Report for HW6 for course Uvod v odkrivanje znanj iz podatkov

Mark Znidar

May 31, 2024

Introduction

In this assigment, various approaches were explored to achieve the best prediction accuracy. The primary aim was to maximize prediction accuracy rather than explainability.

The following were implemented:

- Conducted exploratory data analysis to understand the dataset.
- Implemented and evaluated linear regression models.
- Considered the task as a document retrieval problem by calculating embeddings, identifying the closest articles in the embedding space, and predicting new article comments based on proximity of new articles.
- Explored XGBoost, which consistently outperformed linear regression. A grid search was implemented to optimize its parameters.
- I read articles [1, 2] to get better ideas on how to construct features.
- Investigated neural networks, ultimately leading to the final model.

Extracted Features

- **artexclaim**: Presence of exclamation marks in the title.
- **artquestion**: Presence of question marks in the title.
- **article_word_count**: Number of words in the article.
- **category_new**: Subcategory extracted from URL.
- **figure_count**: Number of figures in the article.
- **daily_sin**, **daily_cos**: Utilizing Unix time.
- **weekly_sin**, **weekly_cos**: Utilizing Unix time.
- **monthly_sin**, **monthly_cos**: Utilizing Unix time.
- **quarterly_sin**, **quarterly_cos**: Utilizing Unix time.
- **half_yearly_sin**, **half_yearly_cos**: Utilizing Unix time.
- **ne_loc_cnt**, **ne_org_cnt**, **ne_per_cnt**: Named entities counts in title and lead.
- **paragraph_count**: Number of paragraphs in the article.
- **title_word_count**: Number of words in the title.
- **category_counts_in_one_hour**: For each article, how many other articles in the same topic published in the interval one hour before or after.
- **emb_0** to **emb_767**: Embedding features representing the whole article content including keywords.

Final Model

Neural networks were constructed for each topic in the dataset, demonstrating an improvement in predictive accuracy compared to a single model applied to the entire dataset. Embeddings for complete articles, including keywords, were generated using the sloberta model from EMBEDDIA. Named entity recognition was performed using SpaCy, and these entities were incorporated into feature construction.

Before inputting the data into the model, a standardization process was conducted using the StandardScaler from the scikit-learn library. Additionally, it's worth noting that no model was constructed for the topic "Kolumne," as it was not present in the test dataset.

Optuna was employed to optimize the hyperparameters and topology of each neural network. The hyperparameter search space was defined based on methodologies proposed in related literature [2]. The hyperparameters explored included the number of layers (ranging from 2 to 4), the number of neurons per layer (ranging from 20 to 140), dropout rate, batch size, and weight decay. The Adam optimizer was used with a learning rate of 0.01, and the loss function applied was L1 loss. To ensure reproducibility, the random seed was set to 1.

Each neural network for each topic has its own topology specifically designed for the task at hand.

Additional Considerations

Sentiment analysis using sloberta-sentinews-sentence was explored, but the model outputs were of suboptimal quality. Named Entity Recognition to extract entities such as names, company names, and places was also implemented, but it negatively impacted prediction accuracy. Sparse NN architecture was explored, but performed worse than weight decay.

References

[1] Alaa Elsakran, Abdullah Al Amin, and Ayman Alzaatreh. Machine Learning Approach to Predict Facebook Comment Volume. In *2019 International Conference on Digitization (ICD)*, pages 64-67, 2019. 10.1109/ICD47981.2019.9105863.

[2] Lihong He, Chen Shen, Arjun Mukherjee, Slobodan Vucetic, and Eduard Dragut. Cannot predict comment volume of a news article before (a few) users read it. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 173-184, 2021.