Computer Science
Defense

Mining User Generated Content: Addressing Data Scarcity in Filtering Tasks

Payam Karisani
Emory University

Abstract: Filtering tasks have a broad range of applications in mining user-generated data. Examples include public health monitoring, product monitoring, user satisfaction analysis, crisis management, and hate speech detection. This dissertation proposes methods and techniques to overcome one of the primary challenges of these tasks, i.e., the lack of enough training data. It has four main contributions. ¡br¿¡br¿First, it employs semi-supervised learning and proposes a novel method based on self-training and pseudo labeling to use unlabeled data. Our model uses the pretraining-finetuning paradigm in a semi-supervised setting to use unlabeled data for model initialization. It also employs a novel learning rate schedule to exploit noisy pseudo-labels as a means to explore the loss surface. We empirically demonstrate the efficacy of these strategies. ¡br¿¡br¿ Second, it proposes a novel active learning model when additional labels can be obtained for a range of tasks. Specifically, we use a multi-view model to extract two views from documents, and then, we propose a novel acquisition function to aggregate the informativeness and the representativeness metrics for querying additional labels. We analytically argue that our acquisition function incorporates document contexts into the active learning query process. We also treat the highly informal language of users in social media as a factor that manifests itself in the output of learners and causes a high variance. Therefore, we employ a query-by-committee model as a variance reduction technique to combat this undesired effect. Our experiments show that our model significantly outperforms existing models.¡br¿¡br¿Third, it exploits unlabeled documents in a multi-view model . We propose a novel algorithm for one of the most challenging filtering tasks in social media, i.e., the adverse drug reaction monitoring task. Here, we propose a pair of loss functions to pretrain and then finetune the classifier in each view by the pseudo-labels obtained in the other view. Therefore, we effectively transfer the knowledge obtained in one view to the classifier in the other view. We empirically demonstrate that this model is the first known algorithm that outperforms the multi-layer transformer models pretrained on domain specific data. ¡br¿¡br¿ Finally, we observe that although in many cases labeled data is not available, annotated data for semantically similar tasks is available. Motivated by this, we formulate a new problem and propose an algorithm for single-source domain adaptation. We assume that in addition to the source and target data, we can access a set of unlabeled auxiliary domains. We empirically show that existing state-of-the-art models are unable to effectively use this type of data. We then propose a novel algorithm based on the uncertainty in output predictions to decompose the target data into two sets. Then, we show that training using the set of confidently labeled target documents along the auxiliary unlabeled data yields a classifier that is highly effective in the regions close to the classification decision boundaries. The experiments testify that our algorithm outperforms the state-of-the-art in this new problem setting.

Friday, December 3, 2021, 10:00 am
https://emory.zoom.us/j/7390068295

Computer Science
Emory University