Research Statement

Elaheh Raisi
elaheh_raisi@brown.edu, elaheh@vt.edu

Learning with Less Dependency to Labeled Data

Most successful machine-learning models, such as deep learning require ground-truth labels for training. The problem, however, is that in many classification and regression tasks having access to the label is costly and time consuming. I am working on designing algorithms to learn when we do not have enough labeled data or information about the label of data.

Multi-Task Learning with Auxiliary Data For problems that often suffer from a lack of labeled data, such as fine-grained visual categorization, I introduce a multi-task learning based framework in which we train two related tasks simultaneously. One is the original task (target), and the other is an auxiliary task (source). The auxiliary task helps to improve generalization capabilities of the deep model. In order to find related tasks to the target data, we leverage a knowledge graph to query for semantically related concepts that are grounded in labeled images. Then, we jointly train our deep network using both target and source tasks. Our framework architecture consists of a shared convolutional neural network followed by two parallel task-specific fully connected classifiers. Shared layers' parameters are updated using both target and source tasks, while each fully connected layer is updated by its corresponding task. Our experiments on two fine-grained visual categorization benchmarks show error rate reduction comparing several baselines. This work published in Visual Learning with Limited Labels (VL3) workshops at CVPR 2020 [1].

Multi-View Co-Training for Cyberbullying Detection In multi-view co-training, we utilize different representations of the data, and jointly optimize all of the learners by maximizing mutual agreements on distinct views to improve the learning performance. I propose a weakly supervised framework for cyberbullying detection in social media. Our weak supervision is in the form of expert-provided key phrases that are highly indicative of bullying. This framework consists of two learning algorithms to improve predictive performance. These learners co-train one another, seeking consensus on whether examples in unlabeled data are cases of cyberbullying or not. Therefore, I refer to the proposed model as the *co-trained ensemble* framework. Each learner has its distinct view of the data; one learner identifies bullying incidents by examining the language content in the message; another learner considers social structure to discover bullying. Intuitively, each learner is using different body of information. And the learning algorithm tries to make them eventually agree whether social interactions are bullying. A fundamental subject for co-trained ensemble framework is choosing diverse learners that look at the problem from different perspectives. Exploiting different learners aligns with the true nature of cyberbullying that can occur in different directions.

Participant-Vocabulary Consistency In my preliminary work, I developed a two-model ensemble refer to as *participant-vocabulary consistency*. This small ensemble method consists of two learners. One learner is based on the tendency of users to bully or be bullied, and the other considers the tendency of language to be used in bullying interactions. Starting with an expert-provided seed set of offensive phrases, the framework is trained so that it can simultaneously discover which users are instigators and victims of bullying, and additional vocabulary that suggests bullying. I applied participant-vocabulary consistency to detect harassment-based bullying on data from three social network services– Twitter, Ask.fm, and Instagram– that rank among the most frequent venues for cyberbullying. I evaluated the proposed method using partially labeled data and post-hoc, crowdsourced annotation of detections by the new algorithms and baselines. The preliminary results were promising; it discovers instances of bullying interactions as well as new bullying language. *I won the best paper award for this work at ASONAM 2017* [2, 3, 6].

Deep Ensemble of Embedding Models I generalized my preliminary work to benefit from nonlinear deep learning methodologies that have advanced significantly in recent years. I applied embedding methods, which represent words, phrases, and nodes in the network, as vectors of real numbers. When word embeddings are trained using deep learning, the vectors created by word embeddings preserve contextual similarities, so we can

extract meaning from text to derive similarities and other relationships between words. I also represent users as a vector of real numbers using node2vec, which is a framework for learning continuous feature representations for nodes in networks. I use word and user vectors as the input to nonlinear language-based and user-based classifiers, respectively. *This deep ensemble framework with preliminary experiments won a best paper award at Learning with Limited Labeled data (LLD) workshops at NIPS 2017, and published in ASONAM 2018 [4, 5].*

Reduced-bias Co-trained Ensemble Model I adjust this framework toward a very important topic in any online automated harassment detection: *fairness* against particular targeted groups including race, gender, religion, and sexual orientations. My goal is to design fair models for our cyberbullying analysis to prevent unintended discrimination against individuals based on sensitive characteristics. To tackle this phenomenon mathematically, I add an unfairness penalty term to the co-trained ensemble framework. The basic idea is to penalize the model when we observe discrimination in the predictions. I explore two unfairness penalty terms: *removal* fairness and *substitutional* fairness. In *removal* fairness, I penalize the model if the score of a message containing sensitive keywords is higher than if those keywords were removed. The other unfairness penalty term is *substitutional* fairness, in which we provide a list of sensitive keywords and appropriate substitutions. For example, for the keyword "Black" in the ethnicity group, substitutions are "Asian", "American", "Middle-eastern", "Native", etc. In a fair model, the score of a message containing sensitive keyword should not change if we replace that sensitive keyword with another one. An ideal, fair language-based detector should treat language describing subpopulations of particular social groups equitably. We quantitatively and qualitatively evaluate the resulting models' fairness on a synthetic benchmark and data from Twitter using post-hoc, crowdsourced annotation. *I won best paper runner up at CSoNET* [7].

Future Research Directions

Semi-Supervised Learning I am enthusiastic to continue my research toward semi-supervised learning in which we have access to small amount of labeled data and huge amount of unlabeled data. The goal is to leverage both labeled and unlabeled data to obtain improvement in generalization performance. I explore several methodologies for these problems. One group of methods is estimating a learner using both labeled and unlabeled data. Those methods include *bootstrapping* or self-training as well as *expectation maximization* and their variants. Another group of methods I am interested in is *scalable graph-based methods* in which we construct a graph on both labeled and unlabeled data and propagate the label based on some distance criterion. Semi-supervised learning can be applied for *few-shot* problems for which we have handful of labeled training set. I would like to approach few-shot learning problems form semi-supervised algorithm's perspective.

Learning with Weakly Labeled Data I would like to continue my research toward learning with weak supervision in which we have access to several weak signals regarding the labels. These weak signals are generally in the form of multiple different weak classifiers that label the data. There are many question in this direction that researchers have been exploring; "how to estimate the true accuracy of these weak classifiers using unlabeled data?," "what is the sufficient condition for the number of learners to estimate the true accuracy considering their error dependencies?," "is a subset of unlabeled data more informative for our purpose?"

Domain Adaptation and Transfer Learning Another areas of my interests, which are connected to weak supervision, and can be related to biased algorithms are *domain adaptation* and *transfer learning*. The major motivation for transfer learning is the lack of labeled training data for many tasks. Transfer learning is a methodology of leveraging the trained model for one task (or domain), then using them for other relevant task (or domain). In this regard, information and knowledge from source task are extracted and applied to target tasks. In some cases, training data is biased. It means there is a difference between the distribution of training data and test data; which results in degradation in model's predictive performance. The aim of domain adaption methodologies is to generalize a trained machine learning model from a source to a target domain. In the *fairness* context, the question will be how to learn a classifier, which is independent of some protected features when we do not have enough labeled data for some particular social groups.

Multi-View Learning– Integration of NLP and Computer Vision The aim of multi-view learning is to exploit various distinct representations of the data to improve the model accuracy. The complementary principle of multi-view learning states that each view contains information that does not exists in the other view, and views could be from multiple sources. I would like to explore *how could we design scalable machines that are able to communicate well with human using multiple resources such as natural language and vision?* The applications in this area I am mostly curious about are visual question answering, visual reasoning, image captioning, searching, and description.

Fairness in Machine Learning Within the past few years, there has been a growing concern about fairness, transparency, and accountability in machine learning models. During my research on cyberbullying detection, I observed some discrimination against particular groups in the framework's prediction. This made me motivated to take some steps in order to make a reduced-biased model for cyberbullying analysis. Many existing machine learning systems are biased against some targeted group. Two main reasons explaining the bias in most of ML models: I) bias in data such as lack of data for some particular groups or existing bias in the data, 2) *algorithmic* bias. After realizing the source of the problem, we could help to decrease the discrimination accordingly. I would like to work on *developing algorithms that do not affect adversely on some particular social groups of people.*

Computational Social Science I am very enthusiastic about studying and analyzing social science phenomena using complex computational approaches. By the help of advanced machine learning techniques, I would like to quantitatively identify and analyze social phenomena and answer social science questions. More specifically, I am interested in *connections* between natural language processing and computational social science. By the contribution of NLP techniques to computational social science, discovering and understanding social behaviors are more efficient and reliable. This is a cross-disciplinary field driven by machine learning, statistics, and social network analysis. My topics of interests include, but not limited to, sentiment analysis and opinion mining, modeling social-network structure, analysis of text in various domains (sociology, psychology, public health, sociolinguistics, etc.,).

[1] E. Raisi and S. H. Bach. Selecting auxiliary data using knowledge graphs for image classification with limited labels. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[2] E. Raisi and B. Huang. Cyberbullying identification using participant-vocabulary consistency. *CoRR*, abs/1606.08084, 2016.

[3] E. Raisi and B. Huang. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 409–416, New York, NY, USA, 2017. Association for Computing Machinery.

[4] E. Raisi and B. Huang. Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 479–486, 2018.

[5] E. Raisi and B. Huang. Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 479–486, 2018.

[6] E. Raisi and B. Huang. Weakly supervised cyberbullying detection with participant-vocabulary consistency. *Social Netw. Analys. Mining*, 8(1):38:1–38:17, 2018.

[7] E. Raisi and B.-H. Huang. Reduced-bias co-trained ensembles for weakly supervised cyberbullying detection. In *CSoNet*, 2019.