Management Summary - NLP Group 8

The primary goal of the project was to develop a Natural Language Processing model using GBERT and enhanced data annotations for effective toxicity assessment in German online comments. The process involved two milestones and a final stage, each contributing to the refinement and evaluation of the toxicity detection model. Our team used the German version of BERT (GBERT), a pre-trained deep learning model introduced by Google in 2018. BERT revolutionised language understanding by capturing bidirectional context in text, unlike traditional models that processed text in a unidirectional manner.

In the **first milestone**, the focus was on data preparation and exploration. We found 522 article titles in the dataset, and first tried to remove all the non-important parts to simplify the classification. Duplicates were removed, and majority labels were preserved. The resulting dataset comprised 37% toxic comments, totaling 1655 instances, and 2818 non-toxic comments.

The **second milestone** went into baseline experiments and detailed error analysis. Baseline refers to the initial set of models and methods used for toxicity assessment before further refinements and advanced techniques were applied. Comments were lemmatized, and a simple bag-of-words approach was employed for baseline models. The remarkable aspect used in the development of a specialised model designed to predict offensive language in comments, focusing on a specific dimension of toxicity. Error analysis involved the examination of confusion matrices, accuracy metrics, and a comprehensive evaluation of precision, recall, and F1-score. Our team encountered challenges, such as dealing with dialects, English components, misspellings, and occasional ambiguity in label interpretation, but successfully navigated through them.

At the **final phase**, the project aimed to refine the existing model and explore more advanced concepts. The goal was to identify toxic comments and provide insights into the underlying reasons influencing model decisions. Our team planned to leverage German BERT models for binary classification, extending the task to Named Entity Recognition (NER) for identifying vulgar speech. The comparison was made with a fine tuned "basic" GBERT model and one already pre trained for toxicity detection. This model is called Toxic GBERT, which is specifically developed for toxicity, so we could also incorporate it into our testing.

The experiments with GBERT classification models involved various setups, including layer freezing and different preprocessing pipelines. Model evaluation revealed that models with two frozen layers exhibited the highest variation and top performance. The exploration of pretrained models,

specifically Toxic BERT German (TOXIBERT), provided insights into the challenges of transfer learning, language differences, and variations in toxicity across different platforms.

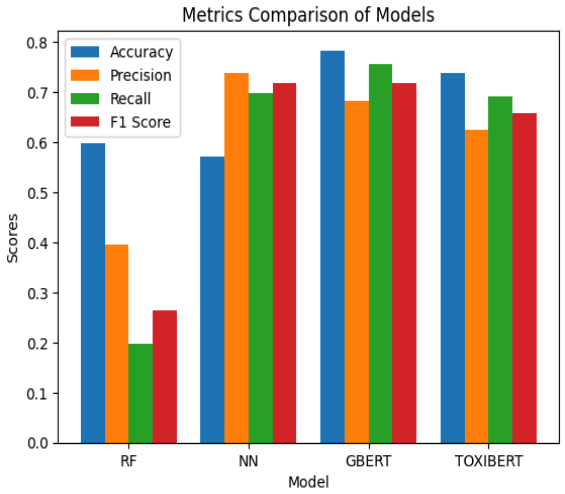

Table 1: Barplot - Metrics comparison

Table 2: Comparison of evaluation metrics

Model	Accuracy	Precision	Recall	F1 Score
RF	0.59854	0.395973	0.197324	0.263393
NN	0.57056	0.738362	0.698027	0.717628
GBERT	0.783455	0.682779	0.755853	0.71746
TOXIBERT	0.737226	0.625378	0.692308	0.657143

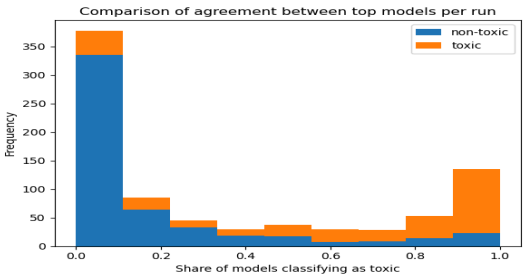

Table 3: Comparison between models / setups

The project also explored Named Entity Recognition (NER) analysis and training, employing a distinctive method by utilising vulgarity tags to identify words considered insults. This approach marked a departure from the methods used earlier in the project. The evaluation of the NER model showed improvements in evaluation metrics, particularly with fewer frozen layers and additional epochs.

As the project reached the present stage, the possible next steps included exploring collaborative learning to enhance the model's performance. Our team planned to train multiple models with different parameters and pipelines, aiming to leverage the combined intelligence of diverse setups. While challenges like managing computational costs and time emerged, these steps could potentially represent a future phase of the project if we could delve deeper and continue our work.

In conclusion, the project made significant steps in data preparation, baseline experiments, and model evaluation. The focus on refining the existing model, exploring advanced concepts, and considering ensemble learning set the stage for continued progress in the quest to effectively detect toxicity in online comments.