Declaration

We , Samyak, Priyam and Atharv B.tech students of **CSE (Evening) of 7th Semester** of the Maharaja Surajmal Institute Of Technology, New Delhi declare that the Minor project Report entitled **"Text Classification using Active Learning"** is an original work and data provided in the study is authentic to the best of our knowledge. This report has not been submitted to any other institute for the award of any other degree.

We hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.

No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism. We hold a copy of this assignment that we can produce if the original is lost or damaged.

Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and given their details in the references.

Samyak Jain - 41896392717

Priyam Gupta - 42296302717

Atharv Mittal - 41796302717

Computer Science & Engineering (Eve)

7th Semester

Place - Maharaja Surajmal Institute of Technology

December - 2020

Acknowledgement

In the present world of Competition there is a race of existence in which those who are having will to come forward succeed. Project is like a bridge between theoretical and practical work . With this Willing we joined this particular project.

This project has been a great learning experience for us and We would like to express our sincere gratitude to all the people to guide us throughout the project and without the valuable guidance and suggestions this project would not be completed successfully.We would like to extend my sincere thanks to all of them.

We also take this opportunity to express a deep sense of gratitude to our mentor **Dr.Adeel Hashmi** (Head of Department , Department of Computer Science & Engineering) & **Ms. Neeti Sangwan** (Proctor, Department of Computer Science & Engineering 7th Semester Evening Shift) for their cordial support,valuable information and guidance which helped us in completing the task through various stages.

We would like to thank all the respondents whom we interacted during my project for their help and suggestions.

Last, but not the least, An honorable mention goes to our family and friends for their understanding , support and suggestions is completing this project

Table of Contents

List of Figures

Abstract

Natural language processing (NLP) and neural networks (NNs) have both undergone significant changes in recent years. For active learning (AL) purposes, NNs are, however, less commonly used – despite their current popularity. By using the superior text classification performance of NNs for AL, we can either increase a model's performance using the same amount of data or reduce the data and therefore the required annotation efforts while keeping the same performance. We review AL for text classification using deep neural networks (DNNs) and elaborate on two main causes which used to hinder the adoption: (a) the inability of NNs to provide reliable uncertainty estimates, on which the most commonly used query strategies rely, and (b) the challenge of training DNNs on small data. To investigate the former, we construct a taxonomy of query strategies, which distinguishes between data-based, model-based, and prediction-based instance selection, and investigate the prevalence of these classes in recent research. Moreover, we review recent NN-based advances in NLP like word embeddings or language models in the context of (D)NNs, survey the current state-of-the-art at the intersection of AL, text classification, and DNNs and relate recent advances in NLP to AL. Finally, we analyze recent work in AL for text classification, connect the respective query strategies to the taxonomy, and outline commonalities and shortcomings. As a result, we highlight gaps in current research and present open research questions.

Active Learning (AL) aims to reduce the amount of data annotated by the human expert. It is an iterative cyclic process between an oracle (usually the human annotator) and an active learner. In contrast to passive learning, in which the data is simply fed to the algorithm, the active learner chooses which samples are to be labeled next. The labeling itself, however, is done by a human expert, the so-called human in the loop. Having received new labels, the active learner trains a new model and the process starts from the beginning. Using the term active learner, we refer to the composition of a model, a query strategy, and a stopping criterion. In this work the model is w.l.o.g. a text classification model, the query strategy decides which instances should be labeled next, and the stopping criterion defines when to stop the AL loop. According to Settles there are three main scenarios for AL:

(1) **Pool-based**, in which the learner has access to the closed set of unlabeled instances, called the pool

(2) **Stream-based**, where the learner receives one instance at a time and has the options to keep it, or discard

(3) **Membership query synthesis**, in which the learner creates new artificial instances to be labeled.

If the pool-based scenario operates not on a single instance, but on a batch of instances, this is called batch-mode AL . Throughout this work we assume a pool-based batch-mode scenario because in a text classification setting the dataset is usually a closed set, and the batch-wise operation reduces the number of retraining operations, which cause waiting periods for the user.

The underlying idea of AL is that few representative instances can be used as surrogate for the full dataset. Not only does a smaller subset of the data reduce the computational costs, but also it has been shown that AL can even increase the quality of the resulting model compared to learning on the full dataset . As a consequence, AL has been used in many NLP tasks, e.g. text classification named entity recognition or machine translation and is still an active area of research.