Natural Language Processing

Project: Detection of Toxicity in Online Comments

Group 8

Roland Ramp / Dario Giovannini / Adrian Ovari

PROJECT GOAL

Construct a robust model using GBERT and added data annotations for effective toxicity assessment in German online comments

Milestone 1 steps

- Dataset Overview:
  - 522 article titles in the dataset, some with minimal comments.
- Column Selection:
  - Primarily used “Comment” only, annotations only for NER (later)
- Data wrangling:
  - Focused on Comments and Label columns for NLP toxicity classification.
  - Edited some special cases (one comment was binary encoded)
- Handling Duplicates:
  - Removed duplicate entries, opting for the majority label.
- Export Process:
  - Utilized stanza-pipeline for tokenization, lemma generation, POS tagging, and sentiment assignment.
  - Preserved labels and titles during export with a helper function

  - Labels: 1655 Toxic and 2818 Non-Toxic comments (37% toxic)

Milestone 2 steps

Data Preparation:
- Lemmatized comments in the dataset.
- Employed a bag-of-words approach for baseline models.
- Pruned dataset by removing singleton words, empty comments.

Train-Test-Validation Split:
- Split data into 60:20:20 ratio for train, test, and validation sets.

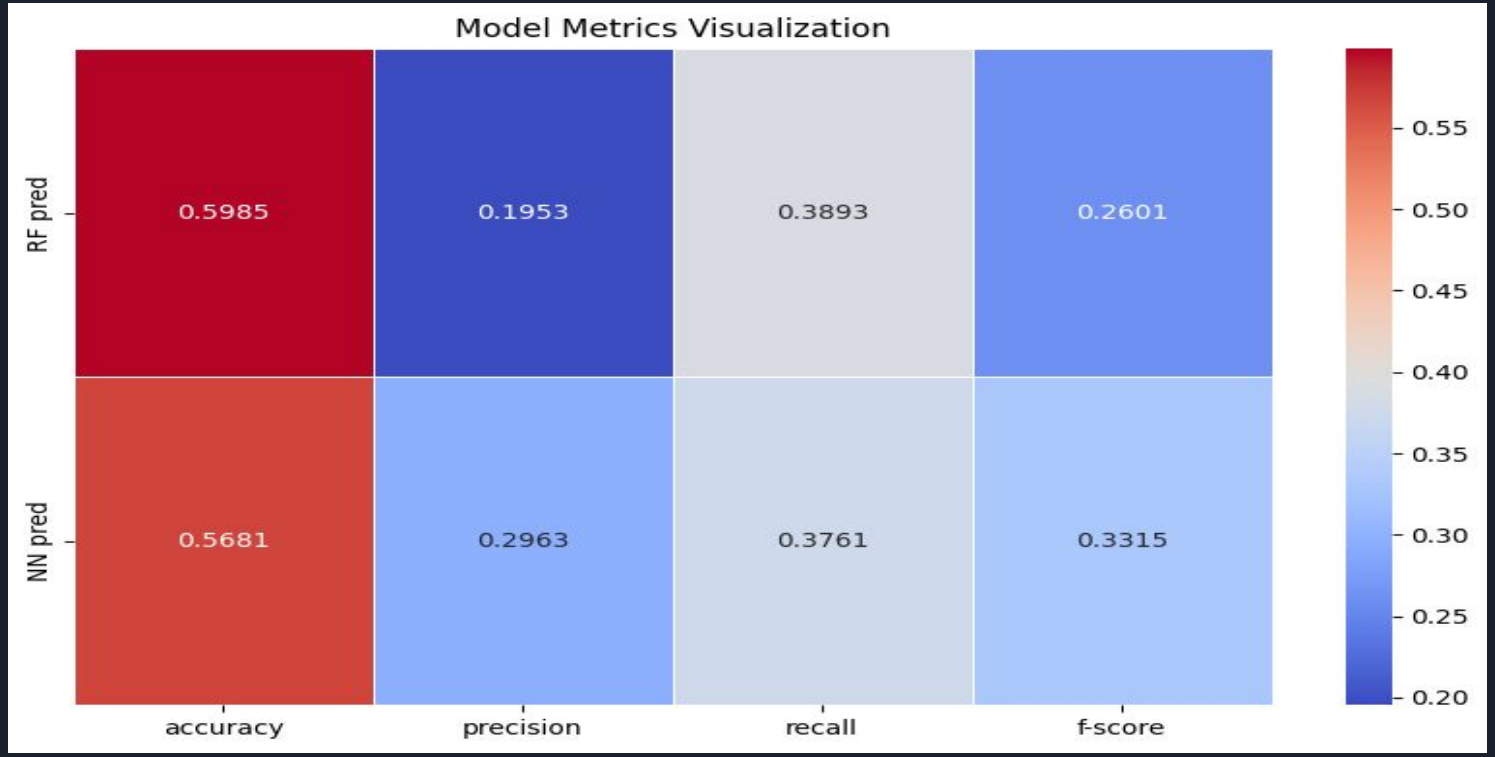
Baseline Experiments:
- Conducted experiments with a Neural Network (2 hidden layers) and Random Forest models.

SWEAR Words Model:
- Implemented a model specifically for predicting swear words in comments.

Error Analysis:
- Confusion matrices, accuracy, precision, recall, and f1-score examined.
- Identified challenges with dialects, English parts, misspellings

Label Interpretation Challenge:
- Labels occasionally open to personal interpretation.
- Binary label limitation noted in capturing the spectrum of toxicity.
- Some misclassifications observed due to variability in understanding toxicity levels.

Heatmap Model Metrics - Milestone 2



Concepts for final phase

- Goal: Identifying toxic comments, insights into “reasons” for model decisions
- Using German BERT models (deepset/gbert-base)

- Basic task: binary classification (toxic / non-toxic)
- Extended task: Named Entity Recognition (NER) to identify vulgar speech
  - Vulgar speech could be an indicator for toxicity
  - Comparison with misclassified comments to possibly gain insights into toxic speech

- Comparison of finetuning of “basic” GBERT model with one already trained for toxicity detection as an additional baseline

Toxicity classification

Challenge:
- Training a transformer model from scratch impractical due to time and computational constraints.
- Our dataset size posed challenges for feasible model training.

Model Selection:
- Utilized GBERT pretrained on Wikipedia for German language understanding.
- Added a sequence classification header for toxicity classification on our labeled data.

Comparison with Alternative:
- Explored Toxic-BERT-German pretrained specifically on toxicity classification.
- Assessed if task-specific pretraining enhances accuracy.

Training Data Approaches:
- Experimented with raw data and preprocessed versions, including lemmatization, stopword removal, and augmentation.

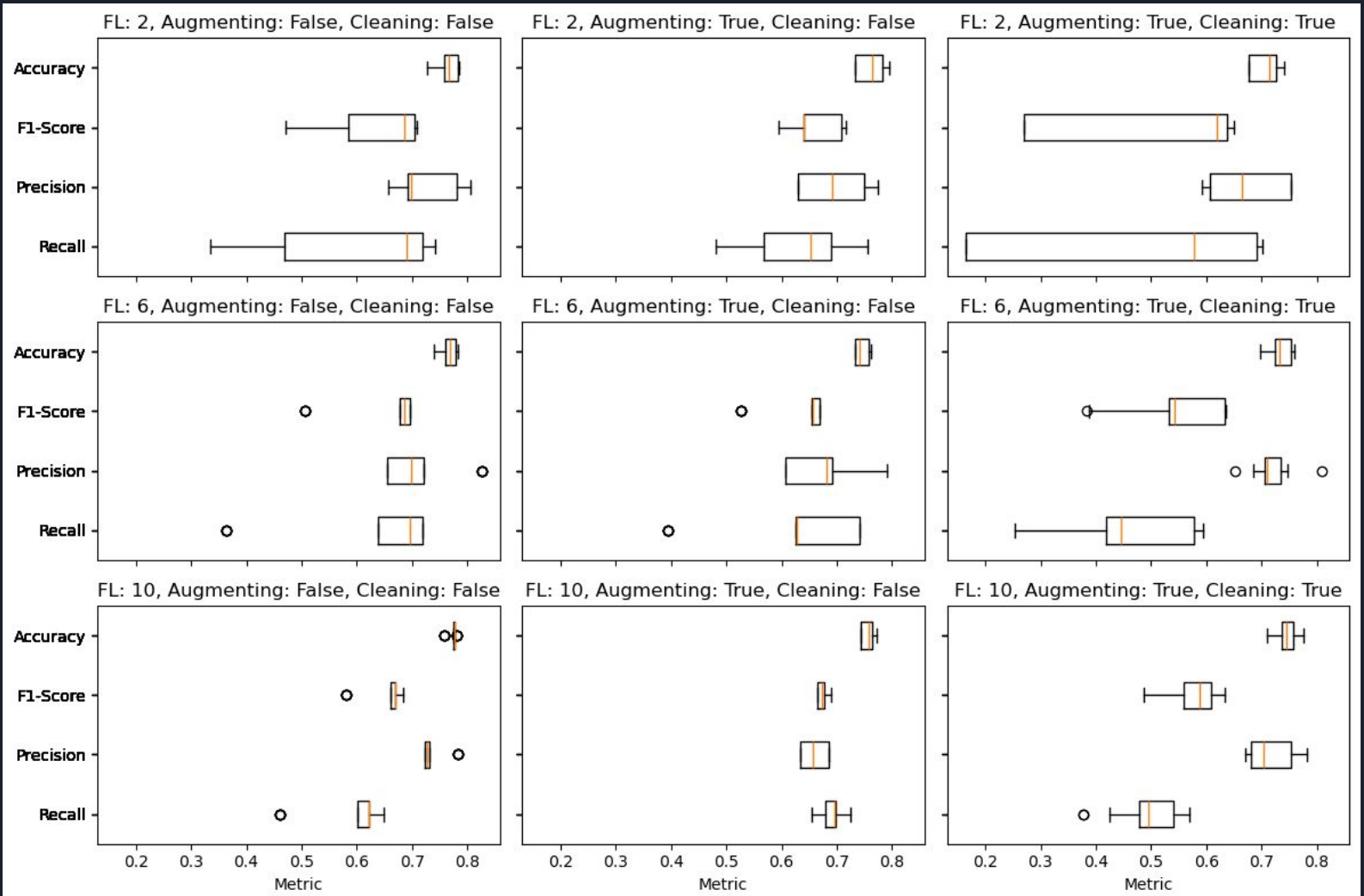
Experiments with GBERT classification model

Training setup

- Variations Explored:
  - Layer freezing (2, 6, 10 layers).
  - Preprocessing pipeline:
    a. Data augmentation (Class balancing).
    b. Data augmentation and cleaning (stopwords & punctuation).
    c. No augmentation or cleaning.

- Training Details:
  - 10 epochs max, with 5 runs per setup.
  - Early stopping after 3 epochs without improvement
  - Model checkpoints saved at every epoch (~150GB total!)

- Data Splitting:
  - Utilized the same split as Milestone 2 for optimal comparability, especially on the test set.


FL: 2, Augmenting: False, Cleaning: False
FL: 2, Augmenting: True, Cleaning: False
FL: 2, Augmenting: True, Cleaning: True
FL: 6, Augmenting: False, Cleaning: False
FL: 6, Augmenting: True, Cleaning: False
FL: 6, Augmenting: True, Cleaning: True
FL: 10, Augmenting: False, Cleaning: False
FL: 10, Augmenting: True, Cleaning: False
FL: 10, Augmenting: True, Cleaning: True
Accuracy
F1-Score
Precision
Recall
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Metric

Model Evaluation

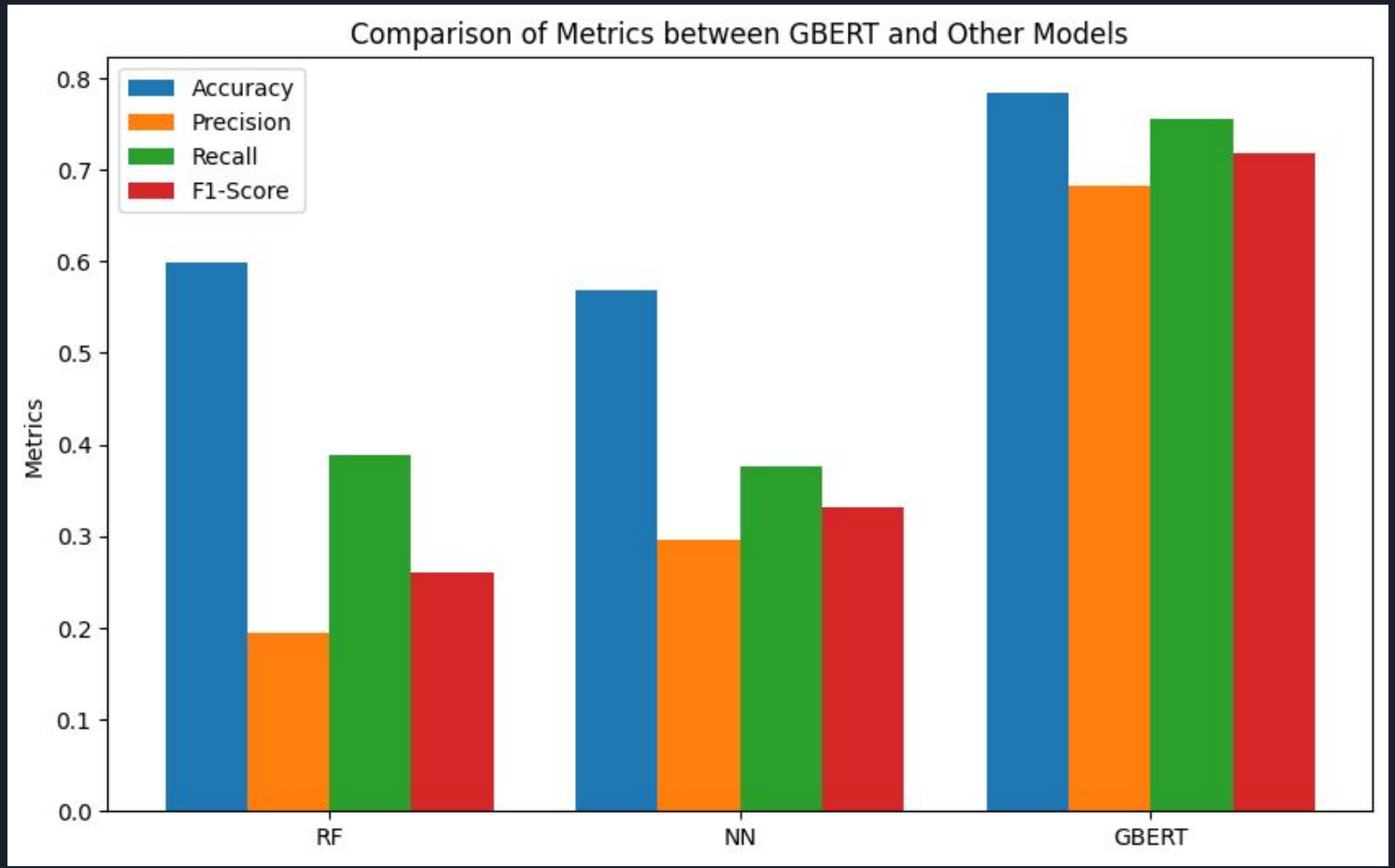
Top-Performing Models:

- Models with 2 frozen layers had the most variation but also the highest top performers.
- data cleaning step seems detrimental
- augmentation showed improved recall but worse precision, similar accuracy & F1.

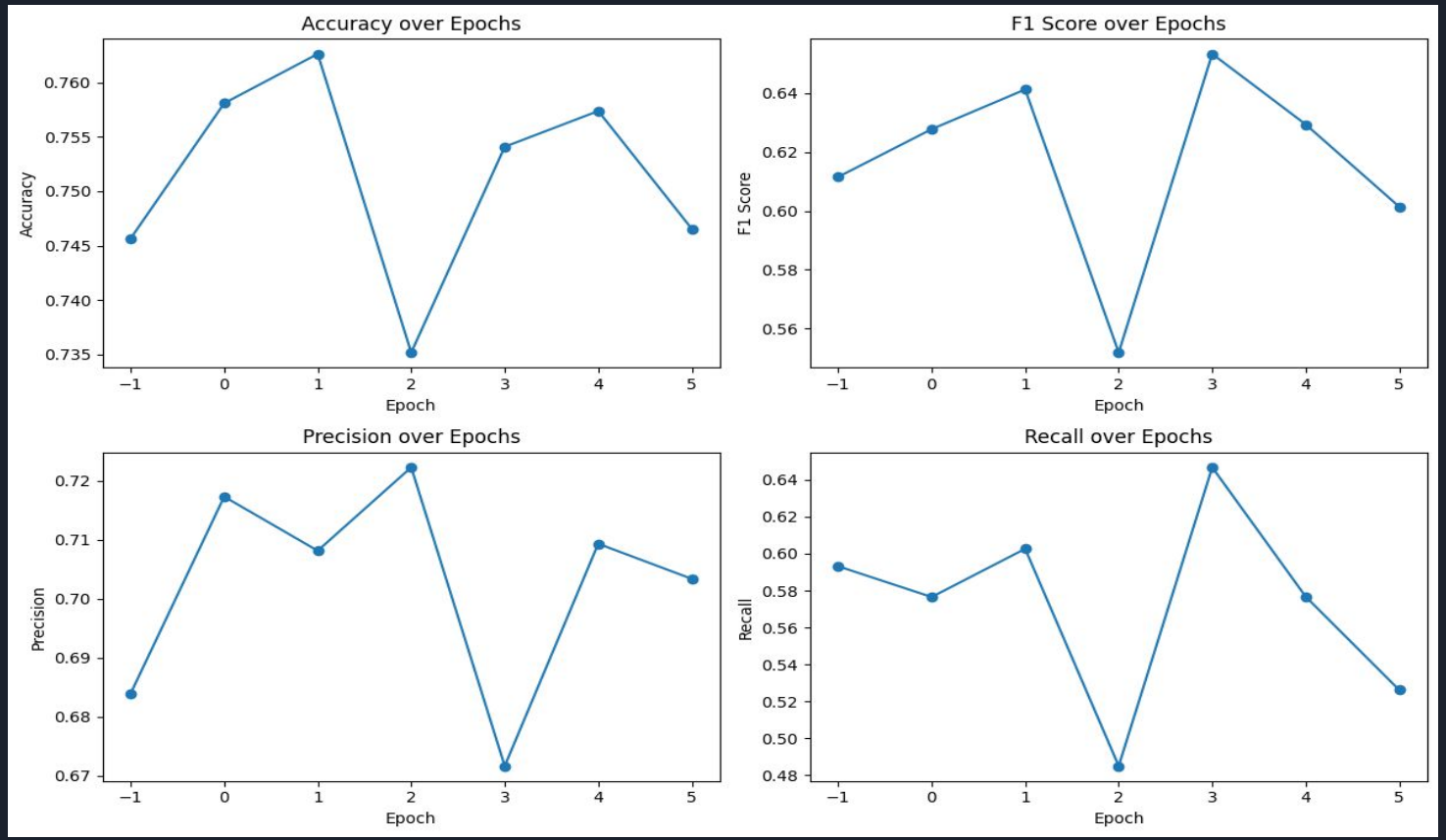
Best Models (Training Set Evaluation):

	id	setupid	fl	clean	augment	accuracy	f1	precision	recall
Best by									
accuracy	97	3	2	False	True	0.7944	0.7091	0.7305	0.6890
f1	99	3	2	False	True	0.7835	0.7175	0.6828	0.7559
precision	228	8	6	False	False	0.7409	0.5058	0.8258	0.3645
recall	99	3	2	False	True	0.7835	0.7175	0.6828	0.7559

Overall best depends on prioritization;
probably Model 99 (best F1: 0.72; near to best accuracy: 0.78)


Comparison of Metrics between GBERT and Other Models
Accuracy
Precision
Recall
F1-Score
Metrics
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
RF
NN
GBERT

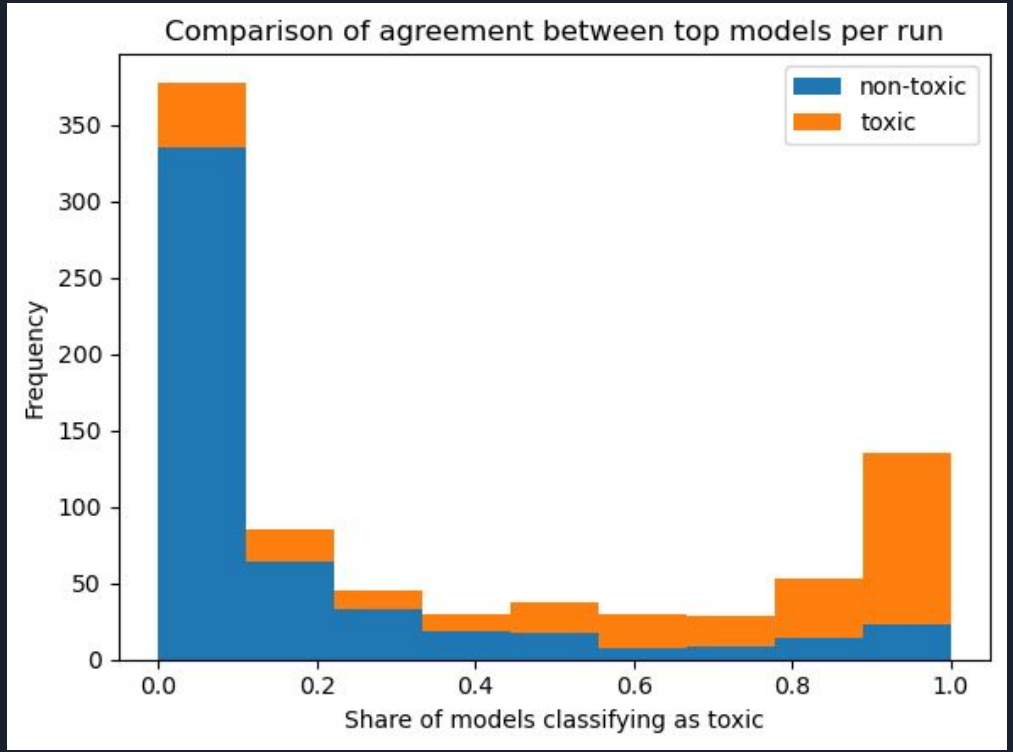
Model Performance Trends Across Epochs



Agreement between models across different setups

- Expectation: Most comments have a high degree of agreement between models
  (most land at 0 or 1)
- Reality: Yes, but there is a considerable portion “in the middle”, where models disagree

- Are these comments inherently ambiguous, or models just bad?



Experiment with a Pretrained model Toxic BERT German

- ankekat1000/toxic-bert-german
- Base model: bert-base-german-cased
- Fine-Tuning Data:
  - 14,000+ German user comments from media sites
  - Binary classification for toxicity
- Toxicity Definition:
  - Inappropriate: Rude, insulting, hateful comments
- Training:
  - Language: German
  - Model size (12GB)
  - Epochs: 4, Batch size: 32, Max tokens: 512
- Labeling:
  - Crowd annotation used

Experiment with a Pretrained model Toxic BERT German

Evaluation results:

- 64.5% accuracy without fine tuning.
- 73.7% accuracy with fine tuning, slightly lower than best result
- Model claims 86% accuracy on its own test set

Insights:

- Transfer learning not effective in this case.
- Possible language differences (german vs austrian)?
- Different kind of toxicity depending on platform / site
- Possible influence of slightly different base model?

NER Analysis and Training

- We decided to use the vulgarity tag to mark the words which are considered to be an insult, to maybe get insight on toxicity of comments
- Out of 4473 Comments 1306 contained a vulgarity tags.
- But the subset of samples marked as vulgar but classified as non-toxic (452) raises questions about the alignment between the 'Vulgarity' tag and the toxicity label. It suggests that not all comments with vulgar content are necessarily harmful or toxic.

Training

- Lower 2 and 6 layers frozen
- 5 and 10 epochs, with a batch size of 8
- Data prepared to have sentence by sentence data set tagged with vulgarity token. [‘O’,’Vul’] (1484 sentences)
- 80/10/10 train/validation/test split used

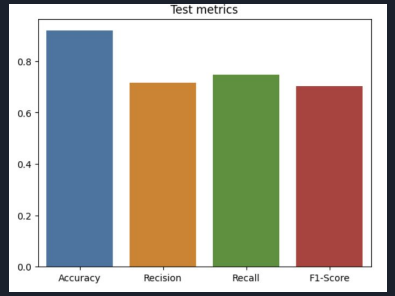
NER Model Evaluation

Predictions generated on a few test runs

- Metrics (Accuracy, F1, Precision, Recall) performed
- Model with few frozen layers showed slightly better results then one with more.
- More epochs lead to better effectiveness

Best model (Training Set Evaluation):

- 10 Epochs and 2 frozen layers (Acc: 0.933, Prec: 0.784, Rec: 0.797, F1: 0.767)



NER Examples

Reiche sind die grössten asozialen Schmarotzer auf dieser Welt (toxic)

Label ['O', 'O', 'O', 'O', 'Vul', 'Vul', 'O', 'O', 'O', 'O']

Pred ['O', 'O', 'O', 'O', 'O', 'O', 'Vul', 'Vul', 'Vul', 'Vul', 'Vul', 'Vul', 'O', 'O', 'O', 'O']

Was für ein Volltrottel (toxic and non toxic)

Label['O', 'O', 'O', 'Vul']

Pred ['O', 'O', 'O', 'Vul', 'Vul', 'Vul', 'Vul']

Ich hasse geradezu einige Teile dieser Menschheit (toxic)

Label ['O', 'Vul', 'O', 'O', 'O', 'O', 'O']

Pred ['O', 'O', 'O', 'O', 'O', 'O', 'O']

Herdimmunität usw. , was für ein Scheiß . (non toxic)

Label['O', 'O', 'O', 'O', 'O', 'O', 'Vul', 'O']

Pred ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'Vul', 'O']

Remark: different length because of bert tokenizer

Future Steps for Improved Predictions

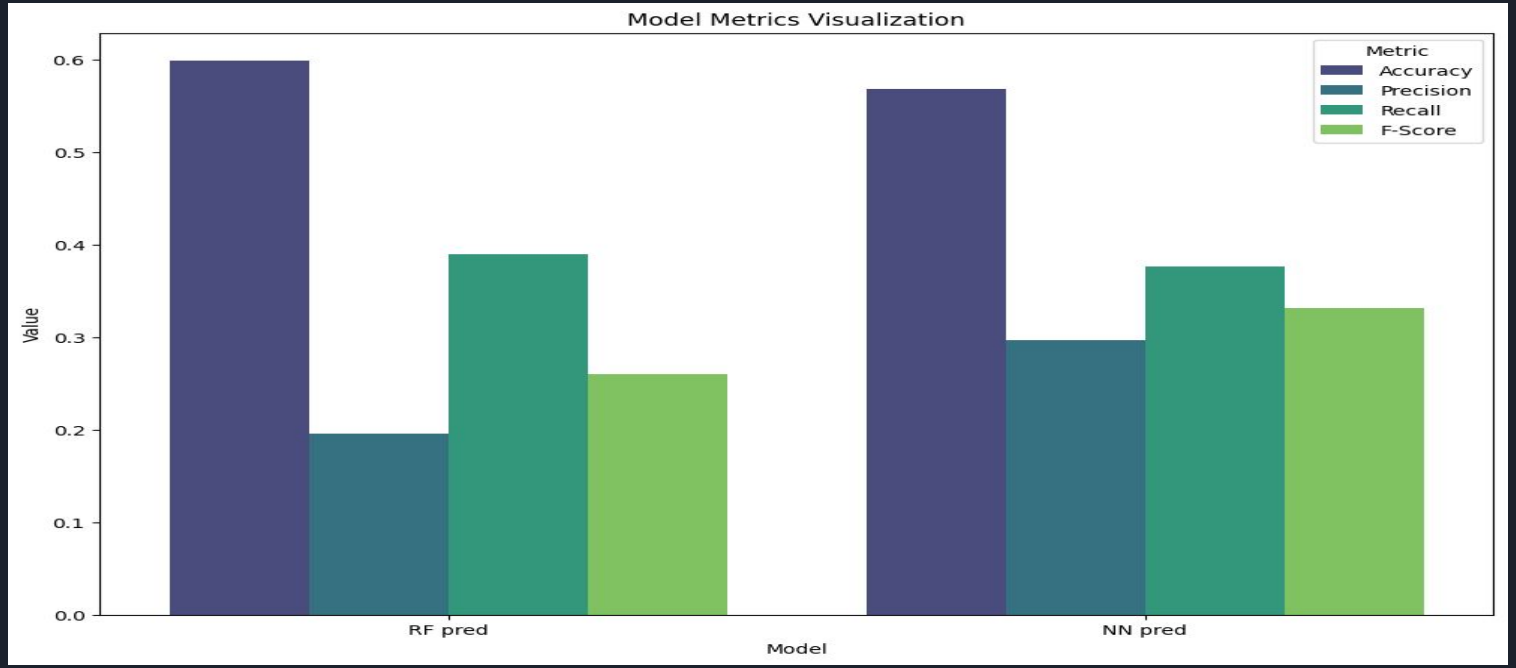
Exploring Ensemble Learning

- Opportunity: Investigate a combined decisionmaking approach for enhanced model performance by combining predictions from diverse models.
- Approach: Train multiple models with different parameters / pipelines, select top performers per setup, aggregate their predictions according to some criterion
- Potential Benefits:
  - (possibly) improved performance through collective model intelligence
  - Taking advantage of diverse strengths of different setups
- Challenges:
  - Computational & storage cost
  - Selection of ensemble decision criterion
- Next Steps:
  - Refine Existing Setup
  - Explore Ensemble Learning for BERT models
  - Perform qualitative error analysis similar to milestone 2

Thank you

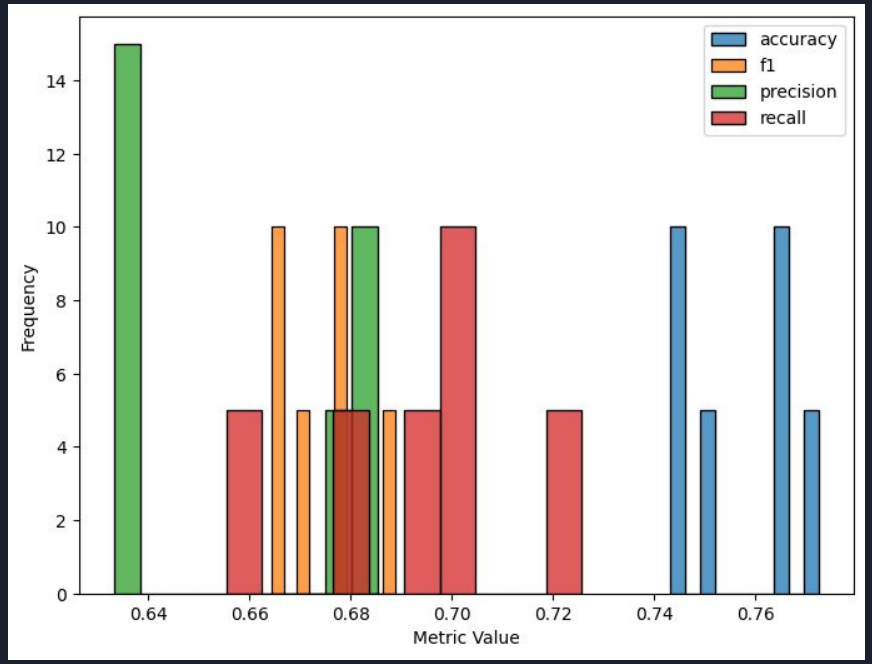
Questions and Answers

Model metrics comparison - Milestone 2



Distribution of Metric Values Across Setup - Example Frozen layer 10 applying data augmentation only

- Visual representation of metric distribution for different setups.
- Each setup is labeled based on frozen layers, cleaning, and augmentation.
- Provides insights into the variability of metrics across different experimental configurations.
- Helps identify trends or patterns in model performance under various conditions.



Summary of Best and Worst Models by Metric - Final phase

METRICS	BEST MODEL ID	SETUP ID	EPOCH	ACTUAL VALUE	WORST MODEL ID	SETUP ID	EPOCH	ACTUAL VALUE
ACCURACY	97	3	1	0.79	118	4	2	0.68
F1	99	3	3	0.72	118	4	2	0.27
PRECISION	228	8	2	0.83	119	4	3	0.59
RECALL	99	3	3	0.76	118	4	2	0.16