Thematic classification of questions

Matthew Harding [Student ID - MHARD94]

May 2024

1 Abstract

2 Introduction

The aim of this project was to benchmark the performance of several techniques for classifying the theme of a question. The hypothesis being that the wording of a questions, without the addition of other context or metadata, is sufficient to classify a question to a general theme.

3 Training Data

The Yahoo! Answers topic classification dataset [1] is a set of set of questions grouped by topic taken from the Yahoo! Answers corpus as of 10/25/2007. The data was obtained through the Yahoo! Research Alliance Webscope program [2].

The dataset contains a training set 1,400,000 questions. Luckily, the dataset is evenly split across the 10 categories so there is no issue of class imbalance. 60,000 additional questions are reserved for testing. Again the test questions are evenly split across the categories.

Topic Code	Topic Label	Training Count	Test Count
0	Society & Culture	140000	6000
1	Science & Mathematics	140000	6000
2	Health	140000	6000
3	Education & Reference	140000	6000
4	Computers & Internet	140000	6000
5	Sports	140000	6000
6	Business & Finance	140000	6000
7	Entertainment & Music	140000	6000
8	Family & Relationships	140000	6000
9	Polotics & Government	140000	6000

The **Datasets** library from Huggingface was used to pull the data. For each question, there is a a topic label, question title, question content and best answer. For this project, we consider the topic to combination of question title and question content to be the classification input and the topic label to be the classifcation output. The best answer value was disregarded for this task.

References

[1] Yahoo! answers huggingface dataset. `https://huggingface.co/datasets/yahoo_answers_topics`.

[2] Yahoo! research alliance webscope. `https://webscope.sandbox.yahoo.com/`.