Research Statement

Elaheh Raisi (elaheh@vt.edu)

Thesis Research

Early detection of harmful social media behaviors such as cyberbullying is necessary to identify these threatening online behaviors and prevent them from increasing. Finding a technical solution to handle this problem is not easy because of the time-varying nature of language, the difficulty of labeling the data, and complexity of understanding the social structure behind these behaviors. In my research, I aim to improve automated detection of cyberbullying, which is key step toward automated systems for analyzing modern social environments that can adversely influence mental health. I developed Co-trained Ensemble Models for Weakly Supervised Cyberbullying Detection. In this weakly supervised framework, two learners co-train each other to form a consensus whether the social interaction is bullying. I target two objectives: incorporating nonlinear embeddings, and building a fair language-based detector against particular targeted. This research represents important steps toward improving technological capability for automatic cyberbullying detection.

Background and Motivation  Social media is an inevitable part of individuals’ social and business lives in recent years. Along with its uncountable benefits, there are, however, various negative implications. One of the major concerns regarding the adverse consequences of social media is the prevalence of detrimental online behavior such as online harassment, cyberbullying, hate speech, and online trolling. Under the advisement of Bert Huang, my Ph.D. thesis was about general machine learning algorithms for cyberbullying identification. Our goal was to address the computational challenges associated with designing automated, data-driven machine learning approaches for harassment-based cyberbullying detection.

Most machine learning methods for this problem consider supervised and text-based cyberbullying detection, classifying social media posts as ‘bullying’ or ‘non-bullying’. They have crowdsourced workers annotate the data, and then a supervised classifier is applied to classify the posts. There are, however several challenges related to these approaches. Fully annotating data requires human intervention, which is costly and time consuming. And without considering social context, differentiating bullying from less harmful behavior is difficult due to complexities underlying cyberbullying and related behavior. I sought to design an algorithm that encodes such complexities into an efficiently learnable model.

Co-trained Ensemble Framework  I propose a weakly supervised framework for which 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.

To train the model, we construct an optimization problem made up of a regularizer and two sets of loss functions: (A) a co-training loss, which is a loss function between the predictions of all pairs of learners, and (B) a weak-supervision loss. In the proposed co-trained ensemble framework, individual learners train each other to come to an agreement about bullying incident, and then the final learner is an ensemble of all learners. When designing the general framework of ensemble model, I address three tasks: First, using minimal supervision to learn the complex patterns of cyberbullying. Second, incorporating the efficacy of distributed representations of words and nodes such as deep, nonlinear models in the framework. Finally, decreasing the sensitivity of the framework to language describing particular social groups including race, gender, religion, and sexual orientations.

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 the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2017.*

**Nonlinear Embedding** 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. Word2vec is a popular word-embedding model that represents words with low-dimensional vectors based on how often words appear near each other. 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. In addition, this work has been published in 2018 in Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.*

**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. *This work has been submitted to European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD-19).*

**Future Research Directions**

My research interests lie in the broad areas of machine learning, data mining, and computational social science.

**Learning From Weak Supervision** I am interested in continuing my research toward weak supervision when we do not have enough labeled data to train the model. 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 would like to work on designing algorithms to learn when we do not have enough information about the label of data; more specifically, learning from weak supervision. One of the popular research directions for learning with limited human supervision I would like to step in is *interactive learning* in which the model interacts with the expert and seeks guidance.
The key point is to ask for the help of experts more efficiently by having them label the most important data points for the specific task.

**Extension of Co-trained Ensemble** The proposed framework of my Ph.D. thesis could be improved in various directions. One way to expand the framework is to take the temporal aspect of social data into consideration, such that the model can decide bullying occurrences based on the time. The other strategy to develop methods for ensemble co-training is learning in an online, streaming setting, where the new data can be added to the model with minimal computational cost. This way, the input data is presented as a sequence, and models are updated by arrival of new data. Another idea is to explore the usage of different user embedding models beyond node2vec. Our weak supervision was only on language of social media; but it could be strengthened by weak supervision over users. In addition, I would like to extend the scope of the *co-trained ensemble* framework in several applications such as bot detection, abnormal behavior (trolling, vandals) as well as illegal activity (online threats, hate crime) detection in social media.

**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.).

**Natural Language Processing** While working on my Ph.D. research, I leveraged natural language processing methods to detect cyberbullying incidents in social media. I have developed an interest in NLP tasks; more particularly, I would like to explore *how could we design scalable machines that are able to communicate well with human using natural language?* The applications in this area I am mostly curious about are interactive question answering, sentence understanding, semantic matching, sentiment analysis, text summarization, chatbots and conversational models, image captioning and description as well as models combine text and visual data such as visual question answering and visual reasoning.

**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. There are two main reasons explaining the bias in most of ML models: 1) lack of data for some particular groups and II) existing bias in the data. 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.

**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.