DOCUMENT CLASSIFICATION WITH NEURAL NETWORKS

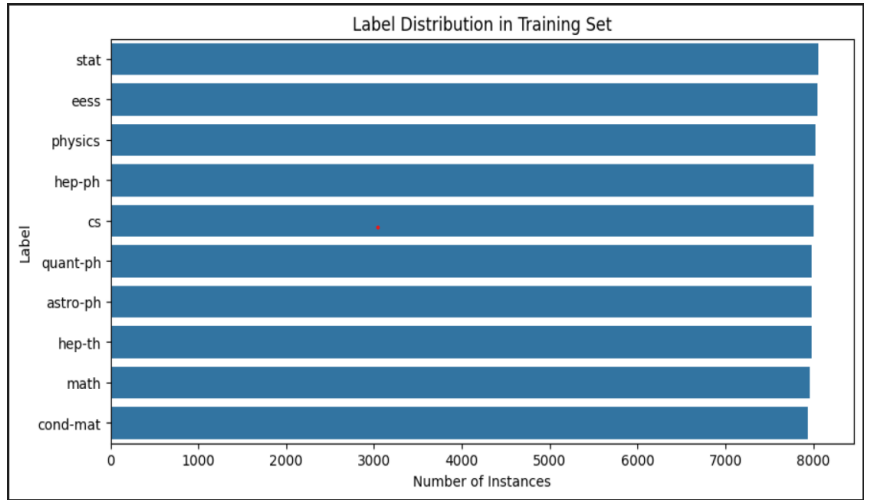
DAT550 Project • Ali Hussain Khan & Said Vagap, University of Stavanger

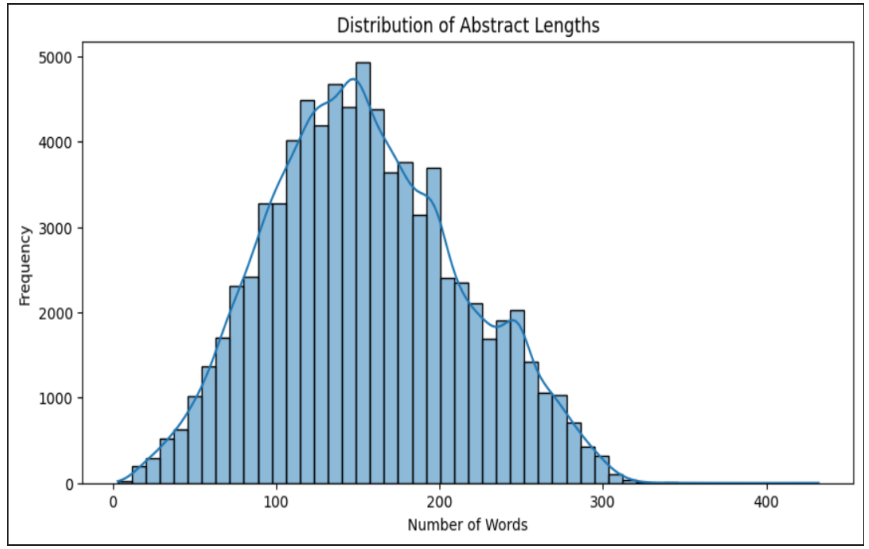
OBJECTIVES

- Classify scientific paper abstracts into their research field
- Compare three modelling strategies:
  - Bag-of-Words + MLP
  - Pre-trained GloVe embeddings with pooling
  - Recurrent Neural Networks (Simple RNN, LSTM, GRU, BiRNN)
- Identify the most efficient, accurate, and generalizable approach

DATASET & PREPROCESSING

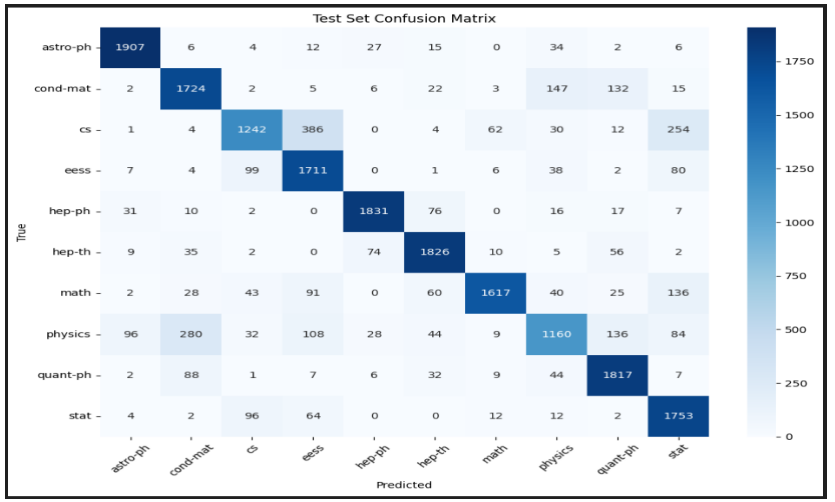
- Dataset: Subset of Arxiv-10 (~80 000 abstracts, 10 balanced fields)
- Abstract length: Most between 100–200 words (avg ~150)
- Preprocessing: Clean text (lowercase, strip whitespace), tokenize, encode labels, stratified 80/20 split



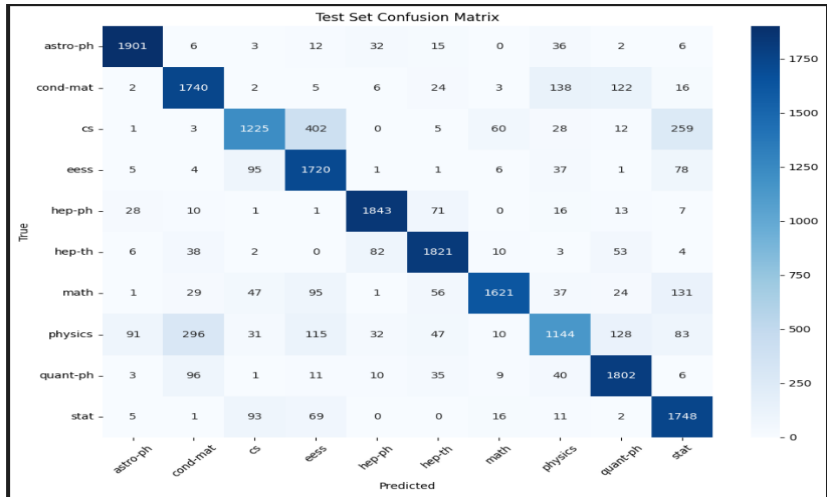


BAG-OF-WORDS + MLP

- Two methods: CountVectorizer and TF-IDF.
- CountVectorizer: Simpler, faster, better accuracy.
- TF-IDF: Slightly better on validation, slower.



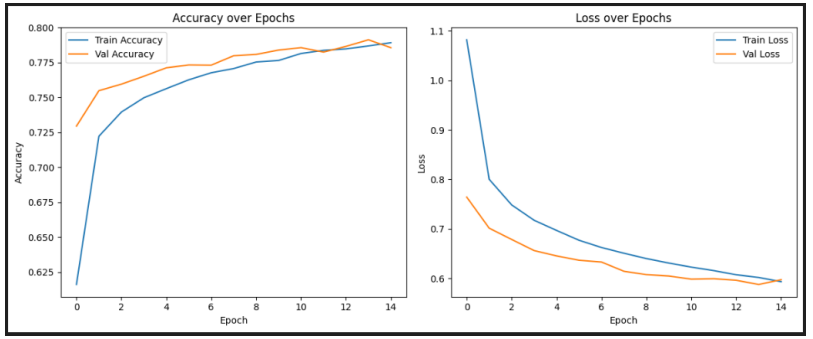
Confusion Matrix of Countvectorizer fed neural network



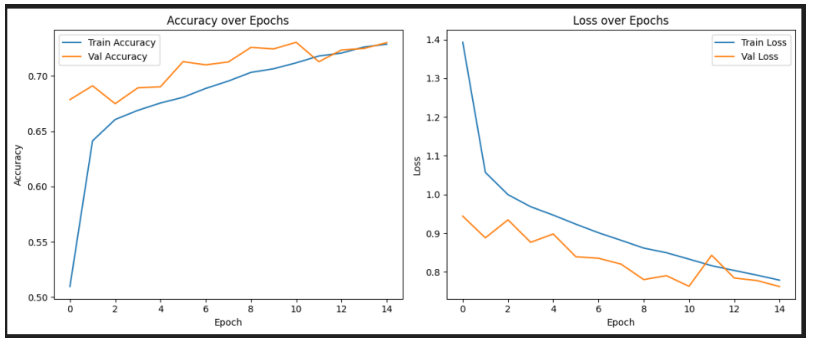
Confusion Matrix of TF-IDF fed neural network

PRETRAINED EMBEDDINGS (GLOVE)

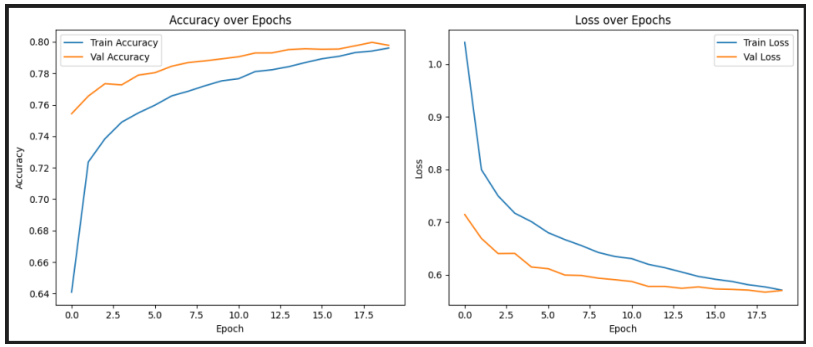
- GloVe 300d used with mean, max, sum pooling.
- Best result: Mean pooling (~78.5%).
- Max pooling underperformed.



Mean pooling through epochs



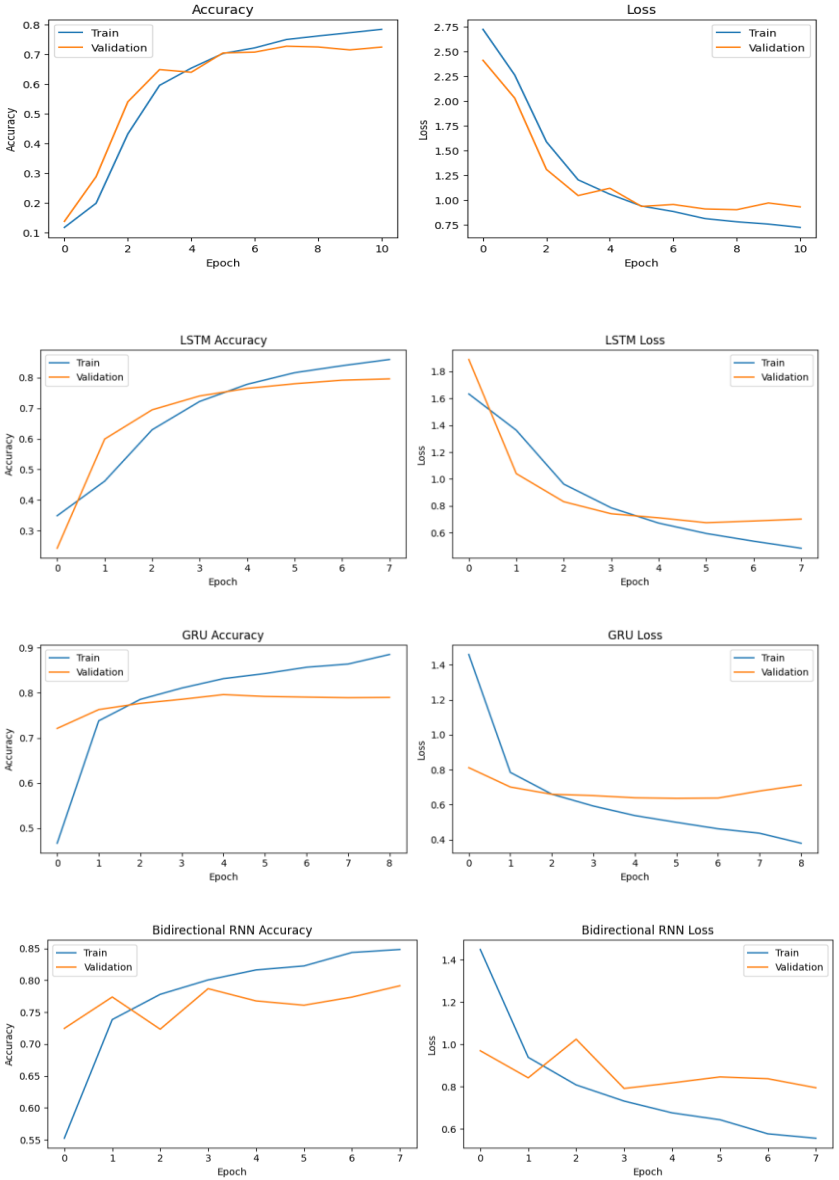
Max pooling through epochs



Sum pooling through epochs

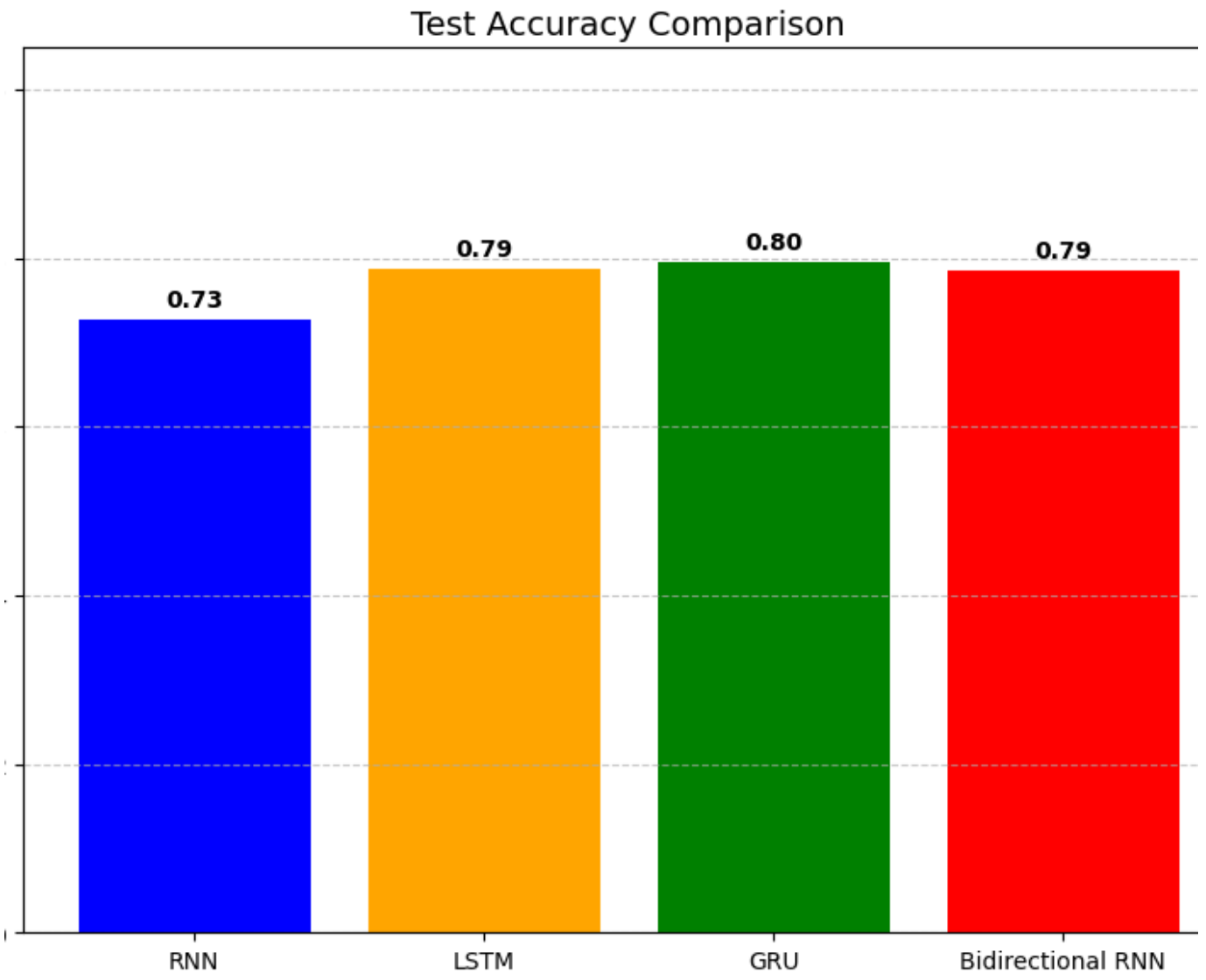
RECURRENT NEURAL NETWORKS (RNNS)

- Models: Simple RNN, LSTM, GRU, BiRNN.
- GRU performed best (~79.6%).
- BiRNN added complexity with marginal gains.



RESULTS – COMPARISON

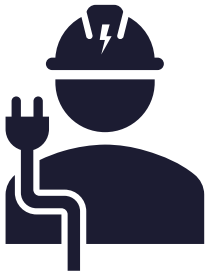
- BoW + MLP: 83%
- GloVe mean pooling: 79%
- Simple RNN: 72.7%
- LSTM: 78.7%
- GRU: 79.6%
- BiRNN: 78.5%



DISCUSSION & FUTURE WORK

Key takeaways: Simple bag-of-words often outperforms more complex models on short texts

Embeddings and RNNs add context but cost more compute for marginal gains



Future directions: Fine-tune transformers (BERT/SciBERT), train domain-specific embeddings, combine feature types