RESEARCH ARTICLE CLASSIFICATION

Presented by: Siang Hostel

INITIAL INVESTIGATION

Our initial investigation on train data the following insights

- Multilabel Classification
- classification of Latex abstracts from arXiv
- where the data exhibited significant bias.
- Abstracts averaged around 150 words.

DATA PREPROCESSING

We employed 2 different approaches of text preprocessing.

one type of preprocessing focusses to take only keywords and
remove the terms and stopwords which are domain independent and special charaters.

as the size of abstract is small
and covers all the important info for the research article we try to conserve
the context by lightly preprocessing the special characters

Initial Experimentation

- In our initial experimentation phase, we explored several traditional and ensemble classification models, including

- We focused on feature engineering using TF-IDF matrices, concentrating on keyword presence and frequency without considering the contextual information. This approach allowed us to assess the effectiveness of different models in handling the sparse and biassed nature of the data.

MULTINOMIAL NAIVE BAYES

COMPLEMENT NAIVE BAYES

LOGISTIC REGRESSION

XGBOOST

SUPPORT VECTOR CLASSIFIER

ENSEMBLE MODELS

OTHER EXPERIMENTS

- Sequential models
- Text agumentation
- Using Different classifiers:
    - One vs Rest
    - MultiOutput

Model Exploration

To address the complexities of the task and leverage contextual information, we experimented with transformer-based models, specifically BERT and XLNET. These models are renowned for their ability to capture semantic nuances and dependencies within text data, making them suitable for our multilabel classification task.

EXPERIMENTS AND EXPLORATION

- Exploring Family of bert models and it’s architecture
- Removing unintended bias
- Heirarchical classfication with bert

OPTIMAL SOLUTION

Vocab and base model- SCIBERT
Loss function : BCEWithLogitsLoss
Optimizer: AdamW
Cross Validation: KFold
Training time - 8 hrs with cross validation