Natural Language Processing Project Phase 2 Report

Diyar Hamedi
Department of Computer Engineering
Iran University of Science and Technology
diyar_hamedi@comp.iust.ac.ir

Summer 2023

Abstract

This report presents the methods employed in the second phase of developing a text topic classification model. This includes analyzing the dataset gathered in the previous phase, extracting different features and designing and training an effective architecture. We will also explain the insights gained from the analysis and showcase the results of the experiments made with different features and architectures. The code for reproducing this file, the results of the experiments and training the models is available at this repository: https://github.com/DiyarH/nlp1402project

1 Introduction

After gathering a dataset for text topic classification from the HowTo100M dataset[1], cleaning up the data and extracting different metrics and statistics from the data, we continue with the process of developing a model suitable for such task. In order to find the best architecture for the model and the features given to it as input, we perform several analysis and experiments. First, we make use of the Word2Vec algorithm to generate a brand new embedding for all of the dataset words and examine patterns and biases which appear in the embedding. Next, we finetune a language model on each of the categorized datasets, and compare the generated sentences of the said model to the actual data. In the following steps, we extract various features from the data and train simple architectures on each of those features separately to gain insight on their usefulness. Using the knowledge acquired, we then develop multiple architectures which take a combination of said features as input, and compare their training performance as well. Finally, we explore the usefulness of Large Language Models (LLMs) for research in Natural Language Processing (NLP), particularly for artificial data generation and zero-shot tasks.

[1] https://www.di.ens.fr/willow/research/howto100m/