Comparing Intrinsic and Extrinsic Evaluation of Sensitivity Classification

Group 15

Karbeutz Gerhard	**Gjorshoska Ivana**	**Aleksic Lucija**	**Palmrich Alexander**	**Romanov Leonid**
12014883	12432889	12202117	0825978	12318347

Availability of data

OHSUMED

- publicly available
- need to merge 3 sources:
  - HuggingFace OHSUMED
  - MESH Categories from NIH
  - original repo
- no single source for all data

AVOCADO

- restricted access
- expensive: $1500
- chose not to use

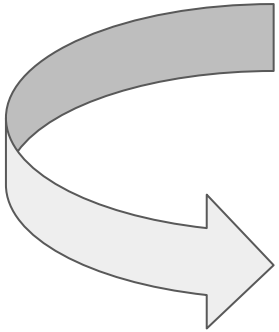
OHSUMED Preprocessing

Sensitivity Labels:

- merged 3 data sets
- regex filtering to label **sensitive**.
- poorly documented in paper!

Relevance Labels:

- Judgements from original dataset repository.
- Relevance status determined based on consensus or fallback logic:
  - `relevant` if majority judgement is positive.
  - `not relevant` otherwise.



Data Preparation:

- Text Cleaning
- TF-IDF Transformation
- Data Split:
  - **Test Set**: both sensitivity and relevance labels (~16,460).
  - **Train/Validation Set**: others split 85:15.

Outputs:

- TF-IDF features saved for modeling.
- Train/validation/test saved for classification.

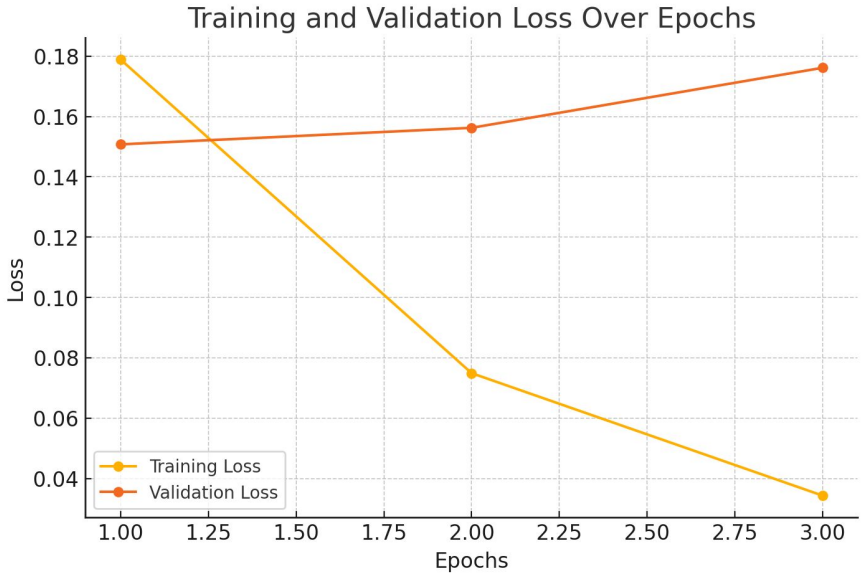
Logistic Regression

- Predict sensitivity: yes/no
- Expectedly poor performance
- Computationally cheaper than distilBERT

	Precision	**Recall**	**F-1**	**F-2**	**Accuracy**
Our model	0.83	0.40	0.54	0.44	0.92
Paper	0.77	0.73	0.75	0.74	0.94

DistilBERT

- Only 10% of data used
- Trained with Google Collab using T4
- Training took approx. 1h
- Results good in spite of smaller dataset



	Precision	Recall	F-1	Accuracy
Our model	0.88	0.69	0.77	0.95
Paper	0.77	0.73	0.75	0.94

Difficulties & Successes

- restricted access data
- poor pre-processing doc
- RAM issues
  - Logistic Regression: solved
  - DistilBERT: open

TO-DO

- combine Logistic Regression & DistilBERT
- extrinsic evaluation
- test statistics