Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Medieninformatik

The Title of the Thesis

Bachelor's Thesis

Yannick Brenning
Born Aug. 27, 2002 in Bamberg

Matriculation Number 3732848

1. Referee: Christopher Schröder
2. Referee: Christian Kahmann

Submission date: February 31, 2022

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, February 31, 2022

. . . .
Yannick Brenning

Abstract

This is the LaTeX template for Bachelor and Master theses at Webis. This template contains several hints and conventions on how to structure a thesis, how to cite the work of others, and how to display your results besides plain text.

Contents

Chapter 1

Introduction

Text is one of the most widespread and important sources of information, but extracting data and tangible knowledge from it can be a difficult and expensive task. With the advent of the digital age, enormous amounts of unstructured texts are available with more being generated by the day. Due to this increasingly large amount of textual data, processing information at a larger scale becomes infeasible and thus demands the use of computer-driven approaches.

The classification of text, meaning the assignment of a category or class to a document or piece of text is one of the most common and useful ways to gain information from a piece of text. As the amount of available text content continues to grow, text classification tasks become an increasingly important area of research within the field of natural language processing.

Thanks to machine learning and data science, we have been able to develop many methods of extracting information from text, and as a result, perform text classification at a larger scale. This possibility for automated organization of data can enhance insights and decision-making across industries such as healthcare, finance, and social sciences, among many others. Active Learning (AL) is a subfield of machine learning in which the learning algorithm is able to perform queries on an information source in order to reduce the total amount of annotated data. This method can offer significant advantages in improving model performances and especially in reducing labeling costs. Though there is no universally good strategy, AL has been proven to be useful in many cases where annotating data is expensive or the total amount of data is very large (?).

Oftentimes, there is a large amount of unlabelled available data for an AL model to learn from. In this case, selecting the most effective data points to be labelled by the information source from this large pool becomes a crucial, but difficult challenge to overcome.

One attempt at improving the effectiveness of AL in this regard is the Core-Set approach (?) This method uses core-set selection to counter the issue of AL ineffectiveness on convolutional neural networks. The proposed approach selects a set of points the pool such that a model learning over this subset can be consistent when provided the remaining data points. The method was shown to have improved results when compared to other approaches in the field of computer vision (?, ?), which encompasses tasks that focus on enabling computers to interpret and understand visual information from the world. This field involves a variety of different methods including image classification, object detection, and semantic segmentation, all of which have significant importance in scientific research, classification tasks, and pattern recognition.

However, Core-Set has been shown to have mixed results in cases of multiclass text classification using BERT (?) and binary text classification using DNNs (?). Particularly in the first case, the experiments show that Core-Set performs poorly even when compared to the random sampling strategy. In addition, the approach has even been shown to be less effective in computer vision tasks in cases with higher numbers of classes as well as higher-dimensional data points (?). The theoretical analysis shown in ? briefly mentions this within the context of higher class amounts, however it does not attempt to provide a potential solution to the problem.

This thesis aims to explore the possibility of improving the Core-Set approach for text classification tasks. By first explaining Core-Set's functionality and the theoretical reasons for why it tends to underperform in certain classification tasks, I aim to then demonstrate the performance difference in comparison to various baseline approaches on large datasets of text content in order to verify this claim. Furthermore, this thesis looks to improve on the Core-Set approach within the context of text classification tasks and demonstrate this improvement as a part of its experiment.

In the following, Chapter 2 explains the background and related work on the topics of text classification (Section 2.1), active learning in general (Section 2.2), and the Core-Set approach to AL more specifically (Section 2.3). In Chapter 3, I will explain my approach to improving the performance of Core-Set for text classification using dimensionality reduction. In Chapters 4 and 5, I will present my experiment as well as discuss its results. Finally, Chapter 6 will conclude the thesis and provide insights on potential future developments of the method.

Chapter 2

Background/Related Work

2.1 Text Classification

Text classification is one of the most fundamental and important tasks in the field of Natural Language Processing (NLP). As a result, developing efficient automatic text classification methods has become an important research topic.

One of the most common applications of text classification is determining whether the opinion associated with a certain document has a positive or negative sentiment, also known as sentiment analysis. This has a wide range of uses, including the possibility for businesses to better gauge customer opinions on products and services (**?**) in order to adapt accordingly. This application is a binary classification task, meaning the classifier has two classes with which each document can be labelled (positive or negative). Similarly, one might apply this binary classification task to the problem of spam filtering in e-mails, text messages and more.

Beyond that, many applications of text classification require multiple classes, such as news and content categorization. In this case, text classification algorithms can organize documents into specific topics or themes (e.g. Sports, Business, Politics, . . .) (**?**). Other applications include information retrieval, recommender systems, and document summarization (**?**).

Generally, text classification methods can be divided into the following phases: data collection and preprocessing, feature extraction, classifier selection, and model evaluation (**?**, **?**, **?**).

2.2 Active Learning

2.3 Core-Set

2.4 Dimensionality Reduction

As mentioned earlier, one of the major challenges of the Core-Set approach (among others) is handling data points with a higher dimensionality (**?**). Broadly speaking, this is a phenomenon coined by Richard E. Bellman known as the *curse of dimensionality* (**?**). As a result, many algorithms have been developed to transform data from a high-dimensional space into a low-dimensional space (**?**). This task directly poses another challenge: managing to reduce the dimensionality of the data while still being able to retain the highest possible amount of information.

Chapter 3

Approach

Chapter 4

Experiment

4.1 Data

This experiment was conducted across three datasets commonly used in the field of AL. These datasets are of three different types: sentiment analysis (S), questions (Q), and news (N). For binary classification, I used **Movie Review**, a sentiment analysis dataset (**?**) containing 5,331 positive and 5,331 negative samples. For multi-class text classification, **AG's News** (**?**), comprised of 120,000 training samples and 7,600 test samples, and **TREC** (**?**), a question dataset containing 5,500 training samples and 500 test samples.

The test set was only provided in the case of AG's News and TREC, in the case of Movie Review I employed a split of the 10,662 samples myself.

4.2 Experiment Setup

4.3 Experiment Results

Dataset	**Model**	**Query Strategy**			
		RS	BT	CS	Unknown
AGN	BERT	0.78 ± 0.171	0 ± 0	**0** ± **0**	0 ± 0
	SetFit	0.862 ± 0.044	**0** ± **0**	0 ± 0	0 ± 0
MR	BERT	0.747 ± 0.089	0.743 ± 0.091	**0.715** ± **0.097**	0 ± 0
	SetFit	0.853 ± 0.015	**0** ± **0**	0.856 ± 0.018	0 ± 0
TREC	BERT	0.668 ± 0.251	0 ± 0	**0.592** ± **0.256**	0 ± 0
	SetFit	0.907 ± 0.076	**0** ± **0**	0 ± 0	0 ± 0

Table 4.1: Final accuracy per dataset, model, and query strategy. We report the mean and standard deviation over five runs. The best result per dataset is printed in bold.

Chapter 5

Discussion

Chapter 6

Conclusion