# 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.

## 2 Experiments

In this section, we give a brief explanation for each of the steps involved in the experiments made on the dataset and different NLP methodologies.

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

## 2.1 Word2Vec

Using the classic Word2Vec algorithm, we generate an original matrix of word embeddings for each of the categorized datasets from scratch. We also train a word embedding using Word2Vec on the entire dataset, which is used in the following experiments.

## 2.2 Language Model

In this section, we select an auto-regressive Language Model (LM) and finetune it on each of the categorized datasets separately. For this experiment we choose GPT2, due to its lightweightness and its availability on HuggingFace. This allows us to leverage the HuggingFace Trainer API, which trivializes the training process. Here, we present some of the samples generated by the finetuned GPT2 for a few of the categories after finetuning for 1 epoch:

- Home and Garden: la, bien, o. The way. They say to be aware, what they really do was take care of this place, they do, very important things at this junkyard so everybody here you might like what I've had before I've left you that there are a large few boxes...
- Cars & Other Vehicles: ia davids orok d1 ikke ww wakw oezyz yaw. What it is with your guys in here is a small motor bike repair team at home who has an 80-year, ronda c/c hodge r4 motor to repair cars or dicierals are you up against ia d1 of your...
- Personal Care and Style: my eyes, or your forehead has, it been infected with the plague on it. The blood I got that there a person has a infection that looks pretty similar but this doesn't exist, as I think about doing so, the rest and they didn't show, maybe this isn't bad though it doesn't, do really take anything and if people really can do better than this...
- Travel: for the word, you did one hundred ten of, twelve thousand eighty eight three ten of thirteen, twelve. twelve thousand number to be used to put up that word for. 12 words because you, it might be different from one thousand...
- Holidays and Traditions: camel of your country's ; I like ********//; And if not, give me to please, yes that was fun that evening when our first lady I POLICE, please get that * the a, * ********.********** for that one? What's. It for? here at hallu is an image and we also like seeing some amazing examples to begin with and keep the ballad here...
- Youth: ick is to give the people something nice it feels nice ixi'm to give her it as usual a thumbs up after some tea at first incense I'll never have enough for another night's entertainment of today is you really need tea tonight tonight? Why yes sir...
- Hobbies and Crafts: with I know and how good it is, oh god thank God uploading. And to make sure I stay away and watch these, and you still to. The morning is a very kind, kind„„ yours know in person, right about. All a long show from when they's, gonna. All this is, about this evening...
- ips. They're a fun for us to listen as we try each new chord. My main thing and some more fun bits of all. third, and four. When they do music for their next musical band:

  So when it hits our fourth birthday dance party I will be my dance to another dance for you're first year to make sure there won't get my song...

It is apperant that while the models each use a few of the words specific to the category they are finetuned on, the samples generated are very different from normal subtitle texts. One possible way to improve the quality of the generated samples is to simply finetune the models for more epochs, as one epoch of finefuning does not seem to be enough.

## 2.3 Feature Engineering

The purpose of this experiment is to compare different features extracted from each sample and then decide on the features selected as input for training the future models. Each of the following features are extracted from each sample of the dataset and then used in isolation as input to the appropriate architecture:

- Sentence length: The length of the sample in words is used as the only feature. A basic Multi-Layer Perceptron (MLP) model is trained using this feature.
- Word lengths: The lengths of each word in the sample is used as the feature.
- Words: The indices of each word in the sample is used as the feature.
- Word bigrams: This indices of each word in the sample and its previous word are concatenated as the feature.
- Word2Vec: The Word2Vec matrix trained in the first section is used to get the vector corresponding to each word in the sample as the feature.
- Word2Vec bigrams: The Word2Vec vectors for each word in the sample and its previous word are concatenated as the feature.
- BERT: Using the embedding weights of the BERT model (specifically the 'bert-base-uncased' model from HuggingFace), the vectors corresponding to each word in the sample are used as the feature.

The architecture of the model used in this experiment for the features other than the sentence length is made of a multilayer, bidirectional LSTM followed by a linear layer to map the LSTM outputs to class predictions.

## 2.4 Classification using OpenAI

In the last section of this report, we propose a simple zero-shot prompt to be given to LLMs such as ChatGPT, GPT4 and LLaMA for text topic classification. The prompt is as follows:

> I need you to analyse a video subtitle and tell me which category out of the following it belongs to:
>
> Finance and Business
>
> Arts and Entertainment
>
> Personal Care and Style
>
> Travel
>
> Youth
>
> Work World
>
> Education and Communications
>
> Pets and Animals
>
> Hobbies and Crafts
>
> Home and Garden
>
> Computers and Electronics
>
> Food and Entertaining
>
> Health
>
> Philosophy and Religion
>
> Sports and Fitness
>
> Cars  Other Vehicles
>
> Holidays and Traditions
>
> Family Life
>
> Relationships
>
> I do not need any explanation for your analysis, so only tell me the name of the category, nothing else. This is the text you must classify:
>
> Insert the text you want to classify here