ALFRED: A NOVEL QUERY SYNTHESIS ACTIVE LEARNING METHOD FOR TEXT CLASSIFICATION

Ihan Dilnath Lelwala

A dissertation submitted in partial fulfilment of the requirement for
Bachelor of Engineering (Honours) degree in Software Engineering

Department of Computing

Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2021

ABSTRACT

What if a Machine Learning (ML) model could have training data generated based on its classification uncertainty and query a human-in-the-loop for their label? In doing so, would we able to construct a better classifier while reducing data acquisition costs for practical, real-world Text Classification applications? This is the premise of Active Learning based on the Membership Query Synthesis scenario, which is coined as Query Synthesis Active Learning in this research. It is an emerging area of research in Active Learning where generative methods and augmentation methods have been proposed for generating new textual training data (Query Synthesis). However, there are several gaps that exist in previous work such as support for Deep Learning architectures in NLP, lack of interpretability in estimation of model uncertainty and query synthesis, and the need to train and finetune a separate model for query synthesis. Therefore, this research aims to design, develop, and test a novel Query Synthesis Active Learning method to address these gaps. To demonstrate the real world, practical application of the Query Synthesis Active Learning method, an interactive data labelling GUI prototype is developed. In addition, another contribution is the source code is designed and developed to be reusable as a Python package that NLP researchers and practitioners can use out-of-the-box for Query Synthesis Active Learning.

Keywords: Active Learning, Text Classification, Interactive Data Labelling