Classifying Categories on the News Category Dataset

Ana Carolina Souza

I. Dataset

With the age of the internet, it's become easier to classify news articles into different categories with the help of machine learning or even manually, so long as one can keep up with the volume. The issue arises with historic journalism, where the sheer volume of articles makes it impossible to classify them all. This is where machine learning comes in.

The case study that inspired this project is a research paper [1] that uses a private dataset to explain and demystify the process of classifying news articles. Since the dataset used in the paper is private, 2 similar datasets from Kaggle will be used.The first one [5] contains around 210k news articles from the HuffPost containing 42 categories. The second dataset [2] is a balanced version of the first one containing only 46k news articles with 10 categories from the original.

II. Classification Pipeline

A new column was created by merging the title of the article with the content. It was then pre-processed by removing punctuation, numbers, and stopwords, tokenized [7] and lemmatized [6]. The original dataset was checked for class imbalance, with the articles heavily skwed towards "Politics" containing 16% of the dataset. To counteract this, the balanced accuracy score was used as the evaluation metric. This issue doesn't occur in the second dataset as that one was doctored to be balanced and contains less classes.

The bag-of-words model [3] was chosen because some categories have unique words, like "Mother" for "Family". However, this can be exploited to confuse the classifier by including words that are not relevant. For example, since the datasets contain overlapping topics, the classifier could stumble on words that are common to both classes like "government" or "election" in both "Politics" and "US News" articles.

To conclude the pipeline, the Logistic Regression model [4] was chosen along with the "Saga" solver due to its ability to handle large datasets.

III. Evaluation

Upon running the pipeline for both datasets with a 80-20 train-test split, the classifier was able to achieve a balanced test accuracy score of 0.44 for the first dataset and 0.80 for the second dataset. The most important words for classification for the top 2 categories for the first dataset were "cheater", "custody" and "stepparent" for the divorce category and "wardrobe", "beauty" and "stylish" for the fashion category. The bottom 3 worse categories were impact, women and taste. The difference between the top and bottom categories due to conflicting words that are common to multiple categories and unique enough words essential to the category. The classifier was able to achieve a higher accuracy for the second dataset due to the balanced classes and the lack of overlapping categories like "Taste" and "Food".

IV. Dataset Size

To test the dataset size, 5 fractions of the dataset were used: 0.1, 0.3, 0.5, 0.7, and 0.9. The dataset was split 80-20, and the error rate was calculated for each fraction. Afterwards, the learning curve was plotted for both datasets. In both cases, the error rate doesn't stabilize, which means that the model would benefit from a larger dataset. However, as seen with the balanced dataset with the lower errors, just increasing the size doesn't guarantee a better model. In order to increase the accuracy, the model would need a dataset with a more balanced distribution of classes. It's feasible to increase the size, but with a careful selection of the articles as the business case doesn't control the amount of news articles released per category.

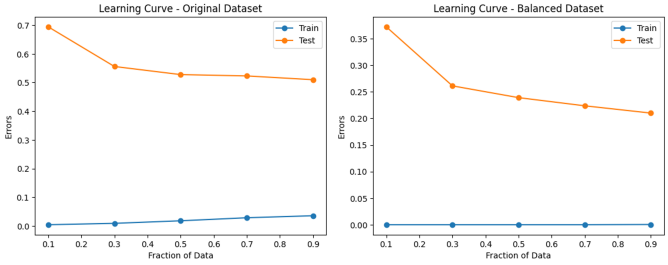

Fig. 1. Figure 1: Scatter plot for Dataset 1 on the left and Dataset 2 on the Right showing their Learning Curves

V. Topic analysis

For topic analysis, a pipeline consisting of a vectorizer and NMF model was used to extract topics from the dataset. The original dataset found 20 topics with a good distribution across. The topics were then used to classify the documents into their respective categories. The second dataset found 10 topics. When retraining the model, the classification error was found to be higher for both datasets. Some topics were found to have a higher error rate than others, like "Travel" and "Food and Drink", which makes sense given the complexity and diversity of the topics in the dataset.

REFERENCES

[1] Aysenur Bilgin, Laura Hollink, Jacco van Ossenbruggen, Erik Tjong Kim Sang, Kim Smeenk, Frank Harbers, and Marcel Broersma. Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history. *ARXiv*, 2018.

[2] Codify. News category dataset, 2021. Accessed: 2024-10-03.

[3] Scikit learn Developers. Countvectorizer documentation, 2024. Accessed: 2024-10-03.

[4] Scikit learn Developers. Logistic regression documentation, 2024. Accessed: 2024-10-03.

[5] Rishabh Misra. News category dataset. *arXiv preprint arXiv:2209.11429*, 2022.

[6] NLTK Team. Wordnetlemmatizer documentation, 2024. Accessed: 2024-10-03.

[7] NLTK Team. Wordnettokenizer documentation, 2024. Accessed: 2024-10-03.