RESEARCH STATEMENT

Mohna Chakraborty

My research interests are in the domain of Data Mining, Natural Language Processing, and Machine Learning, with a focus on extracting and analyzing useful information from large and complex structured/unstructured textual data from multiple data sources such as humans, websites, machine learning models, etc. Specifically, I aim to develop efficient approaches and algorithms to obtain high-quality annotations for diverse data types with minimal or no human effort under a limited budget. *The solutions I developed can be implemented in daily-use systems without expensive hardware, promoting accessibility to everyone.*

Current Research Contributions: Quality Aware Annotations with a Limited Budget.

Data labeling is essential to infer useful information and to train quality machine learning models. When the ground truth is not available, and the budget is limited, the feasible approaches for data labeling are using pre-trained large language models (LLMs) or crowdsourcing. Both these approaches may result in annotations of low quality. Through my research, I proposed methods to improve the quality of annotations obtained using LLMs or crowdsourcing. Below, I discuss my current research work and proposed research plan as a Postdoctoral Fellow.

1.) ***Bias and noise in weakly labeled data:*** In my KDD'22 [2] paper, I address the challenge of analyzing and extracting useful information from the massive influx of daily reviews. Manual extraction is tedious due to the volume and is cost-sensitive, resulting in the need for automatic extraction methods. Prior studies have achieved high performances using deep neural networks, which rely on training data. The motivation for my study stems from the difficulties posed when the budget is limited and the unavailability of labeled training data. To overcome these difficulties, I proposed a double-layer span extraction framework for analyzing the reviews under weak supervision. By leveraging the syntactic and structural patterns in the data, I proposed generic, high-quality rules that use universal dependency parsing to obtain weak labels. Compared with human-annotated labels, the obtained weak labels are biased and noisy. One of the limitations of LLMs is their sensitivity to bias and noise in the training data, which can result in poor quality of the labels. To combat noise and bias in weak labels, I employ canonical correlation analysis as an early stopping criterion to address the issue of noise and a self-training paradigm to enrich the training data iteratively, thus mitigating the issue of bias.

2.) ***Perturbation sensitivity to prompts:*** Recent studies have shown that natural-language prompts can help to leverage the knowledge learned by LLMs for downstream tasks. Prior studies rely on few-shot settings to fine-tune LLMs with manually or automatically generated prompts. The approaches suggested by these prior studies require access to human-annotated samples, and their performance is sensitive to the perturbations of the prompts. My ACL'23 [3] paper addresses these challenges by proposing an automatic prompt generation and ranking approach under a zero-shot setting. I introduce prompt augmentation techniques for automatic prompt generation and rank them using a novel metric based on the intuition that high-quality prompts should be sensitive to the change of certain keywords in the given sentence.

3.) ***Label dependencies to improve data labeling quality:*** Several labeling tasks have implicit or explicit label dependencies between samples in the unlabeled corpus. These label dependencies can help propagate labeling information to dependent samples and enlarge the impact of labels from crowd workers. In my paper published in UAI'23 [5], I take advantage of label dependencies to improve data labeling accuracy for node classification tasks in citation networks. Specifically, I proposed two optimal policies to select the next instance to obtain a crowd worker label by formulating the problem as a Bayesian Markov Decision Process (MDP) and then propagating this information throughout the network employing belief propagation.

Future Research Proposal: Enhancing LLM performance by improving reasoning capabilities.

Despite being treated as black boxes in several studies, understanding the reasoning behind the outputs of LLMs is crucial for deploying them in sensitive applications. LLMs may produce biased and inaccurate outputs due to hallucination, influenced by factors like query misunderstanding, limited domain knowledge, and reasoning capabilities. As a Postdoctoral Fellow, I aim to enhance the quality of data annotations by improving LLMs reasoning capabilities and enabling explainability, interpretability, replicability, and overall robustness in LLMs. I will also continue to apply my research to real-world applications in natural language processing and extend to more interdisciplinary applications since data annotation and analysis are needed in all disciplines. My research will focus on, but is not limited to, the following topics. My research will focus on, but is not limited to, the following topics.

1.) **Explaining and improving the data annotation quality of large language models:** LLMs store the knowledge learned during training as model parameters. One of the techniques to extract the learned knowledge from LLMs is using prompt-based methods. Designing prompts that require LLMs to explain the thought process for their prediction can help explain data annotations by LLMs. I aim to explain the data annotations using LLMs by developing model-agnostic prompting techniques under a zero/few-shot settings. Unlike existing studies such as Chain-of-Thought (COT) [6] and Graph of Thoughts (GoT) [1], my approach will not rely on predefined instructions.
Additionally, LLMs often struggle with reasoning and lack the ability to grasp underlying graphical structures in textual data. Leveraging graph learning techniques, known for their capacity for complex reasoning, scalability, and resource efficiency [4], can complement and enhance LLMs performance. By combining the power of LLMs and graph learning, my goal is to improve the data annotation quality of LLMs by enhancing their reasoning capabilities.
These proposed techniques have the potential to elevate data annotation quality across diverse domains, extending the applicability of LLMs for informed decision-making. As a postdoctoral fellow, I will focus on applying the techniques in the social science domain to learn the behavioral patterns of users from online posts, and the learned patterns can be utilized to help the users in need, thus impacting society as a whole.

2.) **Exploiting textual correlations for cross-lingual data annotations:** The knowledge of LLMs is limited to the language they are trained in, hindering their ability to tackle tasks in other languages. Training them for language-specific tasks is costly due to the need for labeled data, often requiring experts or crowd workers for labeling. To utilize LLMs for cross-lingual labeling tasks, we can take advantage of the intrinsic correlations among languages. During my Postdoctoral fellowship, I will develop a graph-based model to exploit correlations among languages, facilitating the transfer of labeling information from a source language to annotate samples in a target language. This approach will help to efficiently generate high-quality labeled data across diverse languages and tasks.

3.) **Information Extraction from Biomedical Dataset**: Much more information nowadays is carried by massive and unstructured text in the form of news, articles, papers, reports, etc. This format of representing knowledge is rich in information but hard to retrieve and organize and time-consuming for humans to understand. In the biomedical domain, human curation is still the major means to extract information from corpora. This labor-intensive process is slow to keep up with the growing volume of literature and records. People have started to develop information extraction methods to speed up text understanding and build knowledge graphs to mine and organize the information, but the tasks are far from complete. I plan to design practical information extraction tools for massive biomedical corpora to assist in understanding, organizing, and retrieving the knowledge hidden in the text. I also plan to build an interactive information extraction system integrating the power of domain experts and data mining tools.

Broader Impacts. The proposed project will address the limitations of LLMs, including hallucination, help interpret and explain model outputs, and advance research in model explainability, interpretability, and general robustness. The techniques can be applied across domains (social science, biomedical, e-commerce, NLP) to acquire high-quality labeled data economically. These efforts extend LLM applicability to new tasks, addressing current shortcomings and producing research papers, workshops, tutorials, and NSF grant proposals for top conferences.

References:

[1] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. arXiv preprint arXiv:2308.09687 (2023).

[2] Chakraborty, M., Kulkarni, A., and Li, Q. Open-domain aspect-opinion co-mining with doublelayer span extraction. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022), pp. 66–75.

[3] Chakraborty, M., Kulkarni, A., and Li, Q. Zero-shot approach to overcome perturbation sensitivity of prompts. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Toronto, Canada, July 2023), Association for Computational Linguistics, pp. 5698–5711.

[4] Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., and Medya, S. A survey on explainability of graph neural networks, 2023.

[5] Kulkarni, A., Chakraborty, M., Xie, S., and Li, Q. Optimal budget allocation for crowdsourcing labels for graphs. In Uncertainty in Artificial Intelligence (2023), PMLR, pp. 1154–1163.

[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022), 24824–24837.