Response Essay for Cross-Domain Topic Classification for Political Texts

1 Background

In the introduction part of the paper ”Cross-Domain Topic Classification for Political Texts”, authors firstly defined the scenario of within-domain and cross-domain supervised learning. The former one is focused on the same types of text while the latter one is more focused on different categories of text when speaking of classifying texts. Authors argued that if one domain lacks labels, using traditional methods such as hand-labeling and unsupervised topic models could raise high computational cost problems. They thought the solution might be that training a cross-domain classifier is helpful on testing on un-labeled texts. I agree with the statement that such transfer learning is efficient as firstly computational cost is low (only need to train one model for all tasks) and secondly when new political texts or parliamentary documents come, the method is still robust when it comes to few shot learning. I didn’t agree with the statement that the authors named where source domain is Manifesto Project Party Platform and target domain New Zealand Parliamentary Speeches as cross-domain datasets. The former one is a dataset containing public political presentations where the latter one is also political presentations so they belong to the same domain: politics. A classic multi-domain benchmark dataset is called Multilingual Domain-specific Translation Corpus (MDTC) Chen & Cardie (2018), where it contains public medical research papers, court judgments, government investment strategies and travel guides etc. They are collected from scientific, legislation, national affairs, and travelling fields, not just limited to political presentations. I guess authors should better explain what domains exactly are, and differentiate them with ”topics”.

Authors argued that they proposed a metric called feature congruence to measure the similarity between the source and target corpus. Furthermore, they extended the methods to real world applications such as New Zealand’s 1993 electoral reform, and gender of New Zealand parliamentarians. In my opinion, these two datasets are appreciated. For the revolution of parliamentary regime, the contents of speeches are marginalized. Topic models often fail to cluster speeches that are unorganized or have no clear topic. Second dataset are dealing with gender bias in presentation to better focus woman’s opinions, rights and necessities. I suspect if the proposed method to test the document similarity by comparing features is novel or not. It is known that documents with similar representation would lead to similar topics or classes (Doc2Vec)Le & Mikolov (2014). And also authors argued that their methods work well compared with some recent works that utilize supervised machine learning to classify cross domain political texts. I found that those benchmarks are different from the one mentioned in this paper. Is that because the dataset is carefully chosen where the topic words are non-ambiguous so that the prediction performance is better?

2 Methods

Authors listed three methods to assign topics to a document, which are dictionary looking, unsupervised topic assignment, and supervised learning prediction. They argued that the methods combine the pros of unsupervised model where it didn’t require more data to be labelled and the pros of lexicon based models which is explainable. I partially agree with authors since Bayesian models are good at explaining underlying assignment ratio, but this situation only limits to some special interpretable models such as state space models, and logistic regression models. I observed that they’ve also trained other machine learning models including random forest, gradient boosting, and a neural net. While xgboost or random forests could provide the feature importance to illustrate the contribution of features to the final prediction, some of which are not interpretable at all for example neural nets. I also didn’t find the code of those models in the code space so I don’t know if they only trained separate logistic regression models with different hyperparameters (L1, L2). Even if they trained them, neural nets are not explainable at all so just giving the weights or probs of logisitic regression is not convincing unless more probs given by other explainbale models could be accessed to. Besides, by giving the weights of each feature as explanation for the topic assignment is out of date, more advanced explainable structure could be used to embed document vectors before applying your individual prediction model such as BERT Kenton & Toutanova (2019) and Roberta Liu et al. (2019) by extracting the attention score. One another problem would be that they didn't consider the data distribution. If treating input source document as training data and targeted document as testing data, testing data could be looked totally different from the training data, where it could be out of distribution. The authors have mentioned this in their conclusion part, which is good. One suggestion might be that they could do some theoretical analysis on their hypothesis class to make sure that they have large mutual information. Assuming some hypothesis classes for the input and targeted source and making some potential guarantees about how two hypothesis classes are far away from each other Shalev-Shwartz & Ben-David (2014).

3 Datasets and Experiment Setting

Authors have defined 44 fine grained classes (topics) and 8 coarse grained classes for each source document. Authors followed the traditional pipeline for classifying texts: pre-processing, tokenization, removing high frequency and low frequency words by N-grams, using term-frequency/inverse-document frequency (TF-IDF) to build input embeddings, and training separate multinomial logistic regression model for 44 and 8 categories respectively. I am surprised why authors didn't use pre-trained Transformer encoder models such as BERT, or pretrained embedding models such as Glove Pennington et al. (2014), and Word2Vec Mikolov et al. (2013) for encoding text. TF-IDF as an encoding method lacks long-term dependencies, as well as embedding similarities between words. I couldn't find any reason in the paper why these modern models are not being used. Future works to authors might be that compare the TF-IDF embedding with BERT and Word2Vec embeddings to see if the prediction model with TF-IDF performs the best.

4 Evaluation and Results

Authors not only considered the common metrics that machine learning uses for example accuracy score, recall or precision or F1 score, but also considered a top-K accuracy, which is that class probabilities are being ranks and see if top K classes contain the true label. This sometimes happens because a document could be classified to two topics with equal probability. With this metrics, a soft margin of prediction is achieved where a more accurate result could be observed from their experimental results. Authors argued that when applying transfer learning on model from within-domain to cross-domain, the model is still robust if observing top-5 accuracy for instance. Authors have not explained such gap but I felt that training data and test data are not served to the same distribution, and also the model could be over-fitting sometimes. I am not sure if authors have printed the training, test loss curve and training, test accuracy curve to observe the overfitting phenomenon, or it still could be underfitting since F1 macro is only 0.417 for 44 topics for manifesto statements.

Regarding the robustness of an explainable model, authors did bootstrapping to eliminate the variations given by input data, where a mean accuracy score was calculated. This is reasonable and has shown that the input data is not out of distribution. For the evaluation on target data, authors still recruited experts to hand code these documents in order to see whether computer and human have a maximal agreement on the prediction results. The problem is that such computation cost is still high though doing a few-shot learning. It could have been resolved by automatically topics assigning with chatGPT, and comparing it with chatGPT's answers. It's far more cheaper and efficient than human labelling. Regarding the interpretation of the model, authors used Ordinary Least Squares Regression (OLR) to assign each word a contribution score. As mentioned in part 2, it could have done simultaneously with Transformer model by extracting attention score instead of proposing a prediction model and using another model to explain this.

Authors have aruged two points about their proposed diagnostic tools. Firstly they argued that the performance of within-domain dataset is along with cross-domain dataset when looking at the table 1 or 4, where they shared a high correlation score on the performance. They argued that such transfer learning was applicable to empirical analysis. I agree on this but need to make sure that other empirical datasets have the same rich features and topics. My comment on the second part which is the feature congruence is shown in part 1 where I showed much concern about its novelty.

5 CONCLUSION

Authors have successfully applied supervised machine learning methods from within-domain to cross-domain without any large performance change. The source and target datasets are all in the domain of political science but shown great importance in the history of parliamentary progress and gender difference in political presentations. They applied an explainable OLS model to tell word's importance according to each topic. They designed a top-K accuracy to give some slack boundaries. They hoped such transfer learning could also be applied to other empirical datasets. Some suggestions to researchers have been given: 1) carefully access the categories of source datasets 2) determine whether source and target datasets are simila 3) hand coding labels to perform evaluation 4) access to alternative advanced models and more data

References

Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1226–1240, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1111. URL https://aclanthology.org/N18-1111.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR, 2014.

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

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.