Michael Lee

Linguistics 583

5/12/2023

News Category Classification

The dataset I used is the News Category Dataset uploaded by Rishab Misra from Kaggle. It is a JSON file containing a list of 42 different news articles with their headlines, short descriptions, and assigned categories. For the purpose of the project, I only needed the short description and category columns. However, I combined the headlines and short descriptions into a text column to utilize more of the available data while still focusing on the more essential features. Having many categories can make the model more complex and prone to overfitting, especially if some categories have very few samples. To address these challenges and to streamline my analysis, I decided to limit the scope of the study to a select few categories and created a dictionary of new mappings for categories that were similar. This reduction not only simplified the models but also significantly improved their computational efficiencies. I resized the categories to match the size of the smallest one to ensure that no single category dominates the model training in an attempt to have fair and accurate classifications. I tried choosing ones that were not too similar to each other to help the models differentiate between them more effectively.

I established a random baseline using DummyClassifier and tokenized the text using spacy. I then experimented with different models, fine-tuning their hyperparameters for optimal performance. I also created a confusion matrix for each model to better visualize the distribution of news categories. Below is a summary table showcasing the different models I explored and their corresponding performance scores:

Model	Original scores	Fine-tuned scores
SGDClassifier	Accuracy: 0.67 F1 macro: 0.669 F1 weighted: 0.668	Accuracy: 0.707 F1 macro: 0.706 F1 weighted: 0.706
SGDClassifier + fasttext vectorizer	Accuracy: 0.663 F1 macro: 0.658 F1 weighted: 0.657	Accuracy: 0.709 F1 macro: 0.709 F1 weighted: 0.708
Multinomial Naive Bayes	Accuracy: 0.689 F1 macro: 0.693 F1 weighted: 0.692	Accuracy: 0.718 F1 macro: 0.719 F1 weighted: 0.719
Logistic regression	Accuracy: 0.690 F1 macro: 0.690 F1 weighted: 0.689	Accuracy: 0.698 F1 macro: 0.699 F1 weighted: 0.699
SVM (Support Vector Machine)	Accuracy: 0.580 F1 macro: 0.583 F1 weighted: 0.582	Accuracy: 0.657 F1 macro: 0.658 F1 weighted: 0.657

The performance of the various models closely resembled each other, with the majority achieving scores approximately in the 0.70 range. However, the Multinomial Naive Bayes model slightly surpassed the others with a score of around 0.72.

I decided to incorporate topic modeling using LDAModel with the goal of evaluating its ability to accurately classify words into appropriate topic clusters. I was curious to see if the generated labels would resemble the predefined categories, which would also provide an overview of the prevalent themes across the articles. After fine-tuning the model, I created labels and saved them to a CSV file along with the key words for each topic for further reference and analysis. While the labels derived from the most common words in each topic did not directly align with the predefined categories, the relevant words associated with each topic offered a sense of the underlying subject matter, allowing for informed assumptions regarding the category to which it likely belonged.

Overall, the performance of the models was satisfactory. However, there is potential for further enhancement by conducting more extensive tuning or exploring alternative models. By

investing additional efforts in pre-processing the text data, fine-tuning the models, or experimenting with different algorithms, I can strive to optimize their performance and achieve even better results.

Dataset link: https://www.kaggle.com/datasets/rmisra/news-category-dataset

Source: rishabhmisra.github.io/publications

1. Misra, Rishabh. "News Category Dataset." arXiv preprint arXiv:2209.11429 (2022).

2. Misra, Rishabh and Jigyasa Grover. "Sculpting Data for ML: The first act of Machine Learning." ISBN 9798585463570 (2021).