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	Politics & 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.

4 Data Preprocessing

First, the question title and content were combined into a single string. All the text was then converted to lowercase and punctuation was removed.

Using the Python Natual Language Toolkit, stop words were removed. Stop words, such as "the" or "in", are removed to reduce noise and highlight keywords that carry essentual meaning. If the task was one where context is required such as machine translation, stop words would be retained.

Lastly, the text is lemmatized, again using the Python Natual Language Toolkit. Lemmatization is used to reduce a word to it's base form. Unlike stemming which merely removes common suffixes from the end of word tokens, lemmatization considers the context and meaning of a word ensuring the output word is an existing word that can be found in the dictionary. This is important for the zero-shot model where invalid English will cause classification errors.

5 Bag Of Words

The baseline approach, based on

References

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

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