or simply do not have any political or policy meaning.

As shown in Figure 1, some of the question can be as long as 7000+ words, or it can be as short as 27 words. In addition, figure X shows the correlation of labels. it is worth noting that requesting for policy justification and requesting for policy change or actions has higher correlations. This is

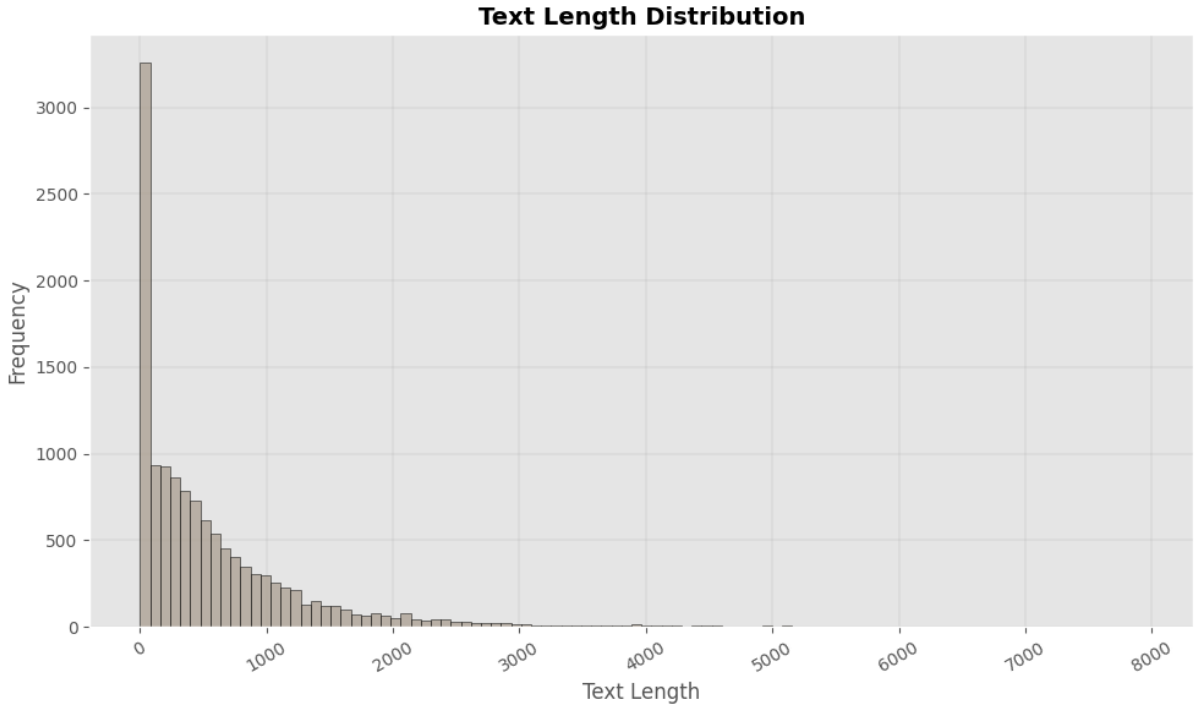

Figure 1: The distribution of text length in the dataset.

Label	Amount
Information	3373
Justification	4907
Change	3996
Sanctions	874

Table 1: Number of questions that contains certain labels.

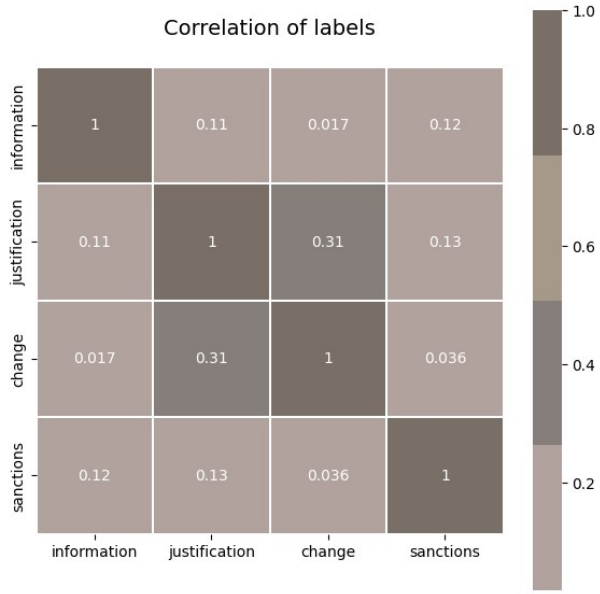

Figure 2: The correlation of different labels in the dataset.

because most of the time when councilors in Kaohsiung City are requesting for policy change or actions, they would also ask for the opinion from the government officials. It is likey a strategy for the councilors to sound more polite by seeking opinion from the bureaucrat first than ask for changes and actions, rather than demanding bureaucrats to change the existing policies.

4 Experiment and Results

4.1 Pre-trained Models and LLM Prompt-Based Learning

Considering that the length of the training data is mostly in the range of 1000 tokens, and the documents are in Traditional Chinese, we refer to the state-of-the-art pre-train models like RoBERTa (Liu et al., 2019) and Longformer (Beltagy et al., 2020). We employed models specifically trained on Chinese data, namely `hfl/chinese-roberta-wwm` (Cui et al., 2020) and `ValkyriaLenneth/longformer_zh` from HuggingFace. The Longformer implementation, optimized for long-text training, adopted RoBERTa's pre-trained checkpoint. It utilized global atten-

tion for special tokens like [CLS] and employed a dilated sliding window for other tokens. The window size was smaller in lower layers, gradually increasing in higher layers. This methodology significantly reduced the traditional self-attention complexity from $O(n^2)$, enhancing special task adaptability (Beltagy et al., 2020).

In addition, to overcome the token limitation and a more accurate classification result, we also employed `yentinglin/Taiwan-LLM-7B-v2.1-chat`, which is a Large Language Model that has been fine-tuned based on LLaMA-2 with corpus data specifically on to the Traditional Chinese language used in Taiwan (Lin and Chen, 2023b). We believe that the result from `Taiwan-LLM` will be more promising than the result from `hfl/chinese-roberta-wwm` and `ValkyriaLenneth/longformer_zh`, as `Taiwan-LLM` was fine-tuned on corpus closer to the data that we used in this research, while `hfl/chinese-roberta-wwm-ext` and `ValkyriaLenneth/longformer_zh` were largely fine-tuned on Simplified Chinese.

In light of the fact that the parameter size of Taiwan-LLAMA reaches 7B, making it infeasible to train on a single-core GPU, we employ the QLoRA (Low Rank Adaptation) method. This approach involves fine-tuning specific parameters of the language model and quantize a pre-trained model to 8-bit or 4-bit (Dettmers et al., 2023). This technique is implemented to enhance the accuracy of multi-label classification tasks for city council member dialogues from the advantage of LLM.

In terms of hardware configuration, we trained on a single V100 in Google Colab with 15GB RAM.

4.2 Result

Given the training data length, our initial trials with inputs of 1024 length yielded a Hamming score of approximately 0.75, lower than anticipated. We attributed this to the granularity of the dataset, as its labeling was based on paragraphs rather than sentences, leading to diminished model learning capability with increased input length due to noise. Moreover, the Longformer model, starting from the RoBERTa checkpoint and optimized with global attention, outperformed RoBERTa-large across various metrics, with nearly an 8% difference in accuracy.

Model	Bsz	Token	LR	GA	Ep
Longformer	4	1024	2e-5	2	8
Longformer	4	512	2e-5	4	10
RoBERTa	32	512	2e-5	4	10
RoBERTa-large	32	512	2e-5	4	10
Taiwan-LLM	4	4096	2e-5	4	2

Table 2: Hyper-parameters of three different models.[4]

To further optimize model capability, we adjusted the input length to 512, continuing with Longformer as the architectural framework. With hyper-parameters set to a learning rate of 2e-5, batch size of 4, gradient accumulation steps of 4, and 10 epochs,, our model achieved a Hamming score of 0.8389. This was approximately 8% lower than the initial manual data annotation accuracy of 90%, indicating near-human classification performance. Additionally, the model demonstrated over 0.89 accuracy in labels like 'information', 'justification', and 'change', with F1 scores above 0.85, although 'sanction' classification was less optimal. The lower recall performance, at 0.75 despite 0.955 accuracy, was primarily due to a lack of training data in this category.

However, we observed the model's susceptibility to punctuation marks in verbatim transcripts of inquiries. Sentences ending with exclamation marks or question marks were more likely to be categorized under 'justification' and 'change,' suggesting learning from punctuation rather than text. Consequently, we removed punctuation from the training data, replacing it with spaces. With the same parameter settings, the hamming score reaches 0.8215, and the accuracy is increased to 0.8795 with 1%. The training curve is shown in Figure 3.

In terms of RoBERTa, with hyper-parameters set to a batch size of 4 and 10 epochs, our model achieved a Hamming score of 0.7355,

[4]Bsz for batch size, LR for learning rate, GA for gradient accumulation step, and Ep for epoch.

Model	Hamming Score	Accuracy	Precision	Recall	F1 Score
Longformer	**83.89**	**72.51**	86.77	**88.65**	**87.70**
Longformer + no punc	82.15	71.88	**87.95**	85.56	86.74
RoBERTa	73.55	60.70	81.28	77.40	79.34
RoBERTa-large	74.28	61.81	70.56	83.97	77.27
Taiwan-LLM + instruction tuning	67.78	56.79	54.66	52.97	51.55
Taiwan-LLM + few-shot	43.64	39.04	49.86	46.22	35.71
Taiwan-LLM + zero-shot	42.29	37.96	50.03	45.97	34.09

Table 3: Metrics of all models we experimented.

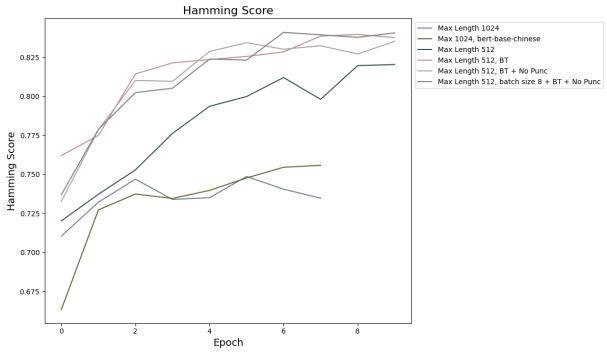

Figure 3: The Hamming score of different settings.

which is lower than the performance from Longformer. Even with training on a larger `hfl/chinese-roberta-wwm-ext`, the model still under perform when compares to Longformer.

Interestingly, when experimenting with Taiwan-LLM, we originally thought that it would have better performance because of its larger parameter and more training data on Taiwanese context, yet it turns out the performance of Taiwan-LLM is lowest among all three models. With our experiment of prompt engineering, the few-shot achieved hamming score of 0.4364 while zero-shot achieved hamming score of 0.4229. Even with instruction tuning of 1000 samples, Taiwan-LLM only achieved hamming score of 0.6778. The poor performance from Taiwan-LLM could be the result of less training sample (only trained on 1000 samples while Longformer and RoBERTa trained on 8000+ samples) and fewer epoch. In addition, this result might indicate that LLM use for language generation are not as capable as pre-trained model like RoBERTa when it comes to classification task, which align with some previous studies.

5 Conclusion

In conclusion, the integration of text-as-data methodologies in political science has become increasingly crucial for understanding the positions and behaviors of political figures, backed by robust empirical evidence. The conventional approach of manually labeling texts is not only resource-intensive but also limited by a focus on Western political contexts, particularly English-language corpora.

Our exploration into automated labeling models for Chinese-language political texts has provided valuable insights into the nuances and challenges of leveraging advanced natural language processing techniques. The transition from manual labeling to automated approaches, particularly with the integration of pre-trained language models such as Longformer, RoBERTa, and Taiwan-LLM, reflects the ongoing evolution in the field of text-as-data methodologies within political science.

Despite having a larger parameter and more understanding on the Taiwanese content, Taiwan-LLM did not perform well in our of classification task when compare to other pre-train model like Longformer and RoBERTa. This result align with some previous studies on classification task in which large language models that were trained for text generation may not have better performance than pre-train model when comes to classification task.

Nonetheless, this study is able to produce a model that is sufficient to predict the type of questions asked by councilors. As the field continues to evolve, the exploration of innovative models, such as those based on pre-trained language archi-

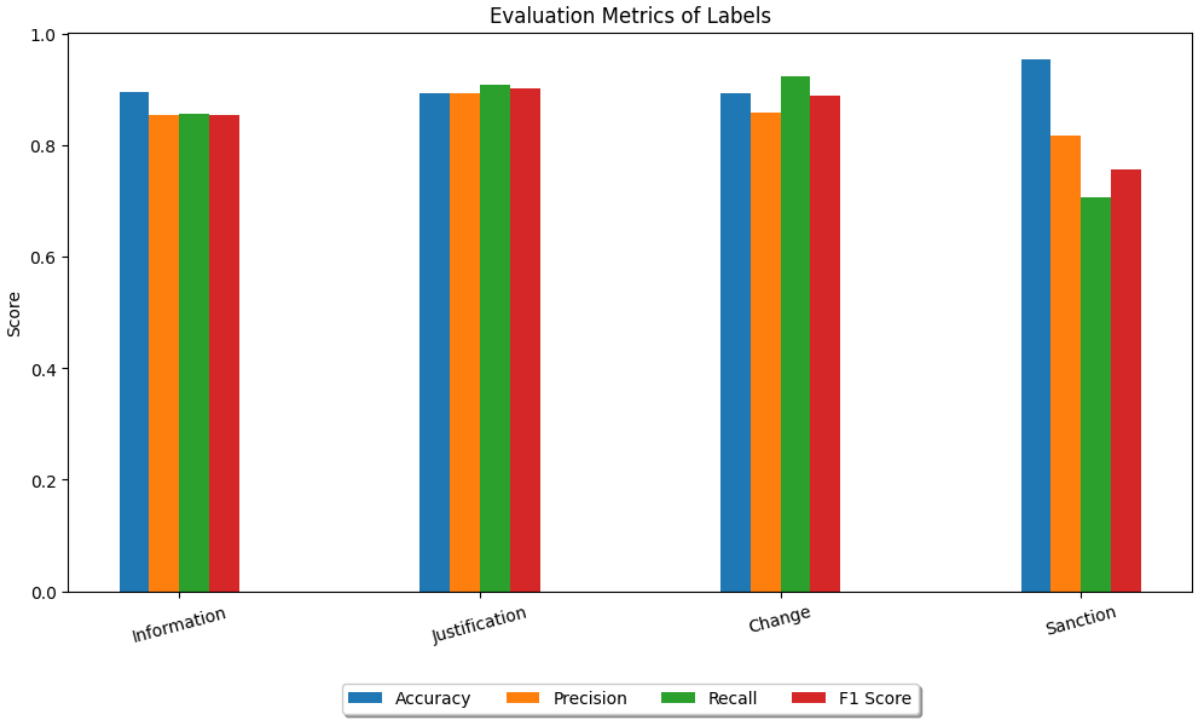

Figure 4: The evaluation of four labels on our model.

tectures, promises to revolutionize text analysis in political science. This study serves as a stepping stone toward a more inclusive and automated approach to labeling political texts, shedding light on the nuances of Chinese-language political discourse and paving the way for future research in diverse linguistic and political contexts.

References

Osnat Akirav. 2011. The use of parliamentary questions in the israeli parliament, 1992–96. *Israel Affairs*, 17(02):259–277.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Claire Cardie and John Wilkerson. 2008. Text annotation for political science research.

Andrea Ceron, Luigi Curini, and Stefano M Iacus. 2015. Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the united states and italy. *Social Science Computer Review*, 33(1):3–20.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Saurabh Dorle and Nitin Pise. 2018. Political sentiment analysis through social media. In *2018 second international conference on computing methodologies and communication (ICCMC)*, pages 869–873. IEEE.

Martin Haselmayer, Sarah C Dingler, and Marcelo Jenny. 2022. How women shape negativity in parliamentary speeches—a sentiment analysis of debates in the austrian parliament. *Parliamentary Affairs*, 75(4):867–886.

Martin Haselmayer and Marcelo Jenny. 2017. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & quantity*, 51:2623–2646.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. 2020. Why attention? analyze bilstm deficiency and its remedies in the case of ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8236–8244.

Yen-Ting Lin and Yun-Nung Chen. 2023a. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Yen-Ting Lin and Yun-Nung Chen. 2023b. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adina Maricut-Akbik. 2021. Q&a in legislative oversight: A framework for analysis. *European Journal of Political Research*, 60(3):539–559.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

H.L Shao. 2019. Machine learning: An application of text mining to xi's grand external propaganda strategy. *Mainland China Studies*, 62(4):133–157.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.