Query-based triggers play a crucial role in modern search systems, e.g., in deciding when to display direct answers on result pages. We address a common scenario in designing such triggers for real-world settings where positives are rare and search providers possess only a small seed set of positive examples to learn query classification models. We choose the critical domain of self-harm intent detection to demonstrate how such small seed sets can be expanded to create meaningful training data with a sizable fraction of positive examples. Our results show that with our method, substantially more positive queries can be found compared to plain random sampling. Additionally, we explored the effectiveness of traditional active learning approaches on classification performance and found that maximum uncertainty performs the best among several other techniques that we considered.

This work was performed at Microsoft Research under the mentor-ship of Paul N. Bennett and Ryen W. White.