Hate Speech Detection Project

Hate speech detection on social media

Objective:

Nowadays spreading opinions, especially hate, has been made easer and easier with the use of social media and the characteristics of anonymity of those channels

1. Can bring **harm** to affected parties
2. **Rapid increase** in hate speech on **social platforms** calls for **effective** automated **detection**

Project **aims** to:

A. explore tools applied to **hate speech detection** with classical **machine leaning** approaches, **transformer models** and **LLM-based** classification

B. use **data science analysis** approaches to get an understanding of what might affect hate speech

Why hate speech detection? – Social relevance and Challenges

- Social impact: HS stimulates **violence**, spread of **misinformation**, contributes to **social polarization**; classically affects minority groups or overall community safety
- Regulatory Pressure: Governments/ platforms are under the **pressure to moderate harmful content**

The need for **interpretable and reliable models** that support **content monitoring** and **policy making** is becoming ever more present.

Objectives and Research Questions

Main objectives:

- Comparison of **multiple modeling approaches** for hate speech detection with **performance and error analysis**

Research Questions:

- Which modeling approach best handles the **nuanced nature of german hate speech** on social media?
- How can **interpretability** enhance the **understanding** of model decision?

Contribution:

- Comprehensive evaluation and comparison of diverse approaches
- Error analysis and potential improvements
- Challenges of hate speech detection in real-world datasets

Detox Dataset – Details 1

- **Twitter comments**, specifically for *German twitter accounts*, specifically downloaded by the use of **specific keywords**, to ensure that the dataset will be a **mostly negative sentiment** and **toxic comments**
- **Twelve different** annotation categories
- Annotated by at least **1-7 experts**, always at least 3, and then the ratio has been taken for the hate_speech column of how many annotated with y or n. (continuous)
- Hate_speech than might be discrimination of **age, gender, ethnicity, religion, and sexual orientation** and more

Detox Dataset – Details 2

- **Rich metadata** like the **legal paragraphs** under which possible prosecution could be filed, when **harmful** comments are annotated
- **Additional Attributes:** sentiment, explicit/implicit expression, target type (person, group, public), and discrimination dimensions.
- **Annotation Quality:** Aggregated annotations from **multiple human evaluators**, capturing subjective nuances under strict evaluation guidelines

Detox Dataset - Limitations

Imbalance and Sparsity:

- Some classes (e.g., extreme hate speech) are underrepresented.

Subjectivity in Annotation:

- Even among multiple annotators, interpretation of hate speech can vary.

Context and Ambiguity:

- Comments may require contextual information (cultural, political) that is not present in the dataset.

Data Quality:

- Presence of noisy text, informal language, and possibly typos or slang

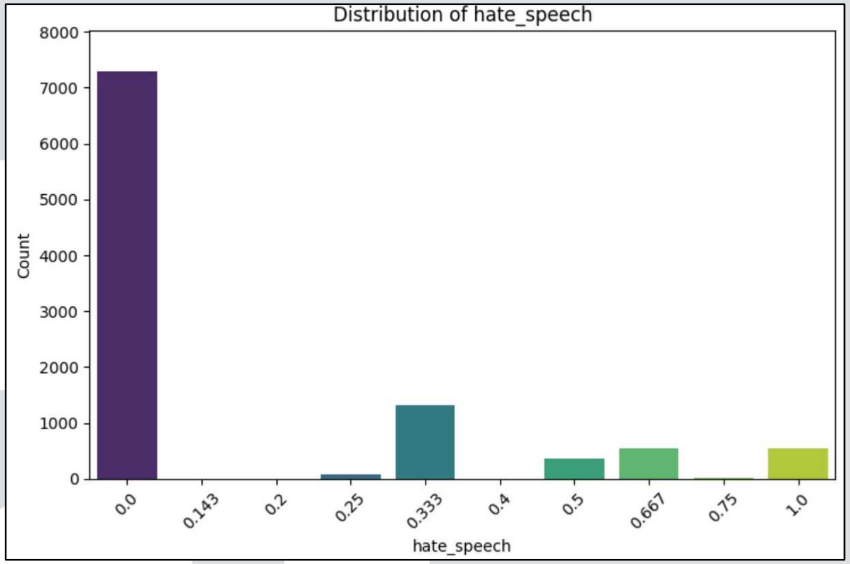
DeTox Overview:

Types of Discrimination that have been annotated with this dataset are:

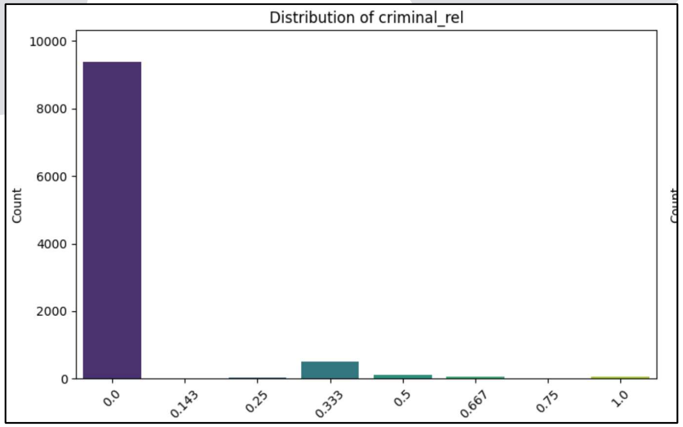
- Job; Political Attitude; Personal Engagement and Interests; Sexual Identity; Physical, Psychological or Mental Characteristics; Nationality; Religion; Social Status; World View; Ethnicity
- Around ~10.000 comments with 41 columns

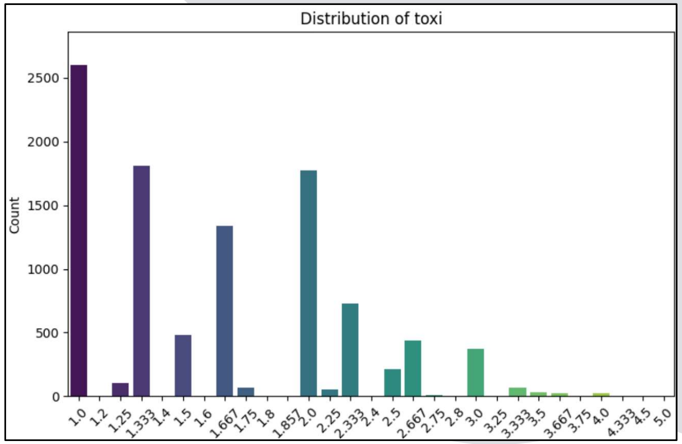
Key attributes:

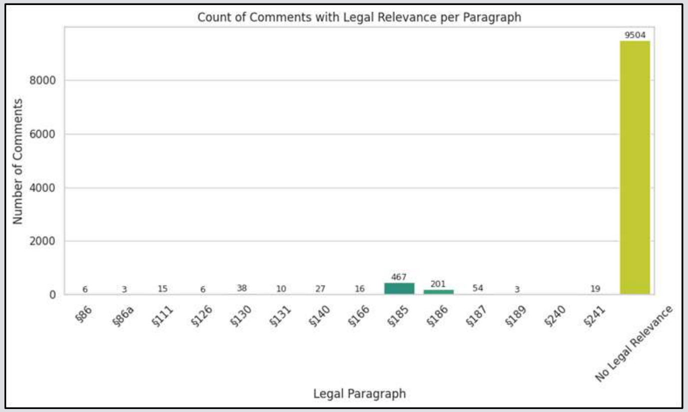
- c_text: the comment text
- Hate_speech: continuous score (later thresholded)

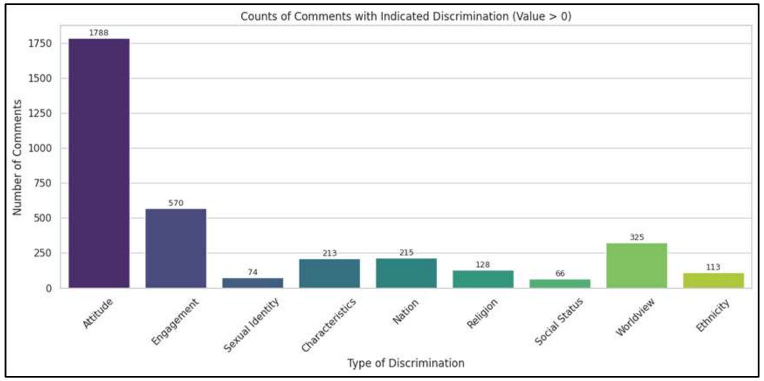


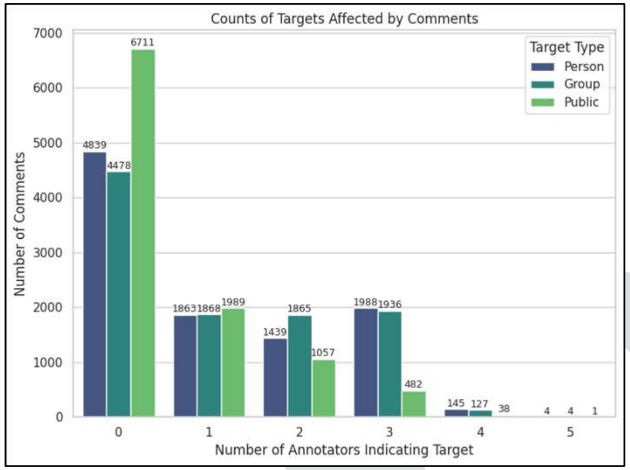
Some other distributions

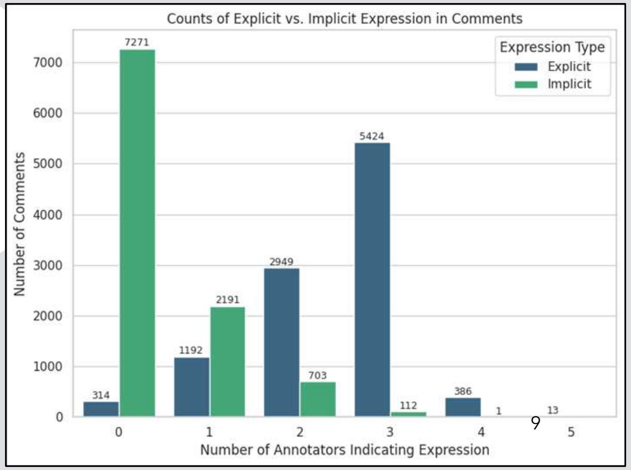




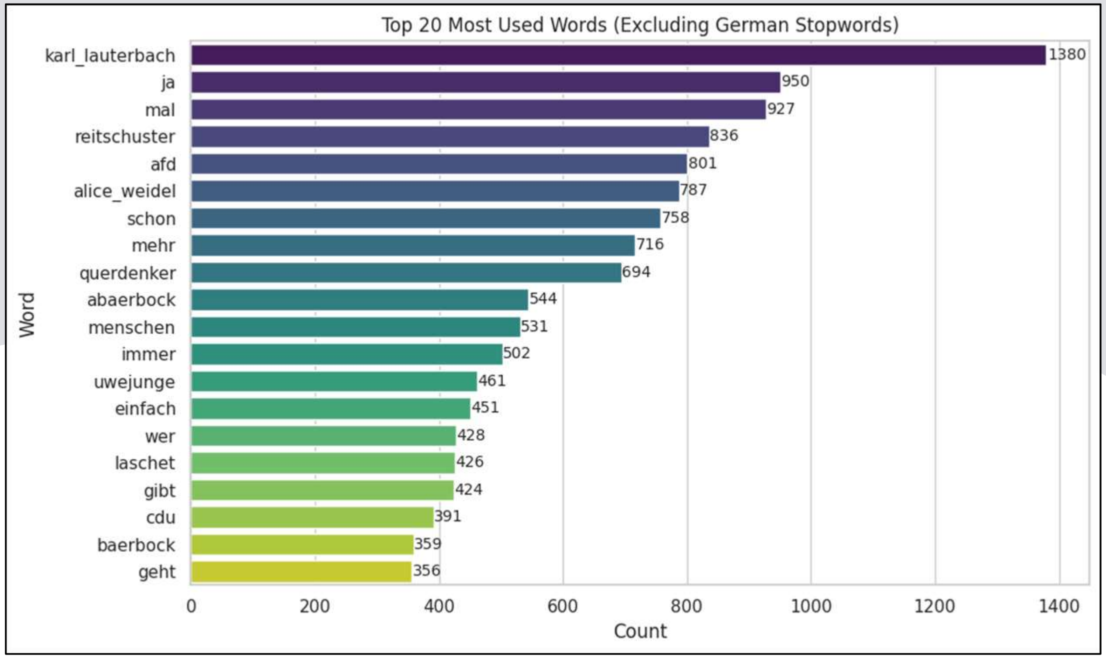


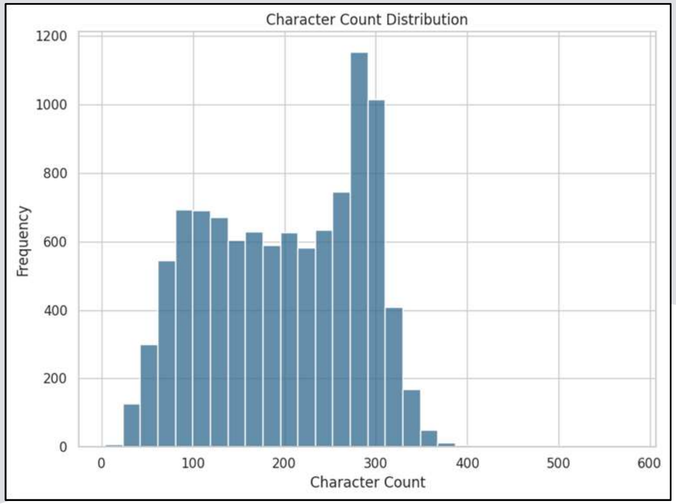


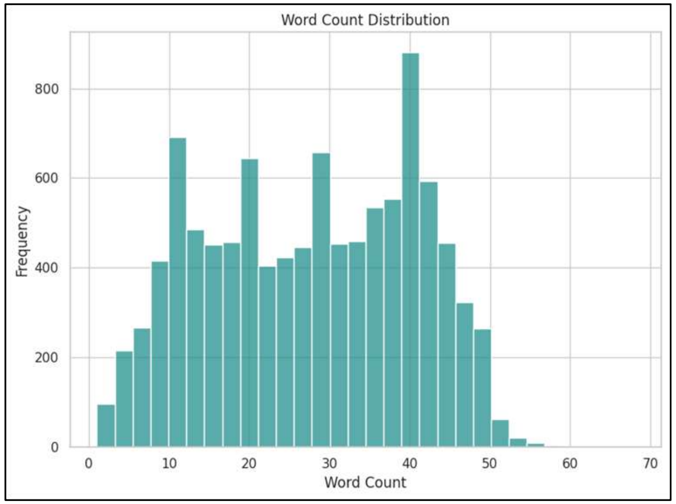




Text characteristics – 1

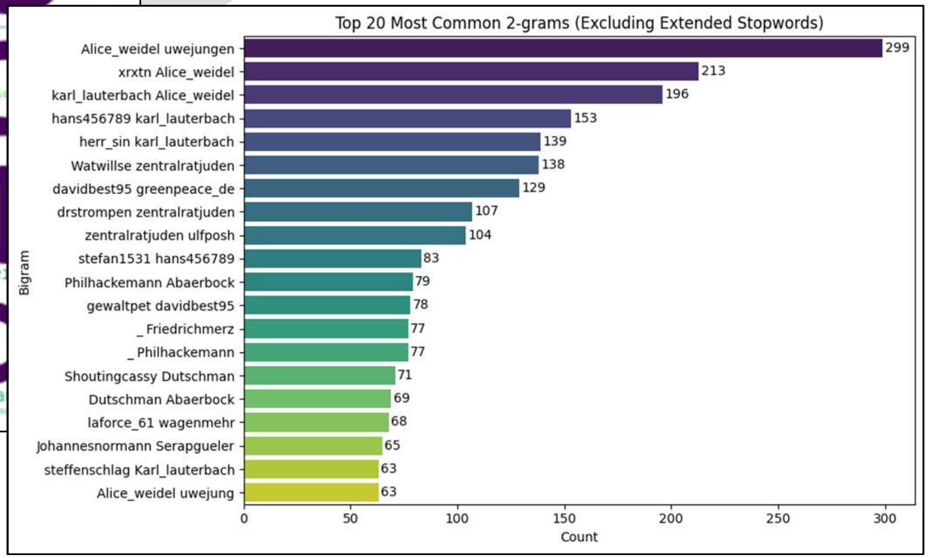




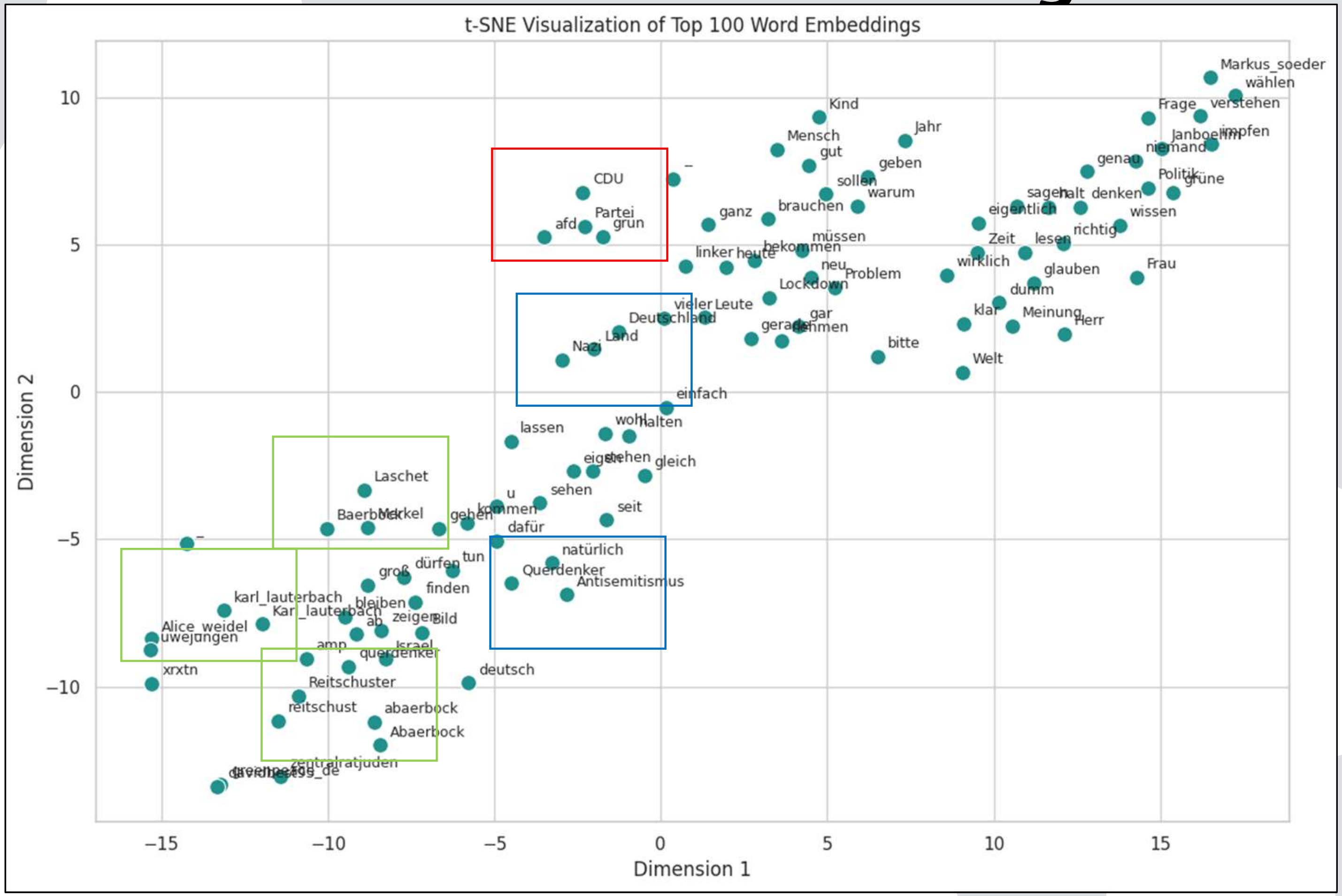


- **Highly political content** ~ most used words include mostly politicians
- On average the **comments** have around **30 words**

Text characteristics – 2



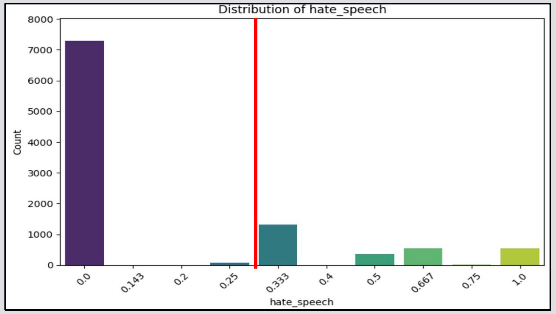
T-SNE word embedding



Observation:

- Cluster of seemingly **semantically similar** top 100 words
- Color red represents german parties
- Color green are politicians
- Color blue represents Germany but also descriptory words of Germany

Threshold for labeling hate_speech – Considerations



<u>Continuous Annotation</u> originally:

- The original dataset provides a **continuous hate speech score**, representing the **proportion** or intensity of hate speech as **rated by annotators** -> need for **binary classification**

<u>Threshold Selection</u>: chose a threshold of **0.33** (i.e. if the hate speech score > 0.33, label as hate speech:

- Reflects the insight that when **more than about 33%** of annotators flag a comment as hate speech, it is **likely to be significant**
- Helps balance the trade-off between sensitivity (capturing subtle cases) and specificity (avoiding false positives)
- **Lower Threshold:** Could capture more subtle or borderline cases but may increase false positives
- **Higher Threshold:** Might reduce false positives but risk missing milder yet concerning hate speech

Overview of Approaches –
Modeling Approaches for Hate Speech Detection

Transformer Models (Hugging Face):

- Leverage **pretrained** language models for **nuanced understanding** of context
- **Two** variants used: a **domain-specific model** (e.g. "oliverguhr/german-sentiment-bert") and a **general-purpose model** ("bert-base-german-cased").

ChatGPT (LLM Approach):

- Uses **prompt-based classification** with GPT-3.5-turbo
- Benefits from few-shot prompting and robust language understanding

Classical Machine Learning (TF-IDF + XGBoost):

- Uses traditional **text vectorization** with TF-IDF followed by an XGBoost classifier
- Provides a **baseline** and **interpretable features**

Hate speech detection tansformer Based

Model – 1 (oliverguhr/german-sentiment-bert)

- **Rare german based** hate speech detection models could be found on hugging face, especially not only trained on the DeTox Dataset
- **Oliver guhr's** is trained on a **variety of comment-based text format** data sets with the use case of detection of hate speech
- Uses the **BERT architecture** and was trained on **1.8 million German-language** samples of **twitter, facebook, amazon, other reviews, etc**.
- **Most** downloaded model found with around **265,000 downloads**

- Trained for **2 epochs** with a **batch size of 16**, using **early stopping (patience=1)** and an **increased weight decay (0.05)** to mitigate overfitting
- Used hugging face 'Trainer' function
- Pre-set validation set of 20% and learning rate of 2e-5 evaluating to select the best model

Hate speech detection transformer Based

Model – 2 (bert-base-german-cased)

- **Rare german based** hate speech detection models could be found on hugging face, especially not only trained on the DeTox Dataset
- **bert-base-german-cased** is trained on a **Wikipedia dump, OpenLegalData and News articles with around 12GB** of data uses the **BERT architecture**
- Performance as a comparison to the pretrained trans

- Same training specifications as previous transformer model

Zero shot LLM labeling

Model 3 - (ChatGPT model o3 – Prompt Engineering)

- Utilizes **GPT-3.5-turbo** with a carefully designed prompt in German, since it seemed to be the **most cost efficient**

German prompt text

"Du bist ein Experte für die Analyse von Hassrede in deutschen Texten. „
"Bitte klassifiziere den folgenden Kommentar als Hassrede oder nicht. „
"Gib als Antwort nur '1' aus, wenn der Kommentar Hassrede enthält, und '0', wenn nicht."

- The prompt instructs the model to **output "1" for hate speech and "0" otherwise**
- Instructs to pretend to be an **expert in German language labelling**

Zero shot LLM labeling

Model 3 - (ChatGPT model o3 – Cost estimation)

- Model has been iteratively interrogated with the prompt composed by the **instruction, as system message**, used to **deliver the task** and **the comment, a human message,** to deliver the comments
- Implemented with **asynchronous API calls and checkpointing to manage rate limits** utilizing the openai and langchain python libraries for communication with OpenAI APIs
- Rate limits were set for **500 requests/min**
- With the use of the asynchronous API calls the **labelling spanned 2 days** (one day ~30min, one day ~5min)
- **Cost estimation** before rendering the code to see if task is feasible:

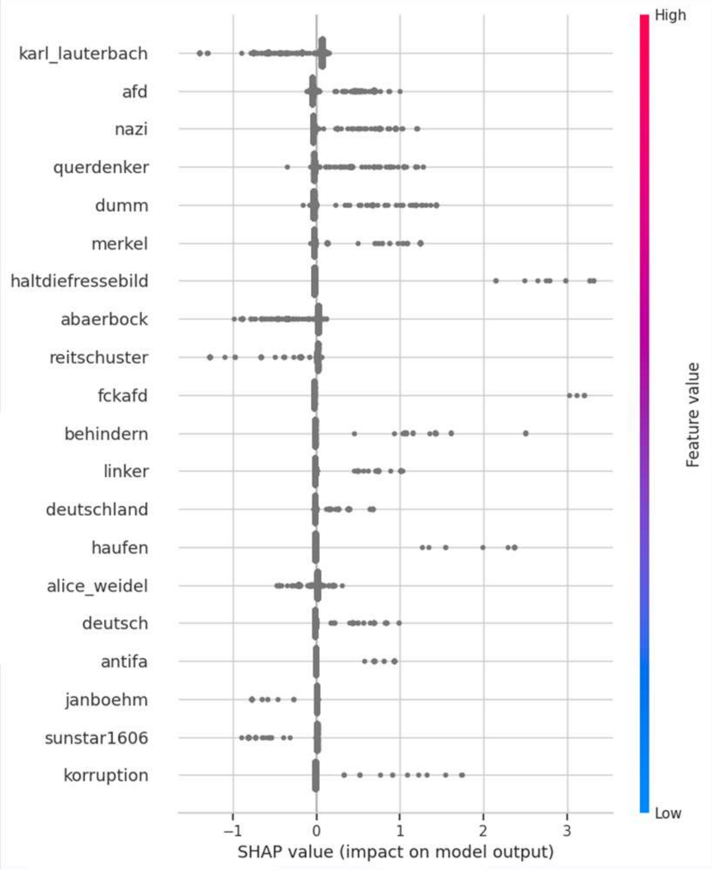
Metric	Value
Total Comments	10,284
Total Tokens	282,396
Average Tokens per Comment	27.46
Price per 1K Tokens ($)	0.0020
Estimated Cost ($)	0.5648

Table 1: Table of Metrics

Model 4 – Classic ML approach (TF-IDF + XGBoost Pipeline)

- **TF-IDF vectorization converts text** into **numerical features**
- **XGBoost** classifier learns to **distinguish** hate speech from non-hate speech
- Provides a transparent baseline with **interpretable feature** importance

- Highly political words show the highst importance in the dataset, most leaning towards a negative sentiment
- First place makes Karl Lauterbauch our favorite politician
- Concerning, that places 2 to 4 are all about far right politics



Performance Comparison

1 - Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Transformer Model 1	0.816	0.696	0.577	0.631
Transformer Model 2	0.790	0.650	0.500	0.565
ChatGPT	0.664	0.469	0.859	0.607
TF-IDF + XGBoost	0.762	0.633	0.298	0.405

Table 2: Model Performance Metrics

Performance Comparison

2 - Model Performance Metrics - Interpretation

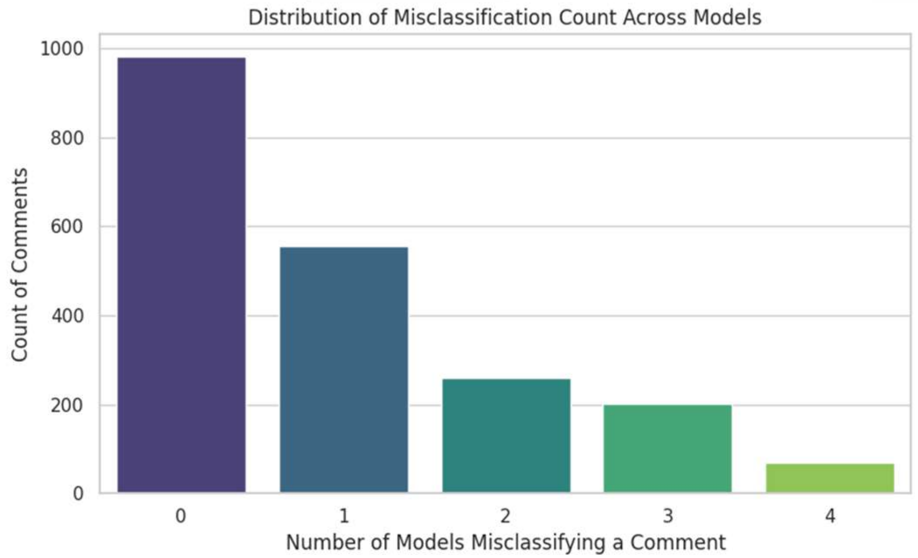
Model	Accuracy	Precision	Recall	F1 Score
Transformer Model 1	0.816	0.696	0.577	0.631
Transformer Model 2	0.790	0.650	0.500	0.565
ChatGPT	0.664	0.469	0.859	0.607
TF-IDF + XGBoost	0.762	0.633	0.298	0.405

Table 2: Model Performance Metrics

- Highest **overall accurarcy** done by **transformer model 1 (hate-speech detection model)**
- Transformer **Model 1 achieves the highest F1 (0.631)**, indicating a **strong balance** between **precision** (0.696) and **recall** (0.577)
- Transformer Model 2 is **slightly less balanced** (F1 = 0.565), with a modestly higher precision (0.650) but lower recall (0.500)
- ChatGPT **prioritizes recall** (0.859) at the expense of precision (0.469), catching **most hate speech** but **flagging more false positives**
- TF-IDF + XGBoost is conservative (precision = 0.633) but misses many true hate speech cases (recall = 0.298), resulting in a lower F1 (0.405)
- **Overall Trade-Off**: <u>No single model is universally best</u>; each suits different priorities (high recall vs. balanced approach vs. fewer false positives)
- **Recommendation**: If catching **most hate speech is critical**, **ChatGPT's high recall** may be preferable; for a balanced approach, Transformer Model 1 stands out

Performance Comparison

3 – Error analysis (0) – Examples



- Total amount of **1088 comments** are misclassified **by at least one** of the models, whereas **70 comments** are misclassified by **all models** in the validation set

Performance Comparison

3 – Error analysis (0) – Examples

- *„@Schmitz441 @Andrea_Stoll_ @PhilHackemann @ABaerbock diese diskussion hat vielmehr mit der armseligkeit derer zu tun, die sie angezettelt haben. vielleicht ist die diskussion ja auch ein ablenkungsmanöver von maskendeals, charaktermängeln von Kanzlerkandidaten und rechtsextremen ex-verfassungsschützern in den eigenen reihen"*

Comment with specific content for which a certain prior knowledge would be needed

- *„@Xaeni2 @ABaerbock Nein, das ist eben nicht so im Umgang Männern und erst Recht nicht unter Männern. Weder bei der umfangreichen Diskreditierung von Qualifikation noch im Hinblick auf Sexismus oder die vielen irren Geschichten die jetzt erfunden werden"*

Without prior context not classifiable

- *„@Karl_Lauterbach Ich finde es nicht besser, wenn Regenwald für Soja-Monokulturen abgeholzt wird, um den vegetarischen Bedarf zu stillen"*

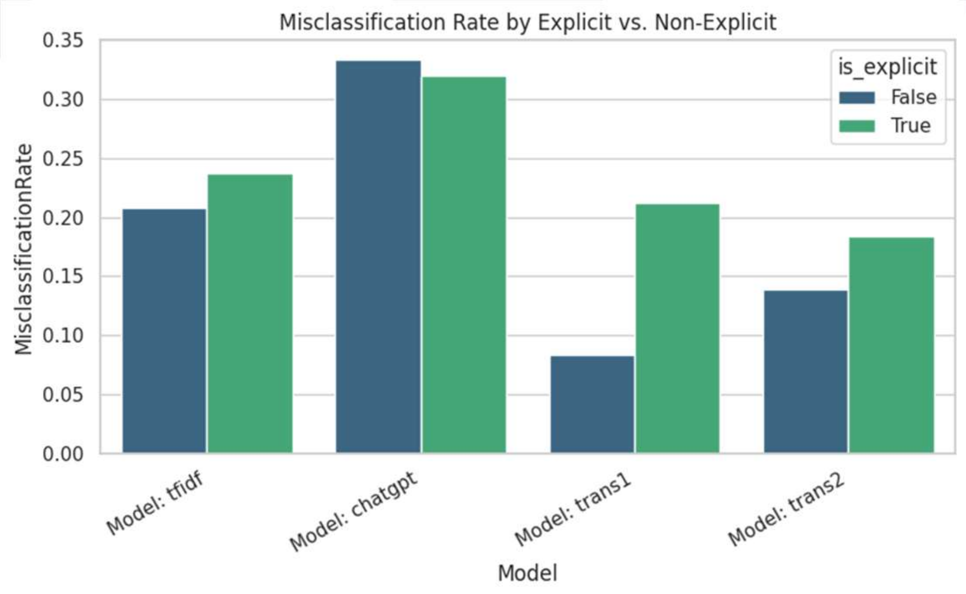
Mostlikely hate speech, but written in neutral way

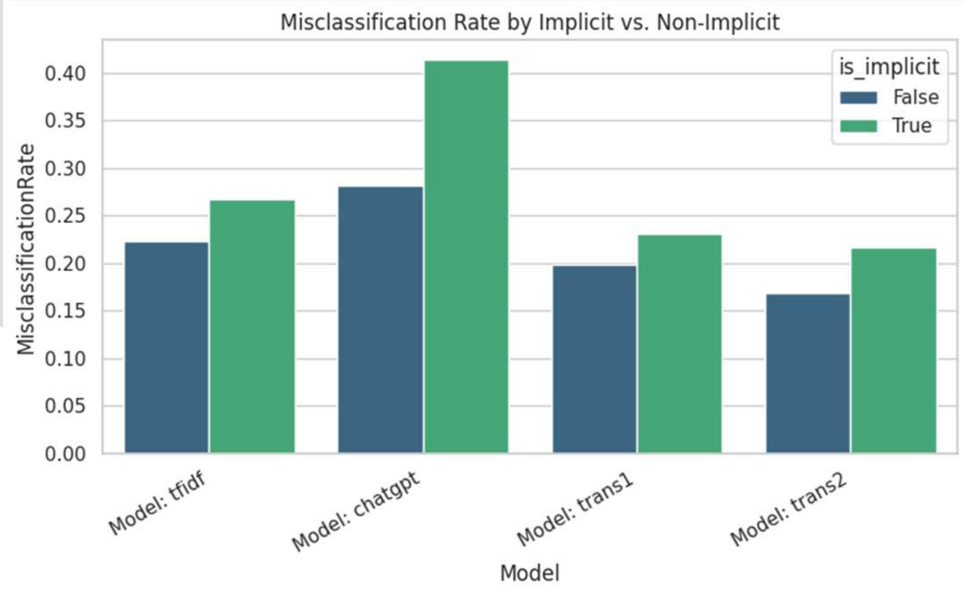
- *„@ErichCzerwonat2 @Mecklenburger11 @LillyBlaudszun @janboehm sieh doch einfach ein dass es tatsächlich ein Problem mit Rechten gibt? Auch 25 sind 25 zu viel"*

No hate speech without context

Performance Comparison

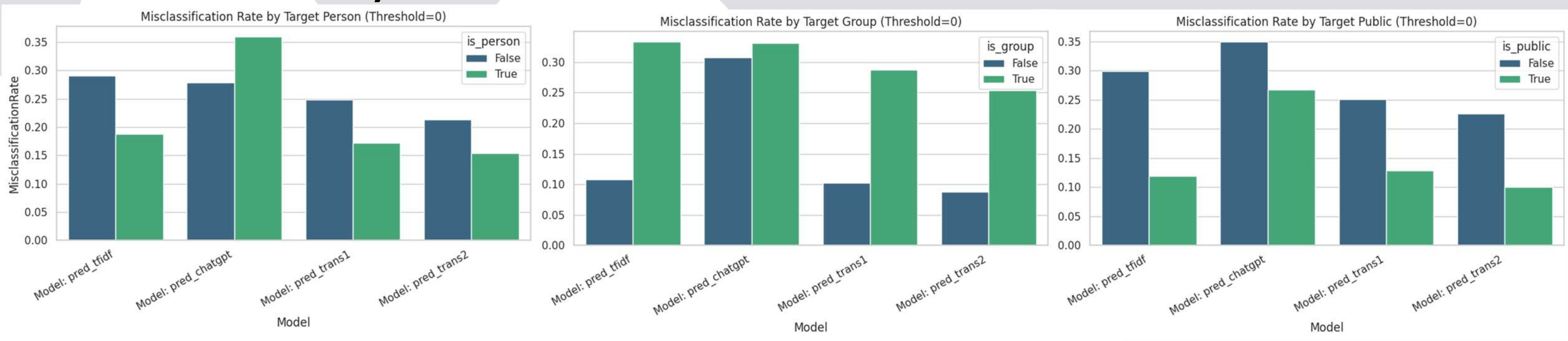
3 – Error analysis (1)





- Total amount of 573 comments marked as both, implicit and explicit
- Across all models, **implicit** hate speech consistently **shows higher misclassification** rates than **explicit**, especially for ChatGPT
- Suggests that **subtler** or implied hateful content is **harder to detect**

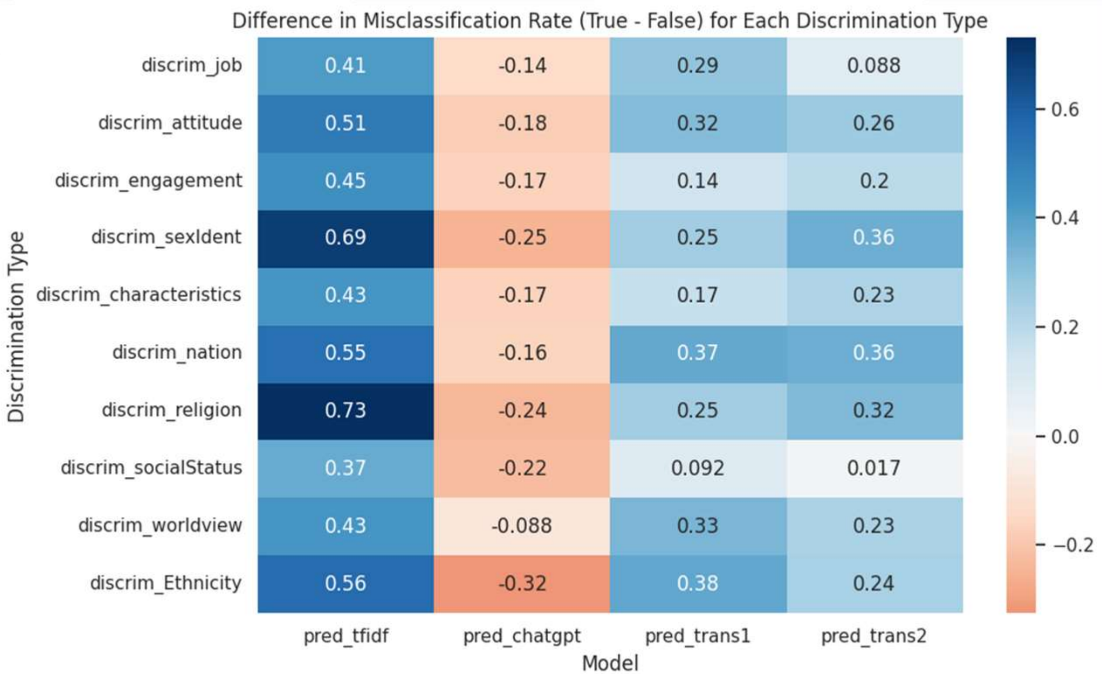
Performance Comparison
3 – Error analysis (2)



- **Group-targeted** comments have a **notable jump** in misclassification for **all models**, indicating the complexity of group-based hate speech
- **Public** targets see **relatively lower** error rates—possibly because public-figure discourse is more explicit or well-defined
- **Person** targets, interestingly, pose **more trouble for ChatGPT**, whereas TF-IDF and the transformers handle them better

Performance Comparison

3 – Error analysis (3) – discrimination types



- **TF-IDF + XGBoost** consistently shows **positive** (and often large) values across all discrimination types, implying it misclassifies **more** often on comments flagged with those discrimination attributes than on others
- **ChatGPT** exhibits **negative** differences for nearly every category, suggesting it actually misclassifies **less** when a discrimination type is present—indicating that ChatGPT may handle overt discrimination cues more effectively than subtle ones

- **Transformer Models** (pred_trans1, pred_trans2) generally have **positive** differences, but to a lesser extent than TF-IDF. This implies they also find discrimination-focused comments more challenging, though not as drastically as TF-IDF, with the largest challenges appearing around categories like nation, religion, or ethnicity

Conclusion & Lessons Learned

- **Diverse Model Strengths:** Transformer Model 1 balances precision and recall best; ChatGPT excels at recall (catching subtle cases) but has more false positives. TF-IDF is straightforward yet struggles with more nuanced or explicit content
  Successfully implemented multiple approaches—transformers, ChatGPT, and TF-IDF—revealing distinct strengths and trade-offs in detecting hate speech

- **Complex, Nuanced Data:** The DeTox dataset's explicit vs. implicit labeling, target attributes, and discrimination flags exposed subtle language cues that challenge all models
  EDA and error analysis showed how implicit content, group-focused discrimination, and overlapping annotations significantly impact misclassification rates

Future Improvements

What could be thought of for the future:

- **Contextual & Multi-Modal Inputs:** Incorporate conversation context, user history, or external knowledge (e.g., social/political context) to handle ambiguous or coded language
- **Attribute incorporation:** Including available attributes of the dataset into classification pipeline to enhance/ level-up hate speech detection performance
- **Augment Training Data:** Expand domain-specific examples, especially for implicit hate speech and minority discrimination types, to reduce data sparsity and better capture subtle patterns