LLVMs4Protest: Harnessing the Power of Large Language and Vision Models for Deciphering Protests in the News

Yongjun Zhang[1,2]

1 Department of Sociology and Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York, United States
2 Version 1, Updated on Nov 29, 2023

* yongjun.zhang@stonybrook.edu

Abstract

Large language and vision models have transformed how social movements scholars identify protest and extract key protest attributes from multi-modal data such as texts, images, and videos. This article documents how we fine-tuned two large pretrained transformer models, including longformer and swin-transformer v2, to infer potential protests in news articles using textual and imagery data. First, the longformer model was fine-tuned using the Dynamic of Collective Action (DoCA) Corpus. We matched the New York Times articles with the DoCA database to obtain a training dataset for downstream tasks. Second, the swin-transformer v2 models was trained on UCLA-protest imagery data. UCLA-protest project contains labeled imagery data with information such as protest, violence, and sign. Both fine-tuned models will be available via `https://github.com/Joshzyj/llvms4protest`. We release this short technical report for social movement scholars who are interested in using LLVMs to infer protests in textual and imagery data.

Introduction

The rapid advancement of large language and vision models has revolutionized the way social movement scholars analyze protest events using multimodal data such as text, images, and videos. Particularly, OpenAI's GPT-4 has empowered scholars with limited computational skills to intuitively examine textual and visual data. However, there are obstacles in adopting these AI tools. One is the cost of accessing advanced data analytic models from OpenAI, which can be prohibitive for analyzing vast quantities of data. An alternative is to use open-source LLMs like Meta's LLaMa 2, but this requires substantial machine learning knowledge and hardware resources. To address these challenges, we fine-tuned large-scale language and vision models for specific tasks, including protest event identification and key attribute extraction. These models are easily operable on standard computing devices.

Next, we will briefly summarize the major development in social movement literature that uses machine learning techniques to automatically identify protests. Then, we will delve into

some technical details on training data and transfer learning to build our models. Finally, we will discuss some potential issues and limitations.

The Machine Learning Approach to Studying Social Movements

Identifying protest events is one of the primary tasks in the field of social movement studies (Earl, Martin, McCarthy, & Soule, 2004). One of the early endeavours made by social movement scholars was to systematically, manually extract protest information from newspaper articles. For instance, the Dynamics of Collective Action project identified protest events in 1955-1995 from the New York Times (Soule & Earl, 2005). This human labeling strategy was both time consuming and expensive. Thus, many scholars attempted to automate the whole process using machine learning techniques. For instance, Alex Hanna developed the Machine-learning Protest Event Data System (MPEDS) to identify protest events and extract protest attributes from news articles (Hanna, 2017). Other scholars such as Han Zhang and Jennifer Pan adopted Convolutional Neural Networks to classify multimodal social media data including texts and images to build Chinese protest database (H. Zhang & Pan, 2019). In recent years, several groups in computer science have also devoted efforts to detect sociopolitical events on social media data. For instance, the shared tasks for Socio-political and Crisis Events Detection have attracted numerous social and computer scientists to develop deep learning algorithms to identify multilingual protests, fine-grained socio-political events, and Black Lives Matter protests in news (Hürriyetoğlu et al., 2019). Still, limited scholars have fine-tuned and shared large-scale language and vision models to identify protest events in news and social media data. One exception is that Davidson (2023) and Y. Zhang (2023) propose to use generative AI tools, specifically GPT-4V to extract protest information from images and texts.

Instead of relying on the recent advancement of GPT-4 models, in this article, we adopt the transfer learning approach to fine-tune existing pretrained large language and vision models for our downstream tasks in the field of social movement studies, specifically protest identification and attribute extraction. One of the underlying logic is that we attempt to take advantage of existing labeled training datasets both in textual and imagery format, which allows us to achieve the state-of-the-art performance.

Training Data Preparation

Dynamics of Collective Action

The widely used training dataset to classify protest events in newspaper articles is Dynamics of Collective Action (Earl et al., 2004; Hanna, 2017). The DoCA project annotated news articles in the New York Times from 1960 to 1995. It is an event database, consisting of detailed article- and event-level information, such as reporting year, article title, protest size, claims, social movement organizations, locations, timing, etc (see `https://web.stanford.edu/group/collectiveaction/`). Unfortunately, DoCA does not

contain the original articles for these events, thus we collected these textual data from Pro-Quest Database by searching article titles. We managed to obtain 16,700 articles that match the DoCA records. We further excluded articles containing multiple protest events in DoCA, which yielded 11,902 positive news articles in our training dataset.

Given that one of the primary tasks to identify whether a news article is protest related, we further retrieved negative articles from the New York Times Annotated Corpus via the Linguistic Data Consortium. We added 27,000 news articles in several categories that are not likely to contain protest related articles such as entertainment, book reviews, business, classified, finance, sports, real estate, leisure, and obituary.

UCLA-Protest Image Dataset

We rely on the UCLA-Protest Image dataset to fine-tune our large vision model. UCLA protest database contains 40,764 images, including 11,659 protest images, with various annotations of visual attributes such as violence and police (Won, Steinert-Threlkeld, & Joo, 2017). We are also working on our own protest image dataset retrieving from Global Database of Events, Languages, and Tones (GDELT).

Fine-tune Large Language and Vision Models

The Longformer Model

Instead of utilizing the pretrained BERT model, we fine-tune the longformer model, which is specifically used for the long documents (Beltagy, Peters, & Cohan, 2020). We fine-tune the longformer-base-4096 model with our training dataset, since it supports sequences of length up to 4096. Note that longformer-base-4096 is a BERT-like model started from the RoBERTa checkpoint and pretrained for marked language modeling on long documents. For more technical details, please check Beltagy et al. (2020)'s paper.

The Swin-Transformer V2 Model

Instead of training CNN models like VGG or ResNet, we fine-tune the Swin-transformer v2 model (Liu et al., 2021, 2022). Swin Transformer V2 has been shown the state-of-art performance in computer vision tasks. For more technical details, please check Liu et al. (2021, 2022)'s articles. We fine-tuned the swinv2-base-patch4-window8-256 model pretrained by the Microsoft research group.

Note that the accuracy rate for both models is over 94%. We will release more details on evaluation metrics when we finalize our models for scholarly community. The baseline models are available to our scholarly community.

Discussion

AI for Social Movement Studies

In this short essay, we attempt to document some of our ongoing work on fine-tuning large language and vision models to classify protest events and extract protest attributes from news articles and images. In recent years, social activists, movement organizations, and bystanders are increasingly sharing protest activities via social media platforms with multimodal data and more visual information is included in news reports. Because of the recent advancement in generative AIs and deep learning techniques, scholars can take advantage of these large language and vision tools to analyze multimodal data to extract event information to address social science questions. Social movement scholars can benefit from adopting these tools to advance movement data collection and theory testing as well. Our work is one of these attempts by fine-tuning the state-of-the-art Longformer and Swin Transformer models to identify protests.

Limitations and Future Directions

There are several limitations that readers should be aware of. First, fine-tuning transformer models requires scholars to have some basic knowledge of deep learning and programming skills. Fortunately, the swift advancement of generative AIs have lowered the barriers to computational social science (Y. Zhang, 2023). Scholars can use ChatGPT to generate, annotate, and debug codes. Second, the training data used to fine-tune longformer model has time constraints. As most of these positive news articles are from 1960-1995 and it might limit the potential of inferring protests in recent years. This warrants further studies. Third, our work still needs more cross-validation. For instance, a comparison between these fine-tuned models and GPT-4 annotation can offer social movement scholars more nuances in terms of choosing generative AIs or transfer learning with pretrained models. Lastly, our classifiers focus solely on protest identification and some limited number of protest attributes. We are still working on expanding protest attributes. These new features will be released soon once we finalize our fine-tuning processes.

Acknowledgments

References

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Davidson, T. (2023). Start generating: Harnessing generative artificial intelligence for sociological research.

Earl, J., Martin, A., McCarthy, J. D., & Soule, S. A. (2004). The use of newspaper data in the study of collective action. *Annu. Rev. Sociol.*, *30*, 65–80.

Hanna, A. (2017). Mpeds: Automating the generation of protest event data.

Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruşan, F., . . . Akdemir, A. (2019). Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting. In *Experimental ir meets multilinguality, multimodality, and interaction: 10th international conference of the clef association, clef 2019, lugano, switzerland, september 9–12, 2019, proceedings 10* (pp. 425–432).

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., . . . others (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12009–12019).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Soule, S., & Earl, J. (2005). A movement society evaluated: Collective protest in the united states, 1960-1986. *Mobilization: An International Quarterly*, *10*(3), 345–364.

Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th acm international conference on multimedia* (pp. 786–794).

Zhang, H., & Pan, J. (2019). Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, *49*(1), 1–57.

Zhang, Y. (2023). Generative ai has lowered the barriers to computational social sciences. *arXiv preprint arXiv:2311.10833*.