World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:04, 2022

Self-Supervised Learning for Hate-Speech Identification

Authors : Shrabani Ghosh

Abstract : Automatic offensive language detection in social media has become a stirring task in today's NLP. Manual Offensive language detection is tedious and laborious work where automatic methods based on machine learning are only alternatives. Previous works have done sentiment analysis over social media in different ways such as supervised, semi-supervised, and unsupervised manner. Domain adaptation in a semi-supervised way has also been explored in NLP, where the source domain and the target domain are different. In domain adaptation, the source domain usually has a large amount of labeled data, while only a limited amount of labeled data is available in the target domain. Pretrained transformers like BERT, RoBERTa models are fine-tuned to perform text classification in an unsupervised manner to perform further pre-train masked language modeling (MLM) tasks. In previous work, hate speech detection has been explored in Gab.ai, which is a free speech platform described as a platform of extremist in varying degrees in online social media. In domain adaptation process, Twitter data is used as the source domain, and Gab data is used as the target domain. The performance of domain adaptation also depends on the cross-domain similarity. Different distance measure methods such as L2 distance, cosine distance, Maximum Mean Discrepancy (MMD), Fisher Linear Discriminant (FLD), and CORAL have been used to estimate domain similarity. Certainly, in-domain distances are small, and between-domain distances are expected to be large. The previous work finding shows that pretrain masked language model (MLM) fine-tuned with a mixture of posts of source and target domain gives higher accuracy. However, in-domain performance of the hate classifier on Twitter data accuracy is 71.78%, and out-of-domain performance of the hate classifier on Gab data goes down to 56.53%. Recently self-supervised learning got a lot of attention as it is more applicable when labeled data are scarce. Few works have already been explored to apply self-supervised learning on NLP tasks such as sentiment classification. Self-supervised language representation model ALBERTA focuses on modeling inter-sentence coherence and helps downstream tasks with multi-sentence inputs. Self-supervised attention learning approach shows better performance as it exploits extracted context word in the training process. In this work, a self-supervised attention mechanism has been proposed to detect hate speech on Gab.ai. This framework initially classifies the Gab dataset in an attention-based self-supervised manner. On the next step, a semi-supervised classifier trained on the combination of labeled data from the first step and unlabeled data. The performance of the proposed framework will be compared with the results described earlier and also with optimized outcomes obtained from different optimization techniques.

Keywords : attention learning, language model, offensive language detection, self-supervised learning
Conference Title : ICMLA 2022 : International Conference on Machine Learning and Applications
Conference Location : Boston, United States
Conference Dates : April 21-22, 2022