St. Pölten University of Applied Sciences

/ medien & digitale technologien

/fh/// st.pölten

Neural Networks for Hate Speech Detection on German Social Media

A comparison of Transformer Models for German Hate Speech Detection in contrast to Human Evaluation

Zweite Bachelorarbeit

Ausgeführt zum Zweck der Erlangung des akademischen Grades
Bachelor of Science in Engineering

am Bachelorstudiengang Medientechnik
an der Fachhochschule St. Pölten

von:
Verena Tiefenthaler
mt191096

Betreuer*in: Jaqueline Böck, BSc

St. Pölten, 08.07.2022

Ehrenwörtliche Erklärung

Ich versichere, dass

- ich diese Arbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfe bedient habe.

- ich dieses Thema bisher weder im Inland noch im Ausland einem Begutachter/einer Begutachterin zur Beurteilung oder in irgendeiner Form als Prüfungsarbeit vorgelegt habe.

Diese Arbeit stimmt mit der vom Begutachter bzw. der Begutachterin beurteilten Arbeit überein.

Neumarkt im Hausruckkreis, 08.07.2022 Verena Tiefenthal

Ort, Datum Unterschrift

Abstract

The global connection of people all over the world via social media networks has not only brought great advantages, but also led to an increase of hate or offensive communication on platforms, such as Instagram or Facebook. The omnipresent occurrence of hate in any culture or language emphasises the urgent need of counter-measurements and regulations for this problem. This is where recent developments in Natural Language Processing and especially pre-trained Transformer models have shown themselves beneficial. Due to the sensitivity of the topic, research in low-resource languages is still at a nascent stage. Thus, this paper focuses on the implementation of methods for detecting German hate speech on social media. An extensive literature research is conducted to determine a good approach for the practical implementation of Hate Speech Detection on a T5 and BERT model. The results of the machine learning models are compared to a human evaluation of the same hate speech data. The overall outcome reveals good, but also divergent performances of the pre-trained Transformer models. Compared to the German models the English ones generate competitive results on the German hate speech data. On the contrary the human evaluation does not keep up with the performance of the neural models. Nevertheless, its findings showed that there is a discrepancy on the definition of hate speech and its individual perception. The research points out significant challenges in data availability and annotation in addition to other obstacles like reliability and ethical issues, which are required to be tackled in future studies to obtain a better generalised approach for Hate Speech Detection.

Kurzfassung

Die globale Vernetzung über soziale Netzwerke von Menschen überall auf der Welt hat nicht nur großartige Vorteile mit sich gebracht, sondern auch dazu geführt, dass Hassrede und beleidigende Äußerungen auf Plattformen, wie Instagram oder Facebook, zunehmen. Die allgegenwärtige Präsenz von Hass in jeglicher Kultur und Sprache hebt die dringliche Notwendigkeit für Gegenmaßnahmen und Regelungen zu diesem Problem hervor. In dieser Hinsicht haben sich die neuesten Entwicklungen im Bereich Natural Language Processing und vor allem vortrainierte Transformer Modelle als sehr nützlich erwiesen. Aufgrund der sensiblen Natur dieses Themas ist die Forschung in nicht so stark forcierten Sprachen noch sehr rückständig. Deshalb befasst sich diese Arbeit mit der Implementierung und Auswertung einer deutschsprachigen Erkennung von Hassrede. Eine systematische Literaturrecherche zum Thema wird durchgeführt, um eine gute Vorgehensweise für die praktische Umsetzung der Hate Speech Detection mit einem T5 und einem BERT Modell zu finden. Die Ergebnisse der verwendeten Machine Learning Ansätze werden mit denen einer menschlichen Beurteilung von Hassrede verglichen. Das allgemeine Resultat der Arbeit zeigt gute, jedoch auch stark schwankende Ergebnisse bei den vortrainierten Transformer Modellen. Beim Testen an deutschen Daten können die englischen Modelle im Vergleich zu den deutschen gut mithalten. Die menschliche Beurteilung hingegen kommt nicht an die Performance der neuronalen Modelle heran. Die Ergebnisse von den Menschen lassen aber erkennen, dass es eine Diskrepanz bei der Definition und individuellen Auffassung von Hassrede gibt. Das Fazit der Recherche hat ergeben, dass es noch wesentliche Herausforderungen in der Datenverfügbarkeit und -annotation gibt, zusätzlich zu Problemen im Bereich Zuverlässigkeit und Ethik. Diese gilt es in angehenden Studien zu adressieren, um einen besseren und generalisierten Ansatz für Hate Speech Detection zu erreichen.

Table of Contents

1 Introduction

The privilege of expressing oneself freely is a valuable human right yet spreading hate or offensiveness is an abuse of this freedom. Owing to the anonymity and distance through online platforms, people tend to feel more empowered to take inappropriate statements. Online communicated hate can spread much faster and potentially cause more harm. The ubiquitous increase of social media usage entails a severe rise of hate speech or offensive language usage on such platforms. The strong motivation to tackle these problems has led researchers to come up with solutions and countermeasures, which are explored in the research area of Hate Speech Detection (HSD) (MacAvaney et al., 2019; Tontodimamma et al., 2020).

Current methods for detecting harmful content are not able to successfully prevent hate postings. Detecting hate speech can be particularly challenging due to imprecise verbalisation of offensiveness or the absence of fixed signal words. That is where Natural Language Processing (NLP) models have proven themselves valuable to better detect offensive language, hate, or fake news (Risch et al., 2019; Yin & Zubiaga, 2021).

This research paper treats the topic of Hate Speech Detection in social media. It comprises a systematic literature research on NLP methods to evaluate the state-of-the-art and introduces commonly used methods for HSD. In addition, this work gives an overview of the current challenges and limitations of this approach. Furthermore, experimental implementations of different methodologies for HSD are conducted and their outcomes are compared with an evaluation done by humans. This thesis is concluded by a brief discussion on the topic and an outlook for future work.

HSD has primarily been investigated for English language, consequently the research for other languages is lacking behind. However, the omnipresent occurrence of hate speech in any language and culture emphasises a necessity of more research for non-English data, as well as a better generalisation on new data (Yin & Zubiaga, 2021). Hence, this paper focuses its empiric part solely on HSD on German language. To achieve this goal, pre-trained as well as already fine-tuned BERT and T5 models are set up for detecting hate on German hate speech data. Within this process, following research questions arise and are answered by the means of this thesis:

- Which common methods are currently applied for detecting hate speech in social media?
- What are the current challenges in the field of HSD, especially when using Transformers?
- How can an existing dataset be enhanced, and which Data Augmentation methods are possible for NLP?
- Which pre-processing methods are commonly used for NLP and how do they affect the results?
- How can HSD be implemented using a German language based pre-trained model?
- How do different neural network models perform, compared to human evaluation?

The purpose of this paper is to investigate how state-of-the-art Transformer models perform in the context of classifying hate speech compared to the results of a detection by humans. Answering these questions provides a better insight into the reliability of HSD by deep neural networks for German social media.

1.1 Motivation

Social media platforms have connected people from all over the world. It facilitates staying in contact with each other, allows to be constantly up-to-date and receive or send messages near real-time. Nevertheless, all the positive effects bring along some obstacles. Amongst these are cyber bullying, fake news and hate speech against all kinds of people. NLP has developed solutions to counter this issues, especially recent technologies, such as pre-trained Transformer models have proven to bring great benefits.

Current NLP models are able to understand one or more languages and with ongoing advancements in their architecture they will understand progressively more peculiarities of a language, such as detailed syntax and semantics. Up to now, this is still a challenge due to the diversity of language systems, cultures, dialect, or slangs. This extensive variety makes it more difficult for language models to comprehend text or speech correctly.

Furthermore, a significant part of the ongoing research is conducted in English language. However, with the interest of addressing these problems on a global perspective, it is essential to get diversification in languages and their cultures.

Hence, this thesis focuses on German HSD to evaluate state-of-the-art models for this language and to promote further research for low-resource languages.

1.2 Method

The first part of this research paper consists of a systematic literature review to gain an overview on the current research. The research is based on sources and papers from scientific online libraries, such as *arXiv.org*, *paperswithcode.com*, *scholar.google.com* and *researchgate.net*. Other resources such as *huggingface.co*, *github.com* or *simpletransformers.ai* provided models, data and help for the code. An overview of the main keywords used for the research is illustrated in Figure 1. The research was conducted from January to June 2022 and its findings are based on the state of the research during this timeframe.

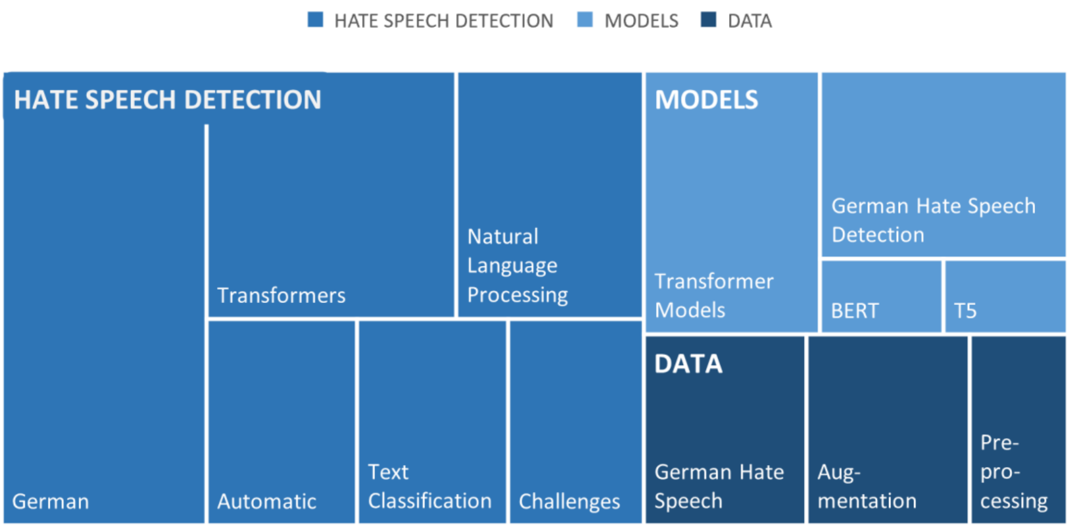


Figure 1: This graphic shows a rough distribution of the keywords in use for the conducted literature research, based on the gathered papers. It serves only demonstrative purposes and does not convey full correctness. Made by author.

The analysis of the papers examines the state-of-the-art in HSD with a focus on approaches in German. It reveals the benefits and opportunities of currently applied models for HSD as well as the challenges within this field.

Subsequently, selected state-of-the-art pre-trained or fine-tuned Transformer models are practically implemented to detect German hate speech. The goal is to fine-tune and test these models on hate speech, using a labelled dataset. Finding suitable datasets, as well as appropriate models is a part of the research process. The results of the models are compared with the results of a human based evaluation on the same test data. The human HSD is performed via an online questionnaire with prior instructions on how the hate speech was annotated to achieve a similar basis as the models. The outcome from the different HSD experiments is put into comparison to determine their differences in performance on German hate speech.

2 Literature Review

This part of the research paper covers the analysis of papers, which were assessed during the literature research. It deals with the definition and the relevance of NLP, investigates its fields of application and common approaches for NLP networks in general, as well as for HSD.

2.1 Definition and Relevance

Natural Language Processing is a computational approach of enabling machines to understand or read texts and extract meaning from human languages[1]. The linguistic analysis of NLP systems can take place on several levels, as there are multiple steps of language processing to perceive a human language (Liddy, 2001). A fundamental understanding of NLP is that every linguistic expression carries an immense amount of information. The topic or word choice, the tone of our voices, the specific order of words within a sentence, or the usage of emojis and certain text elements – all these parameters contribute value to an expression. The enormous number of conversations in this world create an unmaintainable amount of such data. This is where ongoing research in NLP has provided revolutionary solutions to not only interpret a text on its topic, but also to understand the meaning behind it[1].

NLP can be split up into three principal areas:

- **Natural Language Processing (NLP):**
  Transforms the text into structured data
- **Natural Language Understanding (NLU):**
  Is responsible for the understanding of the input and its classification.
- **Natural Language Generation (NLG):**
  Generates text, based on the processed data and task.

  All together they form a working NLP network[2], as illustrated in Figure 2. [3]

[1] https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1, retrieved on 19 April 2022

[2] https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696, retrieved on 27 February 2022

[3] https://datasolut.com/natural-language-processing-vs-nlu-vs-nlg-unterschiede-funktionen-und-beispiele/, retrieved on 8 February 2022

NLP *= NLU + NLG*

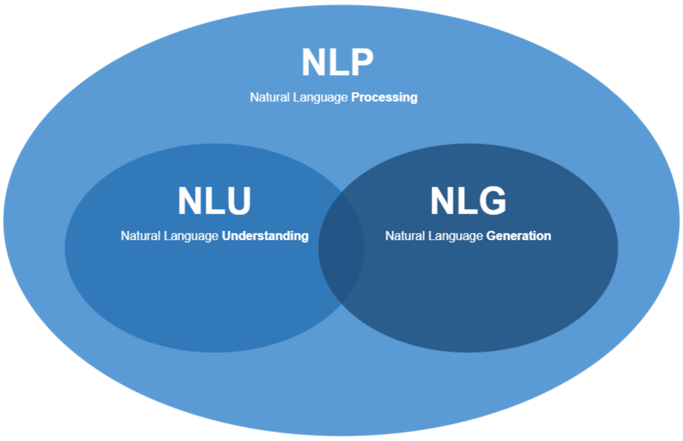

NLP
Natural Language Processing
NLU
Natural Language Understanding
NLG
Natural Language Generation

Figure 2: Correlation between NLP, NLU and NLG. Adapted from datasolut.com[3].

2.2 Fields of Application

While the ability of NLP to understand and handle human language is remarkable, the real value lays in its useful application on many different tasks[1]. NLP machines can compute large volumes of language-based data, consistently and without exhaustion. NLP methods can help to improve the syntactic and semantic understanding in languages and add structure to the data. This permits a broad variety of useful NLP applications[4]. The list below states commonly used fields for NLP:

- **Information Retrieval:**
  Gathering similar information according to an input text from a source text or database (e.g., Google Search[5]).
- **Information Extraction:**
  Extracts related information to a requested subject from a text, for instance documents, e-mails, or short messages (e.g., e-mail filters, social media analytics).
- **Machine Translation:**
  Translates an input from one language to another (e.g., Google Translator[6]).
- **Sentiment Analysis:**
  Analyses the meaning behind a text (e.g., Hate Speech Detection).

[4] https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html, retrieved on 8 February 2022.

[5] https://www.google.at/, retrieved on 8 February 2022.

[6] https://translate.google.com/, retrieved on 8 February 2022.

- **Conversational Systems/Speech Recognition:**
  Enables conversation with a system via voice or text interface (e.g., chatbots).
- **Question Answering:**
  Answers questions to a prior learnt topic.
- **Natural Language Generation:**
  Generates words or sentences from an input image, video, or text (e.g., question generation).
- **Text Analytics:**
  Extracts meaning from the input text (e.g., text summarisation).

(Kalyanathaya et al., 2019; Shetty[7])

2.3 Natural Language Processing Networks

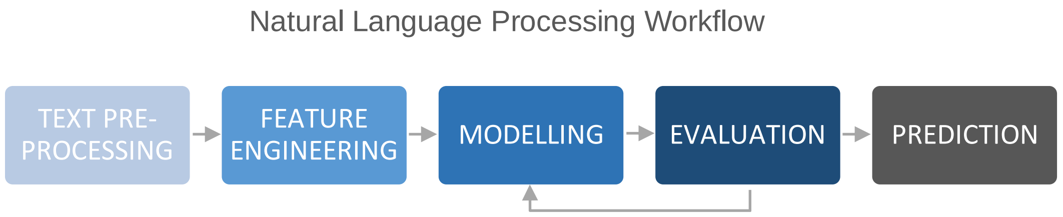


Figure 3: Basic workflow of an NLP Network. Adapted from (Wuttke, 2020)

NLP networks work with a textual dataset (corpus), which is in general a collection of documents or text files (e.g., tweets on Twitter)[8]. This data is prepared with text pre-processing methods and subsequently an algorithm is applied to extract features to train a classifier. The classifier is trained on specific parameters and its respective outcome is evaluated and possibly modified, until the model is suitable for the desired task (see Figure 3).

2.3.1 Text pre-processing

The basic methods of text pre-processing in NLP are briefly explained in the following paragraphs.

[7] https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b, retrieved on 2 April 2022

[8] https://medium.com/geekculture/basics-of-natural-language-processing-for-beginners-d86351df9d09, retrieved on 20 April 2022

2.3.1.1 Tokenisation

Machines do not know how to cut text into phonetic sounds that is why they need rules on how to break the text into smaller sequences called *tokens* in order to process them in future steps. The technique of tokenisation depends on the specific use case. Different languages have diverse semantic rules e.g., some languages do not segment by spaces. Tokenisation can be performed on word-level (see Figure 5), on character-level (see Figure 4), or on subword-level[9]. The subword tokenisation cuts infrequent words into smaller meaningful words (e.g., “boys” is split up in “boy” and “s”)[10].

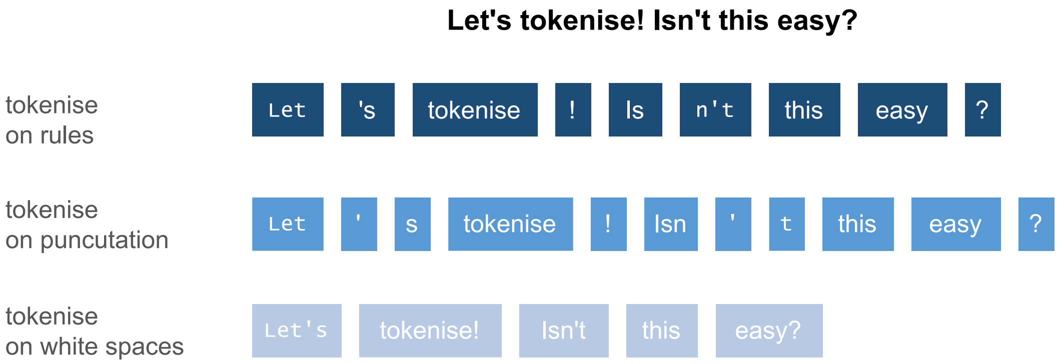

Let's tokenise! Isn't this easy?
tokenise on rules
Let 's tokenise ! Is n't this easy ?
tokenise on puncutation
Let ' s tokenise ! Isn ' t this easy ?
tokenise on white spaces
Let's tokenise! Isn't this easy?

Figure 5: Three ways of word-level tokenisation. Adapted from Cathal[9]

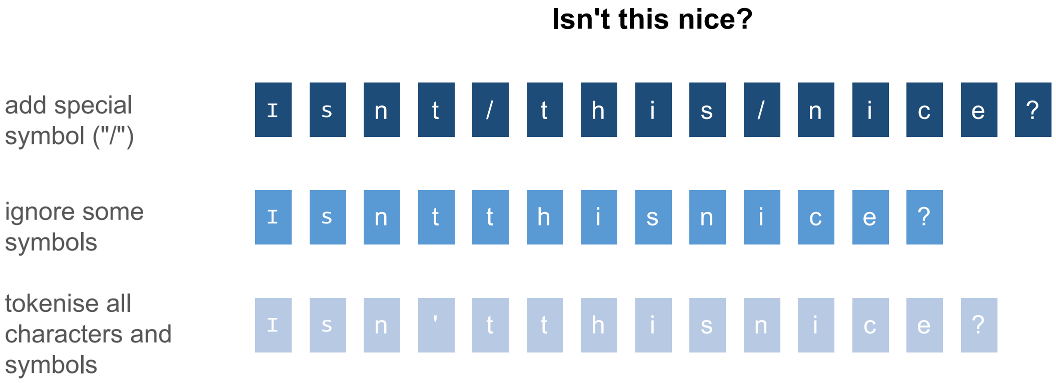

Isn't this nice?
add special symbol ("/")
I s n t / t h i s / n i c e ?
ignore some symbols
I s n t t h i s n i c e ?
tokenise all characters and symbols
I s n ' t t h i s n i c e ?

Figure 4: Three ways of character-level tokenisation. Adapted from Cathal[9]

[9] https://blog.floydhub.com/tokenization-nlp/, retrieved on 7 May 2022

[10] https://towardsdatascience.com/byte-pair-encoding-subword-based-tokenization-algorithm-77828a70bee0, retrieved on 7 May 2022

2.3.1.2 Normalisation

The normalisation process intends to eliminate inflection from texts[8]. Typical normalisation methods are:

- **Stop Words Removal:**
  Removal of stop words from a predefined list (e.g., *and, the, or*).
- **Stemming:**
  Extracts affixes by cutting the beginning or the end of a word (e.g., prefix *eco* of the word *ecosystem*).
- **Lemmatisation:**
  Changes a word to its root form. This homogenises similar words to its base form. (e.g., *best* is changed to *good*).
- **Topic Modelling (Clustering):**
  Clusters text into topics in reference to their contents, by adding values to the words based on their distribution[1].

2.3.2 Traditional Methods

The development of NLP systems began with rule-based models used for simple analysis such as parsing and extraction, followed by statistical models and later neural networks. They can be distinguished between s*yntax analysis* and s*emantic analysis*. Word or sentence structure comprehension is covered by syntax analysis (e.g., word segmentation, part of speech tagging, or parsing). Whereas semantic analysis intends to evaluate the meaning of a message (e.g., named entity recognition, sentiment analysis, or question answering) (Chai & Li, 2019).

A common approach for NLP networks is the Bag of Words (BoW) model. The approach is to count all words in a given text and create an occurrence matrix (a vector) on each word. Thereafter a classifier is trained on the features of the word frequencies. BoW models do not respect word orders or grammar and hence are not able to detect semantic meaning. In addition, high frequency words like "the" add noise to the procedure. The Term Frequency – Inverse Document Frequency (TF-IDF) technique improves this issue but does still not overcome the problem with the loss of semantics[1].

The data-driven approach of NLP relies on statistical and probabilistic computations, which include machine learning algorithms such as Naïve Bayes, k-nearest Neighbors, Hidden Markov Models, Decision Trees, Random Forests, and Support Vector Machines (Otter et al., 2019).

2.3.3 Neural Networks

The linear and sparse inputs of machine learning approaches were dominant in NLP for a long time, until neural models with non-linear and dense inputs emerged and achieved superior results. A neural network consists of layers, at least one input and one output layer. Each layer is responsible for a specific step in the learning process. The layers in turn keep the neurons, which get an input and deliver an output (see Figure 6). These outputs are weighted and generated via non-linear transformation functions. The weights are adapted according to the computed errors or losses of the network. There are various constellations of neural networks available, which vary mainly in the connection of the nodes and the amount of layers (Goldberg, 2016).

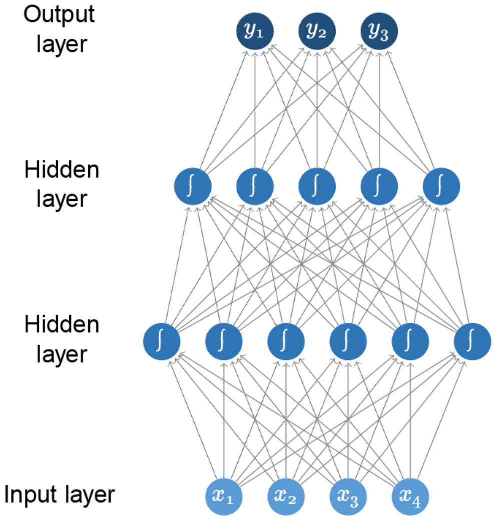


Figure 6: Architecture for a basic neural network with two hidden fully connected layers. The circles represent the neurons, and the arrows are the neuron's inputs and outputs. Adapted from (Goldberg, 2016).

Especially the embedding layer of deep learning models is from particular interest. It maps the features from input words into continuous vectors, while using only low dimensional space. These feature embeddings work as model parameters in training. The transformation into vectors enables the possibility to operate on them, for instance to compute the distance or similarity between words. Further on, the system can also learn to join word vectors and use them for prediction. The training can be done supervised or unsupervised. The unsupervised training is beneficial for feeding the model larger amounts of data, which gives a higher potential to match the requirements of the final task. The selection and adaption of various parameters (e.g., learning rate, batch size or window approach) is also crucial for a suitable network design. (Goldberg, 2016, 2017; Otter et al., 2019).

2.3.3.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) apply non-linear functions know as filters over regions with fragments of text or images. The filter has a n-dimensional vector as output that contains characteristics of the words, and the pooling operation combines all the vectors to a single feature vector (see Figure 7). Networks using convolutional or pooling layers are broadly used in image and video processing, but they also work well for text classification tasks. For classification, the position of the features is not necessarily important but rather whether they appear or not in a certain location, this is where CNNs are beneficial. However, they have the downside of losing most of the structural information when encoding the words into vectors (Chai & Li, 2019; Goldberg, 2016; Otter et al., 2019).

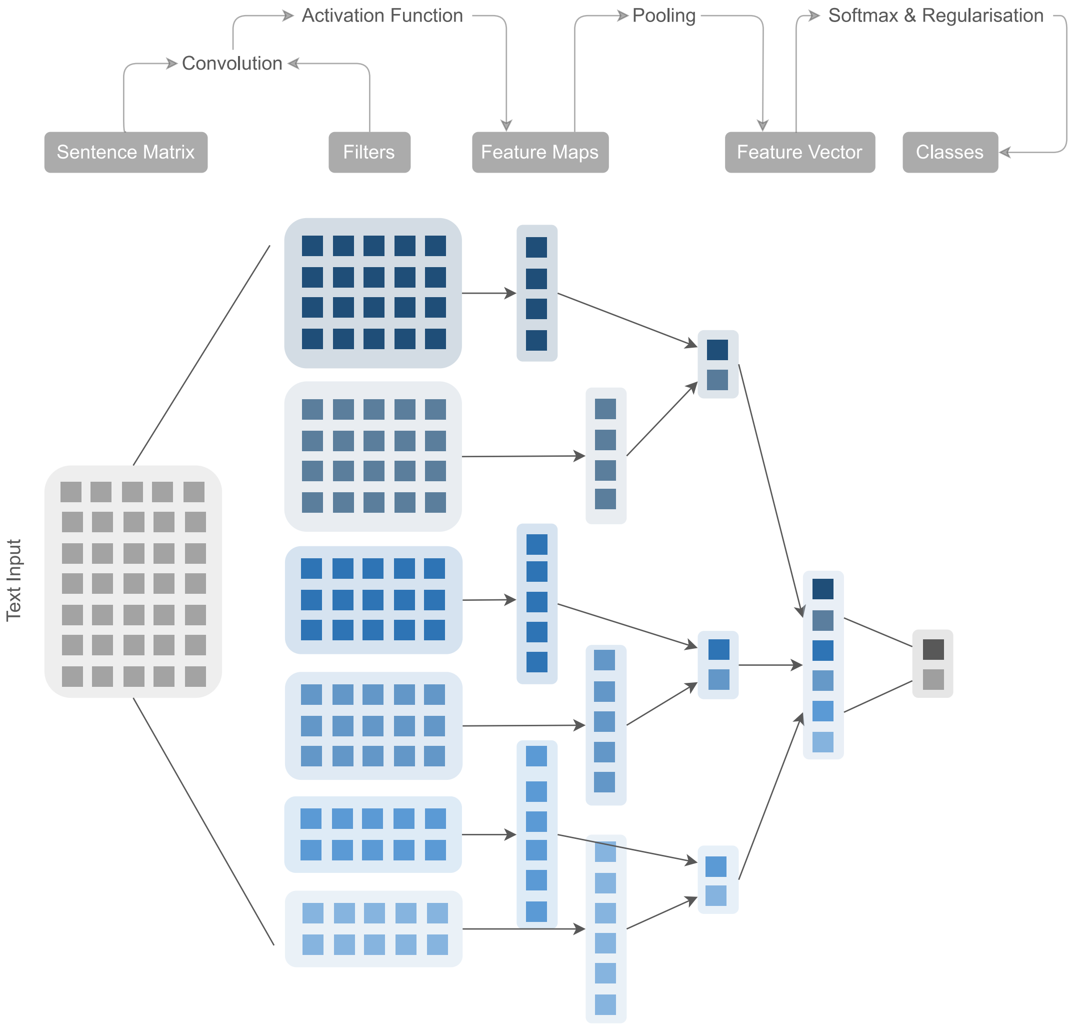


Figure 7: Architecture of a CNN for an NLP task e.g., binary text classification. Adapted from (Chai & Li, 2019).

2.3.3.2 Recurrent/Recursive Neural Network (RNN)

Recurrent and Recursive Neural Networks have the advantage of remembering the order of the input, due to their approach of working with sequences (recurrent) or trees (recursive). Thus, they can provide an understanding for temporal behaviour. The training of an RNN is very labour-intensive. Furthermore, the simple RNNs have a problem with vanishing of gradients over training time, which results in memory loss. That hinders the distinction between older and more recent information (Chai & Li, 2019; Goldberg, 2016).

Long Short-Term Memory Network (LSTM)

To address the memory loss problem Long Short-Term Memory Networks were introduced. LSTMs are based on the architecture of RNNs and use “memory cells” on a gated approach, which means they can decide whether an information is kept or let through (Chai & Li, 2019; Goldberg, 2016). Even though social media texts are not usually very long the advantage of remembering dependencies between sequences allows LSTMs to obtain more contextual semantic information and the ability to drop information improves computational time (Bai, 2018).

Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is based on the gating architecture of an LSTM but uses fewer gates and no memory-cell vectors. That results in more simplicity of the model structure and potentially fewer computational costs, while keeping up with the performance of LSTMs (Goldberg, 2016).

RNN models such as LSTM and GRU have a good performance on long-range context dependencies, as well as text classification tasks. Effective classification is reliable on the understanding of the whole text input. Hence, the best suitable model architecture for text classification depends on whether the conception of long-range semantics is required for the data in use or not (Zulqarnain et al., 2020).

2.3.4 Transformer-based Neural Networks

The encoding of an entire text sequence to a limited length vector, irrespective of the importance of an input, has its downsides. That is where the attention mechanism of Transformer models comes into play (Otter et al., 2019).

In their paper *Attention is All you Need*, Vaswani et al. (2017) initially proposed the Transformer architecture. Transformers are based on the architecture of an Encoder-Decoder structure while using in addition a stacked self-attention and pointwise, fully connected layers (illustrated in Figure 8). The Encoder consists of a stack of layers with corresponding sub-layers, each having a residual connection and a normalisation layer on top. The Decoder on the other hand has an additional third sub-layer on each layer, executing a multi-head attention over the output. The self-attention mechanism is modified by a masking mechanism, preventing the Transformer from “cheating” of subsequent positions. The attention is added by computing weighted sums on key-values of an output. Adding the multi-head attention enables the model to consider information from separate representation subspaces at distinct positions.

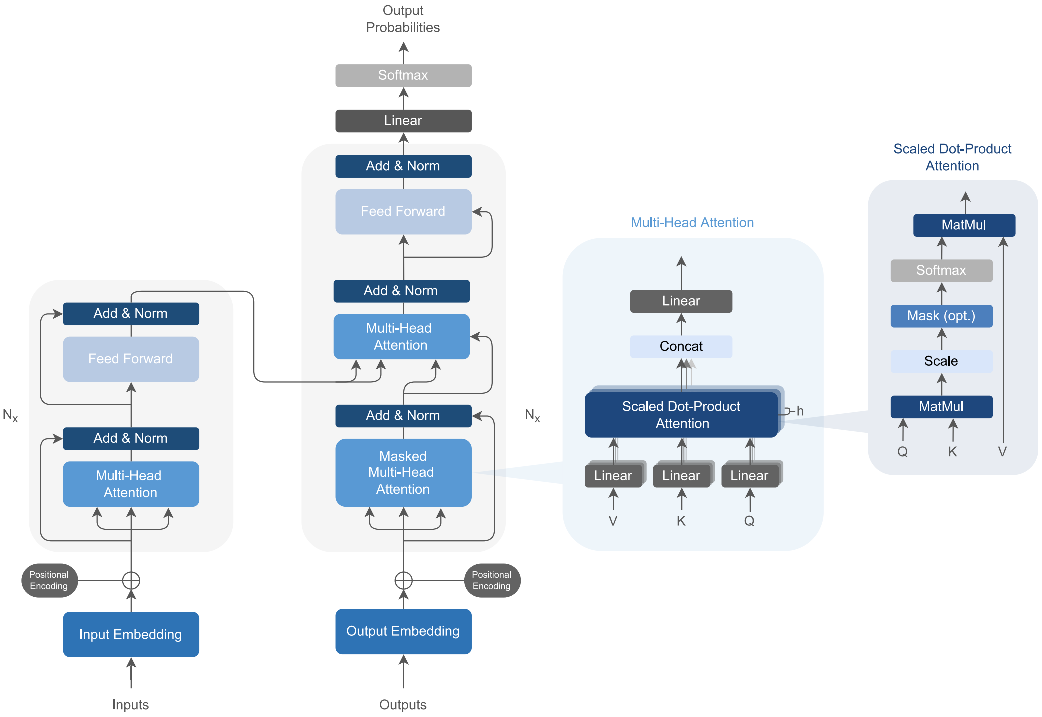


Figure 8: Architecture of the vanilla Transformer model using a “scaled dot-product attention”. Adapted from (Vaswani et al., 2017).

In general, Transformers are trained on a large amount of generic text data, which is called pre-training and can be fine-tuned on specific tasks afterwards while keeping an overall high performance. Numerous variations of the vanilla Transformer architecture have been developed and are still explored. Lin et al. (2021) categorised the pre-trained Transformer models as followed:

- **Encoder only:**
  e.g., BERT (see chapter 2.3.4.1 BERT)
- **Decoder only:**
  e.g., GPT series (Generative Pre-trained Transformer)
- **Encoder-Decoder:**
  e.g., T5 (see chapter 2.3.4.2 T5)

The architectures of BERT and T5 are outlined in the following paragraphs, as they are relevant for the experimental setup of the proposed HSD, covered later in this paper. The GPT models were not selected for the implementation and thus are not treated in more detail.

2.3.4.1 BERT

Devlin et al. (2019) introduced the Bidirectional Encoder Representations from Transformers (BERT) model, which is based on Masked Language Models instead of Unidirectional Language Models and allows bidirectional representations, which means the model learns information in both directions, forwards and backwards. BERT only uses the encoder part of the Transformer architecture. It can be fine-tuned by a single additional output layer. The BERT model comprises two phases, the *pre-training*, where the model is trained on unlabelled data and the *fine-tuning*, where it is initialised with the parameters from pre-training and fine-tuned on labelled data from the specific task.

The initial BERT uses WordPiece tokenisation, which breaks words up into smaller wordpieces, called tokens and can recover the original word from the sequence. The example above shows that some words are broken into wordpieces while others are not. The character “_” is added to tag the beginning of a word.

- **Raw text:**
  Jet makers feud over seat width with big orders at stake
- **Tokenised:**
  _J et_makers _fe ud_over _seat _width_with_big_orders_at_stake

(Wu et al., 2016)

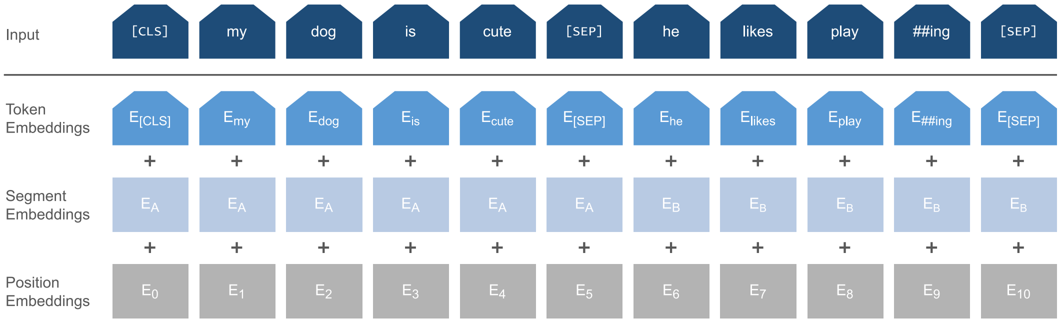


Figure 9: Visualisation of the input representation of a BERT model. The input representations build the sum of token, segment, and positional embeddings. Adapted from (Devlin et al., 2019)

Each input representation from a token is a sum of the equivalent token, segment, and positional embeddings (see Figure 9). The pre-training of BERT is done via two unsupervised tasks, in particular Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). The fine-tuning is depending on the individual downstream task, but it is rather inexpensive on computational resources compared to pre-training. For a classification task the BERT needs the specific input (e.g., text + label) and then all parameters are fine-tuned end-to-end. Subsequently, the `[CLS]` token from the output, which contains the classification embedding, is fed into an output layer for classification, to obtain the prediction (Devlin et al., 2019).

Masked Language Modelling:

Due to the bidirectionality, the BERT model is able to learn information left-to-right and right-to-left. This allows BERT to simply look up the target words. Hence the model would not learn anything new from the prediction. Consequently, the MLM process randomly masks 15% of the tokens, where 80% of these masked tokens are actually replaced with a `[MASK]` token. 10% are replaced with another random token while the remaining 10% are unchanged (Devlin et al., 2019).

Next Sentence Prediction:

Some NLP tasks require an understanding of the relation between two sentences; hence BERT implements NSP to train the model on this matter. Explicitly, this means that for 50% of the sentences of a corpus, the correct next sentence `B` follows the sentence `A`, and for the other 50% a random selected sentence follows the sentence `A` (Devlin et al., 2019).

2.3.4.2 T5

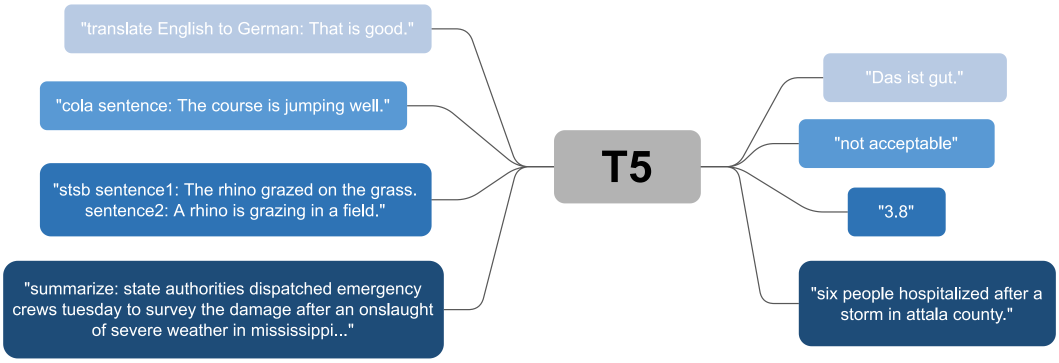

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 10: Text-to-text framework of the T5 model. The model is fed textual input and responds with textual output. Adapted from (Raffel et al., 2020)

The Text-to-Text Transfer Transformer (T5) was proposed by Raffel et al. (2020). It is based on an Encoder-Decoder Transformer architecture with a text-to-text approach, as shown exemplarily in Figure 10. It is trained on unsupervised and supervised tasks on a large common crawl-based data set, the Colossal Clean Crawled Corpus (C4). It comprises about 750 GB of clean, natural English text data. The Encoder and Decoder have a comparable design to the BERT$_{\text{BASE}}$ model. A form of BERT's MLM is applied in the training phase, 15% of random tokens are replaced by a sentinel token (see example in Figure 11). For T5, only consecutive spans are masked, which benefits computational costs.

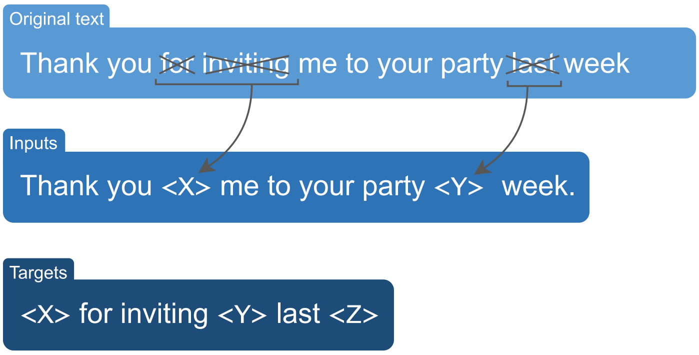

Original text
Thank you ~~for inviting~~ me to your party ~~last~~ week
Inputs
Thank you <X> me to your party <Y> week.
Targets
<X> for inviting <Y> last <Z>

Figure 11: Example of the masking process for the T5 baseline model. Randomly chosen spans are replaced by a unique sentinel token (shown as <X> and <Y>). The output contains the dropped-out spans and a final sentinel token (<Z>) to mark the end. Adapted from (Raffel et al., 2020).

In its original implementation, T5 uses SentencePiece tokenisation, proposed by Kudo & Richardson (2018). It is language-independent and generates subword sequences as followed:

- **Raw Text:**
  Hello_world.
- **Tokenised Text:**
  [Hello] [_wor] [ld] [.]

An advantage of the T5 framework is the reusability of the same model, loss function, and hyperparameters for any NLP assignment (Roberts & Raffel, 2020).

2.4 Traditional Hate Speech Detection

The detection of hate speech is an important NLP task, which has been pursued extensively in the last years, especially for content on online platforms. Often those platforms have a sort of regulation to interdict or prevent offensive language, like banning specific content or profiles. Nevertheless, the execution of these measurements often requires manual intervention. Thus, it is desired to automate such processes (MacAvaney et al., 2019).

2.4.1 Keyword Filtering

Keyword-based approaches rely their filtering for hateful content on an ontology or dictionary which contains possible hateful words. Such databases (e.g., Hatebase[11]) keep a large batch of pejorative expressions against various groups and for different languages. A benefit of keyword-based processes is the easy application and the simple understanding of the technology behind. Yet, the maintenance effort of updating the vocabulary is a severe disadvantage as there is still a lot of manual labour (MacAvaney et al., 2019).

2.4.2 Metadata Extraction

Accessing additional information from users' profiles can help to further understand their postings by means of demographics, location, timestamp, or social engagement. Due to privacy policies this sensitive information is scarcely publicly accessible. This misleads systems to wrong predictions, biased by incidental dataset characteristics. For instance, a demographic biased detection could miss hate postings from users, who do not typically create hateful content. Besides, using sensitive user information raises ethical issues (MacAvaney et al., 2019).

2.4.3 User Reporting

This is a human moderated approach of flagging hate speech. A respective moderator determines the action for the reported content – deletion, suspension, or ignorance. However, the amount to tackle user reporting with manual labour is not manageable. Furthermore, it causes mental stress on the moderator and it is an expensive solution due to personnel costs. (Chaudhary et al., 2021).

2.4.4 Machine Learning Models

The first studies towards basic automatic HSD were conducted by Kwok & Wang (2013), Warner (2012) and Xu & Zhu (2010) (as cited in Hasanuzzaman et al., 2017), which used basic supervised machine learning techniques and sparse datasets. Further on, the range of covered characteristics expanded, demographic

[11] https://hatebase.org/, retrieved on 28 January 2022

or location-based features were integrated, and the models used word embeddings to train the classifiers.

A simple machine learning approach for HSD consists of a data pre-processing phase, where sequences of text are transformed into feature representations. For some datasets it is useful to enhance the data prior to training. Afterwards a classifier is trained on these features. The trained classifier can then be tested against the explicit data to receive the prediction (Lee et al., 2018).

2.5 Neural Hate Speech Detection

The dependence and the workload of feature engineering has steered the research towards neural-based HSD, which have outperformed the traditional methods. (Lee et al., 2018).

The neural detection of hate speech is based on the concept of word embeddings, which means words with a comparable semantic and syntax must be near to each other in an n-dimensional space of embeddings. Vector embeddings, such as Word2Vec and GloVe, or sentence embeddings and Embeddings from Language Models (ELMo) have been used to train classifiers for HSD in recent approaches. Among these neural networks, CNNs, LSTMs, RNNs, GRUs, or a combination of those are commonly used models. Experiments with these kind of networks can be found for example in Deshpande et al. (2022), Dorris et al. (2020), or Geet D'Sa et al. (2020). Transformer-based networks (e.g., BERT) have superseded the feature-based classification method due to their pre-trained model convention (Geet D’Sa et al., 2020).

3 Hate Speech Detection

This chapter explores the definition of hate speech, its challenges, its types, and the state-of-the-art in order to find a good approach for the experimental implementation, which is addressed in chapter 4 Experimental Setup and Results.

3.1 Definition

The interpretation of hate speech can be defined as “any communication that disparages a person or a group on the basis of some characteristics such as race, colour, ethnicity, gender, sexual orientation, nationality, religion” (Tontodimamma et al., 2020, p. 1). The definition is neither globally acknowledged nor are there particular aspects or characteristics which have been reconciled entirely. Agreeing upon a clear definition might facilitate the HSD research and would make it more reliable. Nevertheless, it is not easy to distinguish between offensive and tolerable opinions within a conversation. MacAvaney et al. (2019) summarises a list of definitions for hate speech in his paper. Some of these definitions state that hate is aimed towards a specific group, whilst other definitions, such as the one from the Encyclopedia of the American Constitution[12], determine assault against one individual already as hate speech. Fortuna & Nunes (2019) extracted the following characteristics from various definitions:

- Hate speech is to incite violence or hate
- Hate speech is to attack or diminish
- Hate speech has specific targets
- Humour can be considered hate speech

(Fortuna & Nunes, 2019 as cited in MacAvaney et al., 2019)

3.2 Related Work

A related research area to HSD in the scope of NLP is Fake News Detection, which is also primarily situated on social media platforms and had a massive rise in

[12] “Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity” (Nockleby, 2000, as cited in MacAvaney et al., 2019).

awareness in the last years. Other related areas are the Multiclass (Non-Binary) Detections of hate, comprising the distinguishment between abusiveness, cyberbullying, discrimination, offensiveness, or even more detailed classification of hate. This is also included in the research area of HSD but can be considered in a more separate way. Nevertheless, it is not further outlined in this research paper.

3.3 Challenges and Difficulties

Machine learning techniques have brought up promising solutions to encounter automatic HSD. However, it is a difficulty for humans to interpret why the system has declared something as hateful or not. This is a challenge for automatic censoring processes because they would still need a manual appeal option and hence there would not be a reasonable explanation to the justification of the censoring. Further challenges remain also in the area of dataset availability and construction and existing approaches (MacAvaney et al., 2019).

3.3.1 Data Annotation and Collection

As mentioned before in chapter 3.1 Definition, it is challenging to define specific words or phrases as hateful content. Hence, annotating datasets for HSD is a rather unreliable process. It is dependent on the agreement between annotators, whether a text passage is considered hate speech or not.

Apart from the annotation issues, the access to collected data from social media platforms is limited, due to their strict policies. (MacAvaney et al., 2019). Additionally, data degradation occurs when datasets are published e.g., containing tweet-IDs, which refer to accounts or posts that have been deleted. Consequently, there are few appropriate datasets publicly available. Moreover, multimodal datasets (e.g., posts containing image and text) need further consideration, as image content and memes are common on social media and may also bear hate or information that helps to better detect the hate speech (Madukwe et al., 2020).

3.3.1.1 Class Imbalance

Compared to other text classification tasks, HSD suffers from substantial class imbalances. Madukwe et al. (2020) made this evaluation based on 17 reviewed hate speech datasets. It is, however, a possibility to try to evade this problem in the collection process, but the question is whether it is better to generate balanced datasets or develop models, which cope well with fewer sample size. Applying

Data Augmentation methods to counter the imbalance bears the risk of adding bias to the dataset.

3.3.1.2 Bias

The problem of bias exists due to unavoidable subjective annotation of datasets and training models on “norms”, where society’s biases are reflected in the datasets. There are a lot of kinds of biases, which cause problems (e.g., topic, author, or racial bias). It is important to pay attention to such possible biases when collecting a dataset or “de-bias” existing data using Data Augmentation methods (Yin & Zubiaga, 2021).

3.3.2 Data Pre-processing

Data from social media posts is rather noisy. Pre-processing steps have been introduced to clean the data beforehand. A problem is that these pre-processing steps may affect the data size and thus have consequences on an objective comparison between studies. Additionally, it is difficult to determine whether certain pre-processing is beneficial for the model training or not.

An improvement for this problem might be the introduction of an unbiased publicly available benchmark dataset with defined pre-processing methods (Madukwe et al., 2020). There have been attempts to establish such datasets e.g., HateXplain from Mathew et al. (2020), nevertheless the development is still at a nascent stage.

3.3.3 Language

The prior mentioned scarcity of appropriate datasets is even more troublesome for certain low-resource languages. Therefore, cross-lingual generalisation for HSD is to be encouraged. A solution for this can be Data Augmentation (see chapter 3.5.3 Data Augmentation Methods), cross-lingual approaches, or applying multi-lingual models (Feng et al., 2021; Yin & Zubiaga, 2021).

Another problem are grammar and vocabulary issues on datasets. On social media platforms it is likely that alternative spelling, dialects, or emojis occur. It can be improved by using character or sentence-level features. Implicit expressed hate or irony is another challenging problem, which needs further investigation (Yin & Zubiaga, 2021).

3.3.4 Societal Concerns

Online toxicity is not solely a technological issue, but also a social one. People feel more freedom to express their opinion online than they would offline and may take more severe actions. This phenomenon is called online disinhibition effect. More information can be found in Suler (2004). Other contributors to the polarisation of hate speech are misinformation and/or situational factors, such as bad mood, anger, aggressions, or dissatisfaction.

Consequences of hate speech comprise:

- **Individual harm:**
  Negative psychological effects such as depression, anxiety, and drug abuse, towards suicidal tendencies.
- **Collective harm:**
  Leading to radicalisation and polarisation, which may terminate in organised hate speech against minorities.
- **Societal harm:**
  Growing polarisation may direct to division of the society and potentially culminate in hate crime violence, which brings up security, health care and legal consequences.

(Chaudhary et al., 2021)

3.3.5 Ethical Concerns

A pressing matter when countering hate speech is the dilemma of preventing hateful content whilst keeping the right of freedom of speech. Kiritchenko et al. (2021) provides eight principles to take into consideration for the system design of an ethical HSD.

3.3.6 Legal Concerns

As hard as it is to find an appropriate generalised definition for hate speech, as arduous is it to execute justified penalties when legal action is necessary, due to the intangible nature of offensive language. Besides, hate speech attacks often reach an international spread, which may involve more than one justice system and raises the problem of “who is to penalise”, also considering anonymous user profiles (Chaudhary et al., 2021).

3.3.7 Summary

Researchers are encouraged to put in more effort to obtain less biased and more consistent methods for HSD. Moreover, social media platforms are ought to give better access on their data for research purposes. There is a need of clear regulations concerning ethnicity, legality, and the definition of hate speech (Madukwe et al., 2020). This aims the goal of a more generalisable HSD.

3.4 Types of Hate Speech

Hate speech can be distinguished between *explicit hate*, which is identified by fixed signal words which are collected in dictionaries and *implicit hate*, which is more arduously recognised, as it requires a semantic analysis (Geet D’Sa et al., 2020). Implicit hate speech can be expressed through stereotypes, sarcasm, irony, humour, and metaphors (Yin & Zubiaga, 2021).

- Examples of hate speech: Explicit hate speech:

  “She looks like a tranny.”

  “You Asian, they will deport you when they see your eyes.“

  “We hate niggers, we hate faggots and we hate spics.“

- Implicit hate speech:

  “Affirmative action means we get affirmatively second rate doctors and other professionals.”

  “I will remove all your organs in alphabetical order.”

  “She looks like a plastic monkey doll!”

(Geet D’Sa et al., 2020)

3.5 State-of-the-art Hate Speech Detection

Deshpande et al. (2022) conducted a recent study targeting multilingual HSD. They chose binary classification, language agnostic embeddings and a large dataset from eleven languages for a better generalisation and outperformed prior multilingual approaches. Their chosen evaluation metric was the weighted F1-score. The tested models in their study were (1) LASER embeddings plus a logistic regression model, (2) MUSE embeddings plus a CNN-GRU model and (3) a multilingual BERT (mBERT) model. On their first test “Monolingual-Train Monolingual-Test”, MUSE + CNN-GRU outperformed the other models significantly. The second test “Multilingual-Train Monolingual-Test” showed, that mBERT performs better, when trained on a larger dataset by achieving similar good results as MUSE + CNN-GRU. Overall, the training on multilingual data enhanced performance for all given models.

Xue et al. (2021) carried out an analysis between an mT5 and their self-developed ByT5 model. ByT5 is based on mT5 but uses byte sequences instead of word- or subword-level token-sequences. The proposed ByT5 model can compete with the mT5 model in English as well as on a multilingual basis, especially at smaller data sizes.

An experiment on the HASOC dataset (Mandl et al., 2020) was conducted by Roy et al. (2021) with XLM-RoBERTa (XLMR) on a multilingual basis. They experimented with several different approaches. The one using semantic embeddings as features and a customised two-layer FFNN as classifier received the best results, with the evaluation metric being the macro F1-score.

Caselli et al. (2021) stated that the BERT model may perform well on numerous NLP tasks, but the performance decreases when using more everyday- or slang-language. Thus, they created HateBERT, a re-trained $\text{BERT}_{\text{BASE}}$ model. They evaluated the model with several English datasets and achieved superior results to the standard BERT each time. The metrics in use were precision and recall.

A GPT-3 model is trained on HSD in the paper of Roy et al. (2021) with settings from zero- to few-shot learning, where the accuracy increases. The authors mention a high rate of misclassification with the GPT model. A T5 model does achieve better results at the current state. However, large language models tend to need proper settings and curated examples to identify correctly.

3.5.1 Pre-training vs Fine-tuning

The advantage of pre-trained models is the possibility of transfer learning. This means models are pre-trained on a more general task, where a high quantity of data is available and are respectively fine-tuned on a downstream task, where data availability is limited. It supports state-of-the-art model reusability, benefits computational costs and also the carbon footprint (Kovács et al., 2021).

The feature-based approach is another way of training Transformer models. Peters et al. (2019) stated feature extraction as a good alternative to fine-tuning, depending on the similarity between the pre-training and the target task.

Bigoulaeva et al. (2022) investigated the possibility of cross-lingual transfer learning and received good results. They suggested to put further research in using disparate labelled datasets on the same model to achieve better generalisation.

3.5.2 Suitable Datasets

Several datasets which are available in German language and suitable for a binary classification have been collected and are listed in Table 1.

It must be mentioned that English datasets are accessible in larger sizes, which is a benefit for fine-tuning. Nevertheless, this paper is laid out for a German HSD and thus relies on the datasets available for this language.

Another possibility would be to train or fine-tune models with cross-lingual transfer learning. This would allow the usage of English datasets. This paper, however, does not experiment with cross-lingual approaches, although relevant results can be compared to those of Bigoulaeva et al. (2022) as far as possible in the given setting.

Further information of the chosen dataset is provided in chapter 4.1.1 Dataset.

Table 1: Recent datasets[13] suitable for German HSD

Dataset	Tasks	Size	Source & Reference
Offending Statements / Hate Speech towards Foreigners	Multiclass	5,836	http://www.ub-web.de/research/ (Bretschneider & Peters, 2017)
German Abusive Language Dataset with Focus on COVID-19	Binary	4,960	https://github.com/mawic/german-abusive-language-covid-19 (Wich et al., 2021)
Moderator- and Crowd-Annotated German News Comment Datasets	Binary, Multiclass	85,000	https://zenodo.org/record/5291339#.YmELeNpByBJ (Assenmacher et al., 2021)
HASOC 2020	Binary, Multiclass	4,669	https://hasocfire.github.io/hasoc/2020/dataset.html (Mandl et al., 2020)
German Hatespeech Refugees	Binary	469	https://github.com/UCSM-DUE/IWG_hatespeech_public (Ross et al., 2016)
GermEval 2018	Binary, Multiclass	8,541	https://github.com/uds-lsv/GermEval-2018-Data (Wiegand et al., 2018)
GermEval 2021	Binary, Multiclass	4,190	https://github.com/germeval2021toxic/SharedTask/tree/main/Data%20Sets (Risch et al., 2021)
Hate Speech Dataset	Binary	45,130	https://github.com/manhecht/hatespeech_bac2 (2022)

[13] Links to datasets, retrieved on 29 May 2022

3.5.3 Data Augmentation Methods

The importance of Data Augmentation (DA) in NLP has grown in the last years, due to further research in low-resource domains and the need of larger datasets for recent model architectures. The goal of DA is to mitigate problems, such as bias, class imbalance, or data scarcity. The quantity of data is often an essential parameter for the quality of training, although it does not always result in better learning. A challenging matter is DA for transfer learning, because large pre-trained models are invariant to certain data transformations. The general aim is to create new linguistic patterns the models have not seen before. For HSD and diverse other classification tasks it is important to preserve the initial labels during the transformation. DA is not only beneficial to build better models, but also favours the competitive position against global players (e.g., Google, Amazon, Apple, etc.), who do have access to large amounts of data and consequently take leading positions in these domains. Furthermore, DA can reduce the high manual labour of labelling datasets when the amount of needed data decreases. However, currently there is still a dissonance on *why* certain DA work and *how good* they perform. Another possibility for future research would be to build more robust models against class imbalances, or biases (Bayer et al., 2021; Feng et al., 2021).

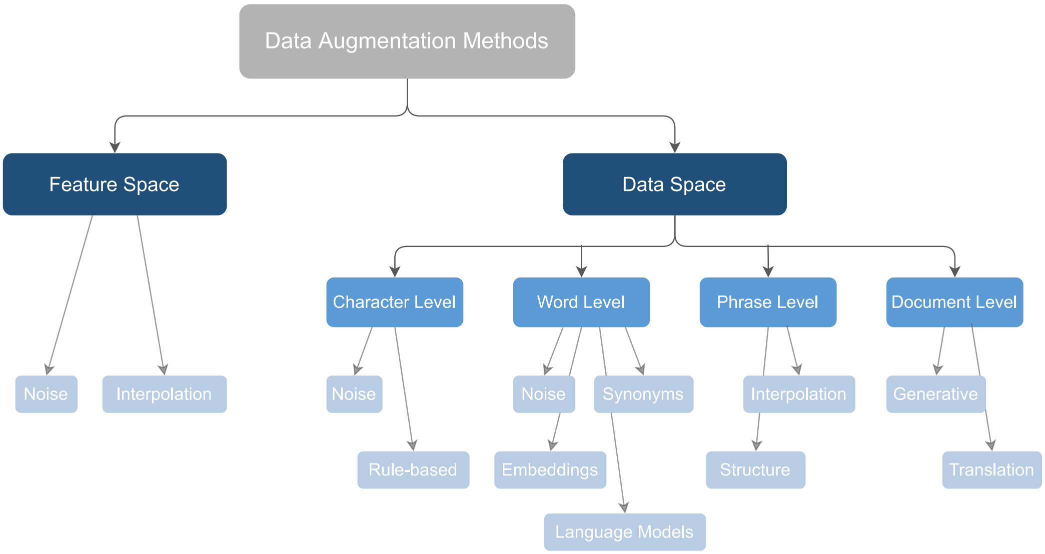


Figure 12: Categorisation of Data Augmentation methods for text classification. Adapted from (Bayer et al., 2021)

DA in the data space transforms the input data. It can be distinguished between an augmentation on character, word, phrase, or document level. Whereas DA in the feature space transforms the feature representations of the input data (see Figure 12). More details on the individual methods are provided in Bayer et al. (2021). They state adversarial training, token/feature/span cut-off, interpolation, and generative methods as promising approaches for pre-trained language models. Thus far there is little research on that matter available.

The DA methods in use for the practical implementation of this paper are outlined in section 4.1.4 Applied Data Augmentation.

3.6 Discussion of Recent Findings

As mentioned in 3.3 Challenges and Difficulties it is rather intricate to compare diverse HSD studies, their models, approaches, or datasets with each other and receive one streamlined knowledge out of it. Thus, it is not possible to conclude with a single solution for a good HSD approach in German language. This paper experiments with the obtained knowledge from previous studies and the results contribute to a further understanding on the application of pre-trained models in low-resource languages. The findings revealed a good overall performance of Transformer models for HSD tasks, but also other neural networks achieved compatible results. However, it is necessary to address the current challenges and concerns in the area of HSD, especially legal and ethical ones, to not drive research further without taking them into account. It is required to create more global and generalisable HSD methods and common approaches and datasets need to be enforced.

4 Experimental Setup and Results

This part of the research paper is dedicated to the empirical research of implementing a German HSD. It comprises a description of used resources and methods, as well as an analysis of the experiment's results, followed by a brief discussion on the experiment.

4.1 Chosen Models and Dataset

The basic setup of the HSD experiment is a comparison between a BERT model, a T5 model and a human assessment. They are all assessed with the same dataset. Several changes on the constellation of the setup are examined as well. These include testing already fine-tuned models, enhancing the train data, and experimenting with English language models on the German dataset.

The models in use are all publicly available on Huggingface[14]. If available, the models were chosen in their uncased variants, as correct lower- and upper-case spelling is not always accurate in social media postings. General information on the T5 and BERT architecture is outlined in chapters *2.3.4.1 BERT* and *2.3.4.2 T5*.

4.1.1 Setup 1 – monolingual

The first setup compares small versions of pre-trained monolingual T5 and BERT models. As monolingual T5 models are only available in English, a respective English BERT model was chosen. For further comparability, a German BERT model is fine-tuned on the data as well.

- **Model 1:** bert-base-german-uncased[15] (German)
- **Model 2:** bert-base-uncased[16] (English)
- **Model 3:** t5-small[17] (English)

[14] https://huggingface.co/docs/transformers/index, retrieved on 13 June 2022

[15] https://huggingface.co/dbmdz/bert-base-german-uncased, retrieved on 9 June 2022

[16] https://huggingface.co/bert-base-uncased, retrieved 9 June 2022

[17] https://huggingface.co/t5-small, retrieved on 9 June 2022

4.1.2 Setup 2 – fine-tuned

This setup evaluates already fine-tuned HSD models. As a first step the fine-tuned models are tested with the test datasets without further ado. Subsequently, they are fine-tuned on the training dataset and tested likewise, to measure if the additional fine-tuning achieved any improvements. With the computational resources available, it was not possible to additionally fine-tune the byT5 model.

- **Model 1:** gbert-germeval-2021[18] (German fine-tuned)
- **Model 2:** dehatebert-mono-german[19] (German fine-tuned)
- **Model 3:** distilbert-base-german-cased-toxic-comments[20] (German fine-tuned)
- **Model 4:** hateBERT[21] (English fine-tuned)
- **Model 5:** byt5-base-tweet-hate-detection[22] (English fine-tuned)

It is suggested to perform another experimental setup for a comparison between multilingual models to evaluate their performance on German hate speech data. Limits of computational resources did not allow to conduct this experiment in the course of this paper; however, results of other studies can still be put into comparison.

[18] https://huggingface.co/shahrukhx01/gbert-germeval-2021, retrieved on 20 June 2022

[19] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-german, retrieved on 20 June 2022

[20] https://huggingface.co/ml6team/distilbert-base-german-cased-toxic-comments, retrieved on 20 June 2022

[21] https://huggingface.co/GroNLP/hateBERT, retrieved on 9 June 2022

[22] https://huggingface.co/Narrativa/byt5-base-tweet-hate-detection, retrieved on 9 June 2022

4.1.3 Conceptual Design

The general approach of the fine-tuning on hate speech classification is outlined in Figure 13. More details on the pipeline and the whole setup can be found in the GitHub Repository[23] accompanying this implementation.

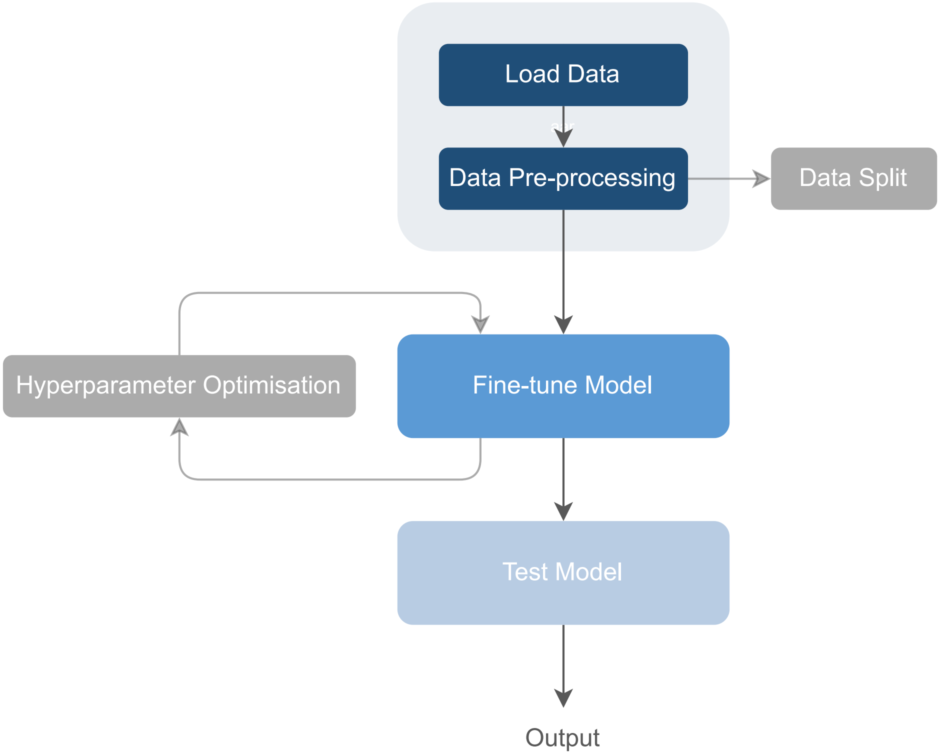


Figure 13: Overall conceptual design of the fine-tuning pipeline for the models. Made by author.

1. **Load Data:**
   The hate speech dataset is loaded into the pipeline.
2. **Data Pre-processing:**
   The dataset is prepared for the model-specific requirements and split up into training, evaluation, and test data.
3. **Fine-tune Model:**
   During the fine-tuning, a hyperparameter optimisation is performed.
4. **Test Model:**
   The models are tested against test datasets and corresponding performance metrics are calculated and logged.

[23] https://github.com/v-tiff/german-hate-speech-detection, retrieved on 25 June 2022

4.1.4 Framework and Tools

The used framework for the fine-tuning on hate speech classification is called simpletransformers[24] and was created by Thilina Rajapakse. It allows an implementation of common NLP tasks and accesses the Transformers library of Huggingface[14]. The code was set up in Python on a Jupyter notebook.

The fine-tuning and testing of the models was carried out via the platform Google Colaboratory[25]. It enables to execute Jupyter notebook services while providing partly free access to computational resources[26]. The processing unit is assigned on availability, but to assure comparability all runs were conducted with a Tesla T4 GPU[27].

Results and metrics of each run were tracked with the platform Weights & Biases (W&B)[28]. It is highly customisable and supports model evaluation, optimisation, and data visualisation.

4.1.1 Dataset

The chosen dataset for the fine-tuning of the models in the following experiments is the Hate Speech Dataset[29] (see also 3.5.2 Suitable Datasets). It is a combination of several already existing datasets for HSD in German language, those are GermEval2018, GermEval2019, Hasoc2019, Hasoc2020, Polly, and HateSpeech_Refugees. The dataset consists of 45,116 postings labelled with `1` for `HATE` and `0` for `NO HATE`. It can be considered slightly imbalanced with 62% hate-labelled postings, however compared to other datasets available it has an appropriate distribution. The author of this dataset offered several pre-processing methods for the dataset, but the outcome of the research showed that best performance was reached with applying no pre-processing at all. Hence, the following experiments apply the dataset with no further elaborated pre-processing.

[24] https://simpletransformers.ai/about/, retrieved on 13 June 2022

[25] https://colab.research.google.com/, retrieved on 13 June 2022

[26] https://research.google.com/colaboratory/faq.html, retrieved on 13 June 2022

[27] https://colab.research.google.com/github/d2l-ai/d2l-tvm-colab/blob/master/chapter_gpu_schedules/arch.ipynb, retrieved on 13 June 2022

[28] https://docs.wandb.ai/, retrieved on 13 June 2022

[29] https://github.com/manhecht/hatespeech_bac2, retrieved on 29 May 2022

4.1.2 Preparations

Some minor data cleaning adjustments are made on the dataset; changing the line breaks to white space and add a “binary classification” prefix to every line, as the T5 models need this prefix for their training. Additionally, for the fine-tuning of the T5 models the labels are converted from Float to String, as they take textual inputs only.

The whole dataset is split into 70% of training data (train_df), and 30% test data, which is again divided into 80% evaluation (eval_df) and 20% test data (test_df). The eval_df is used for a testing at set intervals during the fine-tuning process to check the model’s performance.

For the evaluation of the models a project in W&B was set up to log the data during the fine-tuning and save the models.

4.1.3 Hyperparameter Optimisation

The hyperparameter optimisation takes place during the fine-tuning and is conducted via an hyperparameter grid search. The selected parameters are train epochs, batch size, and learning rate (see Table 2). For each combination of the defined parameters a model is trained. The models are optimised with an AdamW Optimiser configured to optimise on the training loss. Early stopping was implemented, triggered by an MCC (Matthew’s correlation coefficient) score not improving by at least 0.01 upon the best score for five consecutive evaluations[30]. Other hyperparameters were left at the default values provided by the Simple Transformers library.

Table 2: Values of the set hyperparameters for the grid search.

Hyperparameter	Values		
train epochs	2	3	5
batch size	16	16	16
learning rate	$2e^{-5}$	$3e^{-5}$	$5e^{-5}$

4.1.4 Applied Data Augmentation

Ensuing the fine-tuning experiments, two different DA methods on the data space are applied to enhance the training data. The augmented datasets are concatenated to the initial train_df to create a larger training set. The parameter

[30] https://simpletransformers.ai/docs/tips-and-tricks/#using-early-stopping, retrieved 22 June 2022

values from the best performing runs in the prior setups are selected and solely pre-trained T5$_{\text{SMALL}}$ and BERT$_{\text{BASE}}$ are fine-tuned on this new created data sets. To perform the augmentation, the library nlpaug[31] from Edward Ma was used, which provides various DA possibilities for NLP. The two augmentation methods in use were contextual word embeddings and round-trip translation.

4.1.4.1 Contextualised Word Embeddings

A classical method of DA is the exchange of word embeddings. It can be performed via algorithms (e.g., Word2Vec, GloVe, etc.), however these are only comparing the similarity between two vectors. Contextual word embeddings do consider surrounding words as well. The prediction of the target word is conducted with NLP models, in this case a German BERT$_{\text{BASE}}$ [32] was chosen[33].

4.1.4.2 Round-Trip Translation

Round-trip translation (RTT) is the translation of a dataset from the initial language to a target language and backwards to the source language (translating solely backwards is named back-translation). The goal is to augment the training data in the source language by obtaining wording variations due to translations. For this implementation, the chosen target language is English, owing to the higher availability on good performing translation models. For the translation to English the opus-mt-de-en[34] model was used and the opus-mt-en-de[35] executed the backtranslation[36].

The two augmentation functions generated new datasets which were combined with the training dataset. Based on that, three augmented datasets were created for further experiments:

- train_contextual_embedding_df
- train_rt_translation_df
- train_contextual_embedding_rt_translation_df

[31] https://github.com/makcedward/nlpaug, retrieved 22 June 2022

[32] https://huggingface.co/deepset/gbert-base, retrieved on 22 June 2022

[33] https://towardsdatascience.com/data-augmentation-library-for-text-9661736b13ff, retrieved on 22 June 20222

[34] https://huggingface.co/Helsinki-NLP/opus-mt-de-en, retrieved on 22 June 2022

[35] https://huggingface.co/Helsinki-NLP/opus-mt-en-de, retrieved on 22 June 2022

[36] https://towardsdatascience.com/data-augmentation-in-nlp-2801a34dfc28, retrieved on 22 June 2022

4.2 Human Hate Speech Detection

As a reference and to evaluate the accuracy of the neural HSD models a human assessed HSD is put into comparison with the model's results. For the human detection, the participants had to review the data from survey_df (likewise as the models) and are requested to examine it with their judgement on hate speech. The test_df has too many examples for this survey setup. To achieve a better comparability all participants received prior instructions to detect the hate speech to make the results comparable to the ones of the models (see Table 3).

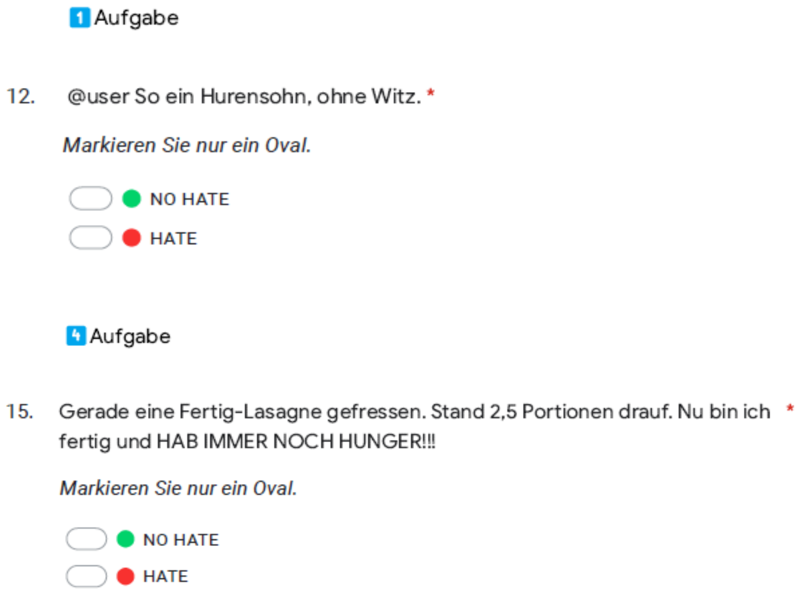
Table 3: Examples from the training data given to the participants prior to their evaluation.

Example	Label
@user Genau die Lügenpresse betreibt Volksverdummung.	🔴 HATE
@user Ich traue Gott das sogar zu, sollte es ihn geben...	🟢 NO HATE

4.2.1 Setting and Used Tools

The survey for the human HSD is conducted via an online questionnaire created with Google Forms[37] (see appendix A Questionnaire for Human HSD). An online setting was chosen to increase the scope of participants as the topic did not demand a personal approach. The questionnaire was distributed via WhatsApp and Slack. A simple and explanatory language was chosen to avoid confusion and increase accessibility for people with less knowledge on the subject. As the paper focuses on German HSD the questionnaire is correspondingly created in German language. It comprises 26 questions, covering a few demographic and social matters and the hate speech examples. As mentioned earlier, the participants received four examples on how to evaluate the sentences on 🔴 HATE or 🟢 NO HATE. Subsequently, they had to assess the 15 examples with their personal judgement on hate speech (see examples Figure 14).

[37] https://www.google.com/intl/de/forms/about/, retrieved on 30 April 2022



1 Aufgabe

12. @user So ein Hurensohn, ohne Witz. *

Markieren Sie nur ein Oval.

○ ● NO HATE

○ ● HATE

4 Aufgabe

15. Gerade eine Fertig-Lasagne gefressen. Stand 2,5 Portionen drauf. Nu bin ich fertig und HAB IMMER NOCH HUNGER!!! *

Markieren Sie nur ein Oval.

○ ● NO HATE

○ ● HATE

Figure 14: Two examples from the survey to be evaluated by the participants. Made by author.

4.2.2 Participants

Due to the online setting a mixed audience responded to the questionnaire, which is good for generalisability. 40 participants responded to the questionnaire and the distribution of the sexes was nearly even. Almost all of them had German as their first language and the majority was between the age of 15 – 29, but it also reached participants over the age of 40 and above. The stated location of the participants is entirely in Austria. The overall educational level is at least “finished apprenticeship” or above. This also fits the result of a quite good understanding on the topic in general as well as the basic understanding of its technology behind. Only five persons indicated to never have heard about hate speech and eight persons do not know anything behind the technology of HSD. More than 90% of the questioned persons use social media with an average usage of two to three hours per day and are primarily on Instagram, Facebook, and Snapchat. All participants stated to have noticed an increase of hate and offensiveness on social media to a certain amount in the last five years. More details on the participants can be found in the appendix B Questionnaire Results.

4.3 Challenges and Limitations

One of the main challenges for this experiment was the comparability or generalisation between datasets, approaches, models, and results of other research. The deficit or absence of benchmarks and common approaches causes difficulties in comparison and particularly, in deciding which approach to use. This is a major issue for all future research. It is required to address these problems near-term to improve the quality of comparison within this research area.

Furthermore, the scarce data availability in German hate speech datasets induces the risk of too less training data for Transformer models and inconsistent annotated datasets, thus a higher chance of creating a bias on the models. The chosen dataset compensates these factors to some extent. Nevertheless, annotation disparities remain, as the results of the survey show.

Access to the required computational resources slightly limited some executions, however, Google Colab served as a great tool for the most part of the planned experiments. Merely the intended comparison between a multilingual BERT and T5 model and the fine-tuning of byT5 were not executable, due to some T5 models being too big for the free Google Colab setup.

Even though the same performance metrics are used for the human and the neural HSD, it must be mentioned that there is not a common basis of knowledge. The models are all fine-tuned on the same training data. Whereas the participants from the survey present a diverse knowledge base on hate speech, due to their experience and education. This fact shows none the less interesting insights, though it hinders an equal comparison.

4.4 Evaluation and Results

The goal of the evaluation is to identify the performance of Transformer models and humans on a German hate speech dataset. The approach of the evaluation and the results of the conducted experiments are treated within this chapter.

4.4.1 Technique and Metrics

W&B was used as tool for the models' evaluation. It provides a good overview of the various runs and visualises the data in interactive and customisable tables. The results of the human detection were manually calculated from the collected data.

The main values for assessing the performance of a prediction are the *actual outcome*, which comes from the ground truth of the dataset, and the *predicted outcome*, which was created by the model or the human. These values were put into a confusion matrix to further compute the performance metrics. The selected metrics for the evaluation of this experiment are accuracy and f1-score. However, it is challenging to choose the right metrics for performance evaluation. Olteanu et al. (2017) did research on abstract evaluation metrics. They allow numeral comparison across domains, nevertheless they do not include insights on the algorithmic decision and human's perception of the algorithm's correctness. He advocated for more human-centred metrics. Although this is an intriguing topic to investigate, it is not applied in this research, but recommended for future studies.

4.4.1.1 Accuracy

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

The accuracy metric measures the part which has been predicted correctly out of all predictions. It ranges between 0 and 1, the nearer to 1 the better the prediction. Accuracy alone is not always a fully revealing metric for potentially imbalanced data and further metrics are needed[38].

[38] https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92, retrieved on 14 June 2022

4.4.1.2 F1-Score

$$f1\text{-}score = \frac{2TP}{(2TP + FP + FNN)} = 2 * \frac{(precision * recall)}{precision + recall}$$

The f1-score indicates the harmonic mean between *precision* (what proportion of positive predictions were correct) and *recall* (what proportion of actual positives was recognised correctly). It also ranges from 0 to 1, with 1 being the highest value[38].

4.4.2 Used Data

W&B logs the results of each iteration (a so called “sweep”) for every composition of the defined hyperparameter grid search. It is individually configurable which metrics are to be logged in the sweeps. For this setup, each sweep logs the calculated accuracy and f1-score on train_df, survey_df, and test_df, and additionally some further evaluation values. This data can be read out directly from the dashboard of W&B or downloaded as a report.

The data from the survey was collected by Google Forms, which provides likewise a graphical overview of the results and a download option for a data table. This data was processed to obtain the values for evaluation metrics.

4.4.3 Results

The results of the hate speech evaluations are outlined in the following chapters and presented in tables. Each run of the hyperparameter grid search is listed in a separate row. The values of the metrics are formatted column-wise like a heat map, the darker the blue the higher is the achieved value.

4.4.3.1 Baseline

As a baseline the English BERT$_{BASE}$ was tested on the survey_df and test_df datasets without any fine-tuning (see Table 4). The English pre-trained model was chosen to create a competitive baseline for all model setups.

Table 4: Results of baseline computation.

Model	Test_acc	Test_f1	Survey_acc	Survey_f1
BERT$_{BASE}$ (English)	0.6014	0.7506	0.5333	0.6957

4.4.3.2 Results Setup 1

Table 5 shows the results of the iterations for setup 1 (see 4.1.1 Setup 1 – monolingual). Overall, the highest values have been achieved by model 1 (German BERT$_{BASE}$). The BERT models showed better results with a learning rate of 2e^{-5} or 3e^{-5}, whereas the T5 model get in average better results with the learning rate of 5e^{-5}. Model 2 (English BERT$_{BASE}$) had the lowest results. It must be noted that, a high score on the test_df does not necessarily entail a good score on the survey_df as well. The low count of 15 examples in survey_df could be considered a factor to these observations. The best model run on the test_df was with the third model, T5$_{SMALL}$ with a test_acc of 0.8755 and a test_f1 of 0.8941. The average best model run however was with model 1 on learning rate 3e^{-5}, achieving the highest results with 2 train epochs.

Table 5: Results of setup 1.

Model	Epochs	LR	Batch	Test_acc	Test_f1	Survey_acc	Survey_f1
BERT$_{BASE}$ (German)	5	5e^{-5}	16	0.8593	0.8789	0.7333	0.7500
	3	5e^{-5}	16	0.8629	0.8814	0.8000	0.8000
	2	5e^{-5}	16	0.8607	0.8774	0.6000	0.6250
	5	3e^{-5}	16	0.8648	0.8848	0.8667	0.8750
	3	3e^{-5}	16	0.8666	0.8850	0.8667	0.8889
	2	3e^{-5}	16	0.8644	0.8827	**0.9333**	**0.9412**
	5	2e^{-5}	16	0.8666	0.8896	0.7333	0.8000
	3	2e^{-5}	16	**0.8722**	**0.8913**	0.7333	0.7778
	2	2e^{-5}	16	0.8703	0.8885	0.8000	0.8235
BERT$_{BASE}$ (English)	5	5e^{-5}	16	0.8460	0.8606	0.5333	0.5333
	3	5e^{-5}	16	0.8463	0.8566	0.6000	0.4000
	2	5e^{-5}	16	0.8467	0.8648	**0.7333**	**0.7143**
	5	3e^{-5}	16	0.8504	0.8712	0.6000	0.6250
	3	3e^{-5}	16	**0.8585**	**0.8785**	0.6000	0.6667
	2	3e^{-5}	16	0.8530	0.8756	0.4667	0.5000
	5	2e^{-5}	16	0.8581	0.8777	0.5333	0.5882
	3	2e^{-5}	16	0.8537	0.8679	0.6667	0.6667
	2	2e^{-5}	16	0.8570	0.8722	0.6667	0.6154
T5$_{SMALL}$ (English)	5	5e^{-5}	16	0.8648	0.8933	0.6667	0.7368
	3	5e^{-5}	16	**0.8774**	0.8933	**0.8000**	**0.8235**
	2	5e^{-5}	16	0.8641	0.8827	0.7333	0.7778
	5	3e^{-5}	16	**0.8755**	**0.8941**	0.6667	0.7368
	3	3e^{-5}	16	0.8633	0.8859	0.7333	0.7778
	2	3e^{-5}	16	0.8489	0.8607	0.7333	0.7143
	5	2e^{-5}	16	0.8703	0.8886	0.7333	0.7778
	3	2e^{-5}	16	0.8456	0.8575	0.7333	0.7500
	2	2e^{-5}	16	0.8415	0.8555	0.7333	0.7500

4.4.3.3 Results Setup 2

Table 6 contains the results of the setup 2 (see 4.1.2 Setup 2 – fine-tuned). It shows, that selected HSD models, which are already fine-tuned do overall not perform as good as the preceding models. Some values are even below the calculated baseline. The best overall performance was achieved by the byT5 model, although it was only fine-tuned on English hate speech. The gBERT and the hateBERT (also English) performed on a similar basis.

Table 6: Results of fine-tuned HSD models.

Model	Test_acc	Test_f1	Survey_acc	Survey_f1
gBERT (German)	0.5172	0.4431	0.8000	0.8000
dehateBERT (German)	0.4016	0.0380	0.5333	0.2222
distilBERT (German)	0.4477	0.1984	0.8667	0.8750
hateBERT (English)	**0.6025**	**0.7520**	0.5333	0.6957
byT5 (English)	0.4667	0.6000	**0.8992**	**0.9469**

As a next step all these models were additionally fine-tuned on the train_df with the same hyperparameter grid as in setup 1. In general, all model performances were enhanced by further fine-tuning- (see Table 7). The best results were achieved with model 1 (gBERT) and model 3 (distilBERT). The distilBERT in particular achieved an exceptional increase in performance. For unknown reasons, the dehateBERT did not show significant improvements.

Table 7: Results of additional fine-tuned HSD models.

Model	Epochs	LR	Batch	Test_acc	Test_f1	Survey_acc	Survey_f1
gBERT (German)	5	$5e^{-5}$	16	0.9073	0.9227	0.8667	0.8750
	3	$5e^{-5}$	16	0.9069	0.9214	0.8667	0.8889
	2	$5e^{-5}$	16	**0.9121**	0.9253	**0.9333**	**0.9412**
	5	$3e^{-5}$	16	0.9028	0.9163	**0.9333**	0.9333
	3	$3e^{-5}$	16	0.9058	0.9205	0.8000	0.8235
	2	$3e^{-5}$	16	0.9095	0.9232	0.8000	0.8235
	5	$2e^{-5}$	16	0.9088	0.9222	**0.9333**	0.9333
	3	$2e^{-5}$	16	0.9091	0.9234	0.8667	0.8750
	2	$2e^{-5}$	16	0.9117	**0.9256**	0.8667	0.8750
dehateBERT (German)	5	$5e^{-5}$	16	**0.7333**	**0.7778**	0.8482	0.8718
	3	$5e^{-5}$	16	0.4667	0.4286	0.8629	0.8789
	2	$5e^{-5}$	16	0.6000	0.5714	0.8596	0.8759
	5	$3e^{-5}$	16	0.4667	0.5556	**0.8655**	**0.8863**
	3	$3e^{-5}$	16	0.4667	0.5000	0.8618	0.8804
	2	$3e^{-5}$	16	0.6667	0.6154	0.8648	0.8812
	5	$2e^{-5}$	16	0.6667	0.6667	0.8629	0.8804
	3	$2e^{-5}$	16	0.6667	0.6667	0.8600	0.8784
	2	$2e^{-5}$	16	0.6000	0.5714	0.8618	0.8788

distilBERT (German)	5	$5e^{-5}$	16	0.8966	0.9132	0.8667	0.8750
	3	$5e^{-5}$	16	0.9036	0.9185	**0.9333**	**0.9412**
	2	$5e^{-5}$	16	0.9110	0.9244	0.8667	0.8889
	5	$3e^{-5}$	16	0.9051	0.9211	0.8000	0.8421
	3	$3e^{-5}$	16	0.9080	0.9231	0.8000	0.8421
	2	$3e^{-5}$	16	0.9132	0.9268	0.8667	0.8889
	5	$2e^{-5}$	16	0.9040	0.9191	0.8667	0.8889
	3	$2e^{-5}$	16	**0.9124**	0.9252	**0.9333**	**0.9412**
	2	$2e^{-5}$	16	0.9117	**0.9253**	**0.9333**	**0.9412**
hateBERT (English)	5	$5e^{-5}$	16	0.8593	0.8789	0.7333	0.7500
	3	$5e^{-5}$	16	0.8629	0.8814	0.8000	0.8000
	2	$5e^{-5}$	16	0.8607	0.8774	0.6000	0.6250
	5	$3e^{-5}$	16	0.8648	0.8848	0.8667	0.8750
	3	$3e^{-5}$	16	0.8666	0.8850	0.8667	0.8889
	2	$3e^{-5}$	16	0.8644	0.8827	**0.9333**	**0.9412**
	5	$2e^{-5}$	16	0.8666	0.8896	0.7333	0.8000
	3	$2e^{-5}$	16	**0.8722**	**0.8913**	0.7333	0.7778
	2	$2e^{-5}$	16	0.8703	0.8885	0.8000	0.8235

The additional fine-tuning on the byT5 was not executed, because of the computational limits of Google Colab. It is still recommended to do further experiments with the byT5 or different T5 models, as they produce promising results.

4.4.3.4 Results Human Hate Speech Detection

The results of the human HSD were processed with some functions (see appendix C Code Listing 1) and put into a confusion matrix (see Table 8) to calculate the accuracy and f1-score of the outcome. The average accuracy of 0.783 and average f1-score of 0.695 do not reach the performance of the HSD models. Only four people scored 13 correct predictions out of the 15 examples, in contrast, five people predicted merely 8 right guesses or below.

Table 8: Confusion matrix of survey results.

		Actual Values	
		HATE	NO HATE
Predicted Values	HATE	217	72
	NO HATE	103	208

Obtaining these results with a human detection shows a great disparity in hate speech perception. Especially when looking at single performances of individual participants there is a large discrepancy. Some examples were even assessed with 50% `HATE`🔴 and 50% `NO HATE`🟢. These results point out the challenges of annotation issues and especially the disagreement on the definition of hate in general. The detailed results of the human HSD are outlined in the appendix B Questionnaire Results from Question 11 to 25.

4.4.3.5 Results with Data Augmentation

The experiments with Data Augmentation were conducted with the best run of German BERT$_{\text{BASE}}$ (3 train epochs, 16 batch size, 2e^{-5} learning rate) and of T5$_{\text{SMALL}}$ (3 train epochs, 16 batch size, 5e^{-5} learning rate) on test_df. Both models were fine-tuned on three augmented datasets (see 4.1.4 Applied Data Augmentation). The results of this process are shown in Table 9 as well as the values of the previous run without augmentation.

Table 9: Results of fine-tuning with augmented datasets

Model	Test_acc	Test_f1	Survey_acc	Survey_f1
No Augmentation				
BERT$_{\text{BASE}}$ (German)	0.8722	0.8913	0.7333	0.7778
T5$_{\text{SMALL}}$ (English)	**0.8774**	**0.8933**	**0.8000**	**0.8235**
Contextual Word Embedding Augmentation				
BERT$_{\text{BASE}}$ (German)	0.8678	0.8868	**0.8000**	**0.8235**
T5$_{\text{SMALL}}$ (English)	0.8814	0.9008	0.6667	0.7368
Round-Trip Translation Augmentation				
BERT$_{\text{BASE}}$ (German)	0.8618	0.8826	0.8000	0.8421
T5$_{\text{SMALL}}$ (English)	0.8840	0.9028	0.6667	0.7368
Contextual Word Embedding & Round-Trip Translation Augmentation				
BERT$_{\text{BASE}}$ (German)	0.8726	0.8983	0.7333	0.8000
T5$_{\text{SMALL}}$ (English)	**0.8929**	**0.9103**	0.6667	0.7368

The outcome of the DA shows no significant increase in the performance of the models. The German BERT$_{\text{BASE}}$ hardly improved on the augmented fine-tuning. The T5$_{\text{SMALL}}$ had increased results on the test_df, although it performed poor on the survey_df. The analysis of this outcome shows a rather bad performance of the augmentation methods. The reasons for this might be problems with noise on the data, a shift of the essential meaning of a sentence, or the generation of not fully linguistically correct sentences after the augmentation. All these factors contribute to a drop of quality in the data. Another problem might be, that Transformer models are already rather good in understanding linguistic patterns and these slight augmentations do not change much for their hate speech understanding. Bayer et al. (2021) indicated similar observations in their paper and suggest more elaborated transformations (e.g., on feature space, or generative models) to gain higher improvement on large pre-trained models. Due to the survey_df being a small dataset the outcomes of the test_df are more decisive. The highest improvement was achieved with the combination of contextual word embedding & round-trip augmentation. This leads to the assumption that more training data helps the model improve even if the data is not the best quality.

There is always a risk in DA methods relying on other language models, as their language understanding cannot be proven to be alike. Additionally, the possible incorrectness of “social media language” increases errors in the translation. All these uncertainties need to be addressed to obtain better results with DA, especially for Transformer models.

4.5 Discussion of the Experiments

The results of the HSD experiments showed that there is no significant difference on the performance between state-of-the-art BERT and T5 models, the same counts for the hyperparameter settings. However, T5 models are not widely investigated for HSD, yet still had promising outcomes in this implementation and need to be taken into further consideration. The human HSD on the contrary shows a scattering and was not competitive to the models. This undermines the ongoing challenges in defining hate speech and its subjective perception. On account of the human based detection two test datasets were required, with the survey_df being a small dataset. This resulted in disparities on the survey metrics. A larger test dataset gives the model a better chance to predict more correctly.

This paper focused on the application of smaller models, since they require less processing time. It should be investigated if the potential faster training of larger models with the right parameter settings can even further reduce computational costs, as this should also be of relevance for not only HSD, but for all deep neural network implementations. According to this exploration the most essential key for a good detection is the dataset with respect to its size, quality, and annotation. Especially the additional fine-tuning on hate speech models proved that more training data increases correct predictions.

The textual DA methods applied on this implementation did not indicate a good improvement on the model performance. This claims for more advanced approaches and methods, as the problems of (hate speech) data scarcity will still exist in the future. Thus, the goal is to enhance the available data and focus on better imbalance- and bias-invariant model architectures.

5 Conclusion and Future Work

This research paper gives an overview on methods and the availability of state-of-the-art German HSD as well as suitable datasets. It provides a framework for fine-tuning BERT and T5 models on hate speech data with the aim of comparing it to a detection by humans. While the performance of the models was inconsistent, they still achieved appropriate results. The outcome of the human HSD was not able to compete with the models. Nevertheless, it revealed discrepancies on the overall hate speech consensus and proved the ongoing challenge of hate speech definition and annotation.

The fine-tuning was executed on downsized forms of the models with limited computational resources. Thus, it is recommended to extend this research on a broader scope with larger model sizes. Particularly a multilingual setup might imply high potential, as the English trained models already showed good performance. Further research with large pre-trained models also depends on more hate speech data. This is where DA needs more focus for future studies. The experiments with DA in this paper showed common textual methods do not greatly enhance the data. This might be led back to potential errors of the used models in the DA or further to pre-trained language models demanding more complex data transformations.

The listed challenges of this research area like annotation disparities, data issues, inconsistent definition, or challenging comparability do need a higher focus in the future. Especially, the divergent results of the human HSD and a human's individual perception of hate speech raise sensitive and ethical questions, which need more attention in the further progress of HSD research.

A major problem is the generalisability of data processing, model training and evaluation. It complicates comparability and reusability of studies and approaches. Moreover, a higher generalisable model architecture potentially enables more robust detection.

The stated difficulties point out that the research has still high potential of development. The practical implementation is further ahead as the theoretical understanding and explainability of why some approaches provide better results than others. This raises ethical and legal issues for an automated HSD, especially when they get increasingly more responsibility.

For future studies it is recommended to extend the research on more low-resource languages and target a higher generalisation. As this HSD was only binary, a multiclass, or even multimodal HSD should be further addressed, especially for German and other low-resource languages. Annotation discrepancies between datasets make it difficult to simply add more data to the model training, hence there is a need for a method on how to annotate datasets on a common basis. For this to achieve, there is a requirement of a universal and cultural independent definition of hate speech, which should be prioritised.

References

Assenmacher, D., Niemann, M., Müller, K., Seiler, M. V., Riehle, D. M., & Trautmann, H. (2021). *RP-Mod &amp; RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets* (Version v2) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.5242915

Bai, X. (2018). Text classification based on LSTM and attention. *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 29–32. https://doi.org/10.1109/ICDIM.2018.8847061

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2021). *A Survey on Data Augmentation for Text Classification*. 42. https://doi.org/10.48550/arXiv.2107.03158

Bigoulaeva, I., Hangya, V., Gurevych, I., & Fraser, A. (2022). Addressing the Challenges of Cross-Lingual Hate Speech Detection. *ArXiv:2201.05922 [Cs]*. http://arxiv.org/abs/2201.05922

Bretschneider, U., & Peters, R. (2017). *Detecting Offensive Statements towards Foreigners in Social Media*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2017.268

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. *ArXiv:2010.12472 [Cs]*. http://arxiv.org/abs/2010.12472

Chai, J., & Li, A. (2019). Deep Learning in Natural Language Processing: A State-of-the-Art Survey. *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 1–6. https://doi.org/10.1109/ICMLC48188.2019.8949185

Chaudhary, M., Saxena, C., & Meng, H. (2021). Countering Online Hate Speech: An NLP Perspective. *ArXiv:2109.02941 [Cs]*. http://arxiv.org/abs/2109.02941

Chiu, K.-L., & Alexander, R. (2021). Detecting Hate Speech with GPT-3. *ArXiv:2103.12407 [Cs]*. http://arxiv.org/abs/2103.12407

Deshpande, N., Farris, N., & Kumar, V. (2022). *Highly Generalizable Models for Multilingual Hate Speech Detection*. https://arxiv.org/abs/2201.11294v1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dorris, W., Hu, R. (Roger), Vishwamitra, N., Luo, F., & Costello, M. (2020). Towards Automatic Detection and Explanation of Hate Speech and Offensive Language. *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, 23–29. https://doi.org/10.1145/3375708.3380312

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *ArXiv:2105.03075 [Cs]*. http://arxiv.org/abs/2105.03075

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Geet D’Sa, A., Illina, I., & Fohr, D. (2020). Classification of Hate Speech Using Deep Neural Networks. *Revue d’Information Scientifique & Technique*, *25*(01). https://hal.archives-ouvertes.fr/hal-03101938

Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, *57*, 345–420. https://doi.org/10.1613/jair.4992

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

Hasanuzzaman, M., Dias, G., & Way, A. (2017). *Demographic Word Embeddings for Racism Detection on Twitter*. 11.

Kalyanathaya, K. P., Akila, D., & Rajesh, P. (2019). *Advances in Natural Language Processing – A Survey of Current Research Trends, Development Tools and Industry Applications*. *7*(5), 3.

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *Journal of Artificial Intelligence Research*, *71*, 431–478. https://doi.org/10.1613/jair.1.12590

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of Hate Speech Detection in Social Media. *SN Computer Science*, *2*(2), 95. https://doi.org/10.1007/s42979-021-00457-3

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *ArXiv:1808.06226 [Cs]*. http://arxiv.org/abs/1808.06226

Kwok, I., & Wang, Y. (2013). *Locate the Hate: Detecting Tweets against Blacks*. 2.

Lee, Y., Yoon, S., & Jung, K. (2018). Comparative Studies of Detecting Abusive Language on Twitter. *ArXiv:1808.10245 [Cs]*. http://arxiv.org/abs/1808.10245

Liddy, E. D. (2001). Natural Language Processing. *Syracuse University*, 15.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A Survey of Transformers. *ArXiv:2106.04554 [Cs]*. http://arxiv.org/abs/2106.04554

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Madukwe, K., Gao, X., & Xue, B. (2020). In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 150–161. https://doi.org/10.18653/v1/2020.alw-1.18

Mandl, T., Modha, S., Shahi, G. K., Jaiswal, A. K., Nandini, D., Patel, D., Majumder, P., & Schäfer, J. (2020). *Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages*. 25.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *ArXiv:2012.10289 [Cs]*. http://arxiv.org/abs/2012.10289

Olteanu, A., Talamadupula, K., & Varshney, K. R. (2017). The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. *Proceedings of the 2017 ACM on Web Science Conference*, 405–406. https://doi.org/10.1145/3091478.3098871

Otter, D. W., Medina, J. R., & Kalita, J. K. (2019). A Survey of the Usages of Deep Learning in Natural Language Processing. *ArXiv:1807.10854 [Cs]*. http://arxiv.org/abs/1807.10854

Peters, M. E., Ruder, S., & Smith, N. A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *ArXiv:1903.05987 [Cs]*. http://arxiv.org/abs/1903.05987

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv:1910.10683 [Cs, Stat]*. http://arxiv.org/abs/1910.10683

Risch, J., Stoll, A., Wilms, L., & Wiegand, M. (2021). *Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*. 113.

Roberts, A., & Raffel, C. (2020, February 24). Exploring Transfer Learning with T5: The Text-To-Text Transfer Transformer. *Google AI Blog*. Retrieved on 19 April 2022 from http://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *ArXiv:1701.08118 [Cs]*. https://doi.org/10.17185/duepublico/42132

Roy, S. G., Narayan, U., Raha, T., Abid, Z., & Varma, V. (2021). Leveraging Multilingual Transformers for Hate Speech Detection. *ArXiv:2101.03207 [Cs]*. http://arxiv.org/abs/2101.03207

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2020). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157–179. https://doi.org/10.1007/s11192-020-03737-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Warner, W. (2012). *Detecting Hate on the World Wide Web*. 6.

Wich, M., Räther, S., & Groh, G. (2021). German Abusive Language Dataset with Focus on COVID-19. *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 247–252. https://aclanthology.org/2021.konvens-1.26

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). *Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language*. 11.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv:1609.08144 [Cs]*. http://arxiv.org/abs/1609.08144

Xu, Z., & Zhu, S. (2010). *Filtering Offensive Language in Online Communities using Grammatical Relations*. 10.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2021). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *ArXiv:2105.13626 [Cs]*. http://arxiv.org/abs/2105.13626

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598. https://doi.org/10.7717/peerj-cs.598

Zulqarnain, M., Ghazali, R., Mazwin, Y., & Rehan, M. (2020). A comparative review on deep learning models for text classification. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*. https://doi.org/10.11591/ijeecs.v19.i1.pp325-335

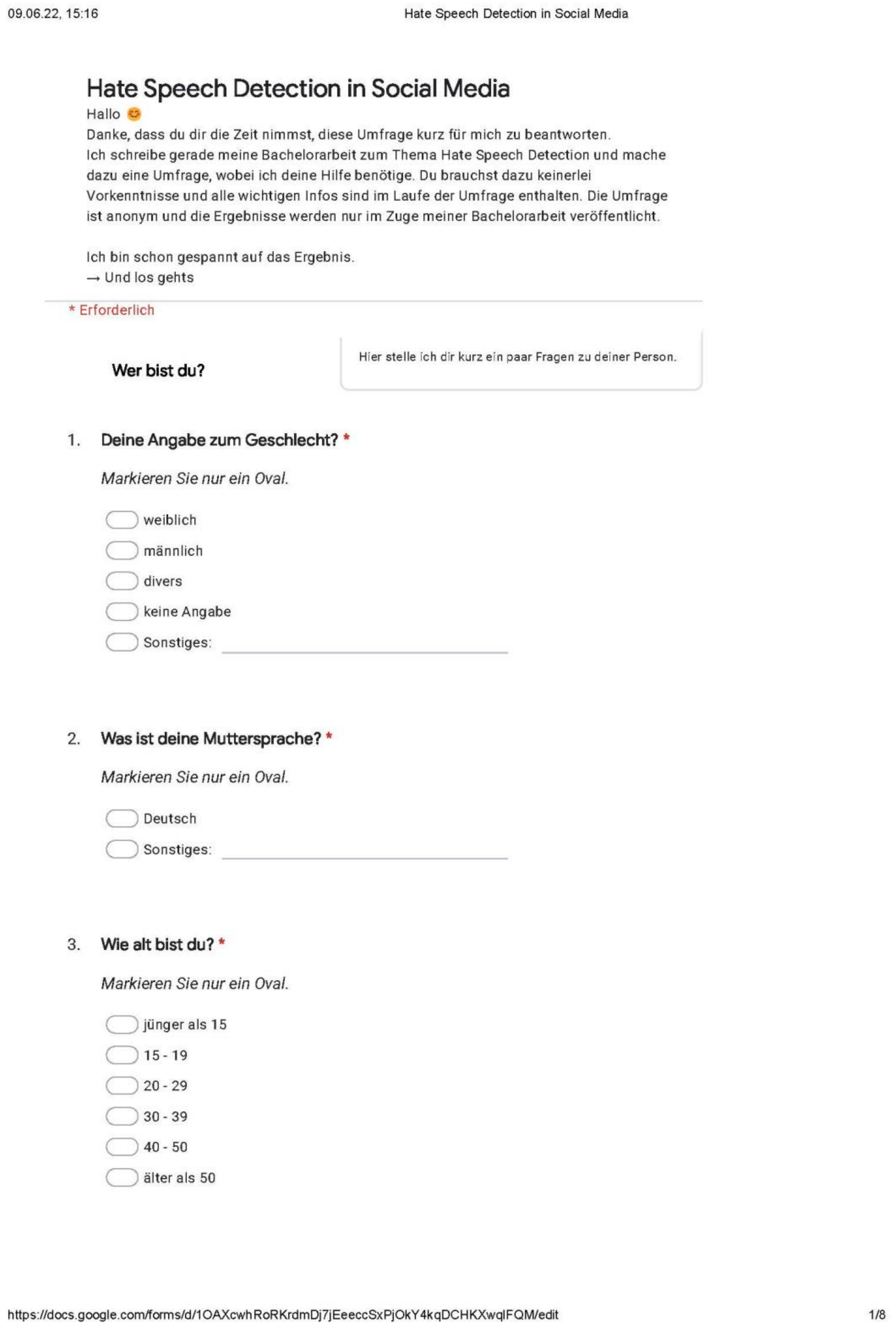
List of Figures

List of Tables

List of Listings

Appendix

A. Questionnaire for Human HSD



Hate Speech Detection in Social Media

Hallo 😊

Danke, dass du dir die Zeit nimmst, diese Umfrage kurz für mich zu beantworten.
Ich schreibe gerade meine Bachelorarbeit zum Thema Hate Speech Detection und mache dazu eine Umfrage, wobei ich deine Hilfe benötige. Du brauchst dazu keinerlei Vorkenntnisse und alle wichtigen Infos sind im Laufe der Umfrage enthalten. Die Umfrage ist anonym und die Ergebnisse werden nur im Zuge meiner Bachelorarbeit veröffentlicht.

Ich bin schon gespannt auf das Ergebnis.
→ Und los gehts

* Erforderlich

Wer bist du?

Hier stelle ich dir kurz ein paar Fragen zu deiner Person.

1. **Deine Angabe zum Geschlecht?** *

Markieren Sie nur ein Oval.

○ weiblich
○ männlich
○ divers
○ keine Angabe
○ Sonstiges: ____________

2. **Was ist deine Muttersprache?** *

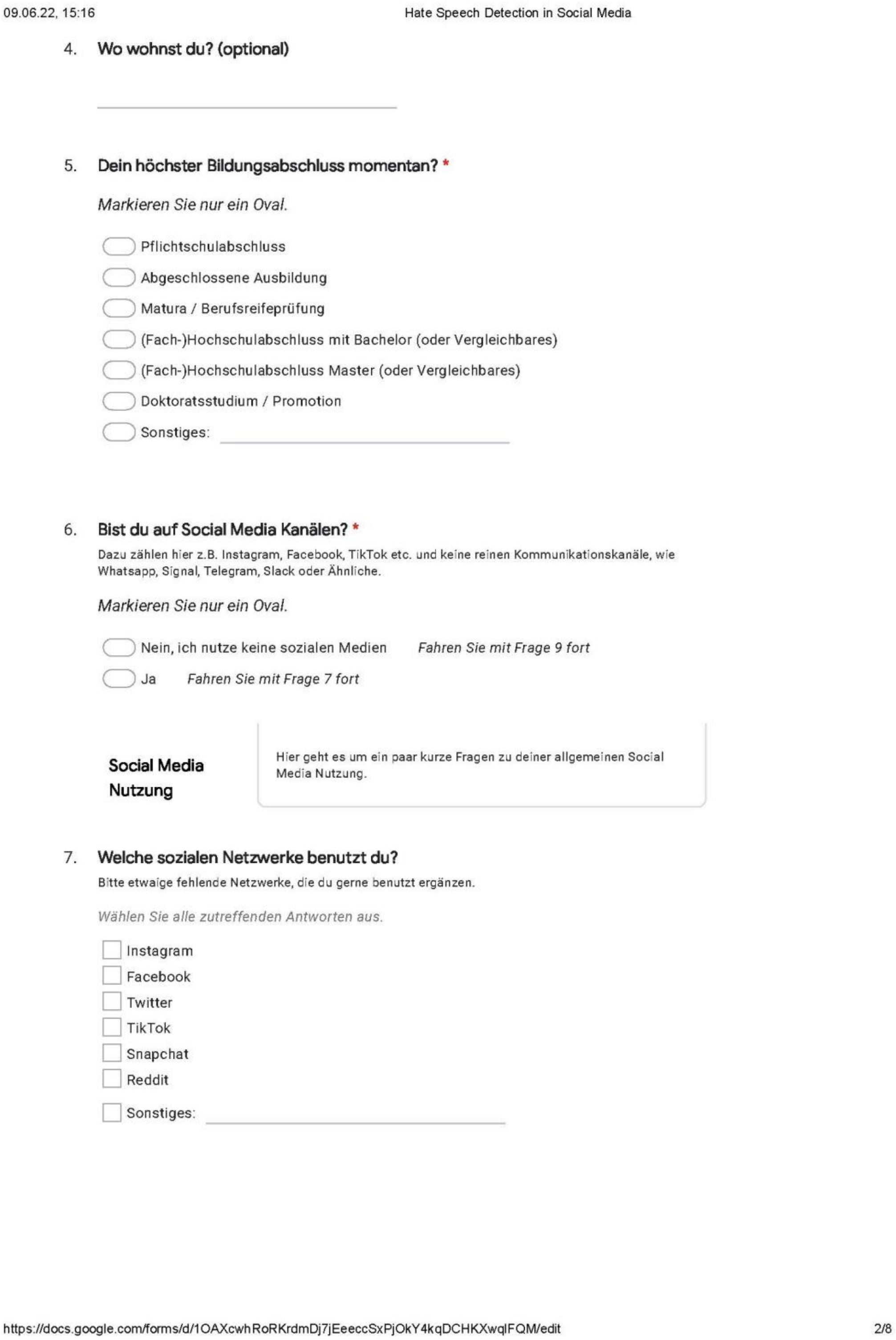
Markieren Sie nur ein Oval.

○ Deutsch
○ Sonstiges: ____________

3. **Wie alt bist du?** *

Markieren Sie nur ein Oval.

○ jünger als 15
○ 15 - 19
○ 20 - 29
○ 30 - 39
○ 40 - 50
○ älter als 50



4. **Wo wohnst du? (optional)**

5. **Dein höchster Bildungsabschluss momentan? ***

Markieren Sie nur ein Oval.

○ Pflichtschulabschluss
○ Abgeschlossene Ausbildung
○ Matura / Berufsreifeprüfung
○ (Fach-)Hochschulabschluss mit Bachelor (oder Vergleichbares)
○ (Fach-)Hochschulabschluss Master (oder Vergleichbares)
○ Doktoratsstudium / Promotion
○ Sonstiges:

6. **Bist du auf Social Media Kanälen? ***

Dazu zählen hier z.B. Instagram, Facebook, TikTok etc. und keine reinen Kommunikationskanäle, wie Whatsapp, Signal, Telegram, Slack oder Ähnliche.

Markieren Sie nur ein Oval.

○ Nein, ich nutze keine sozialen Medien *Fahren Sie mit Frage 9 fort*
○ Ja *Fahren Sie mit Frage 7 fort*

Social Media Nutzung

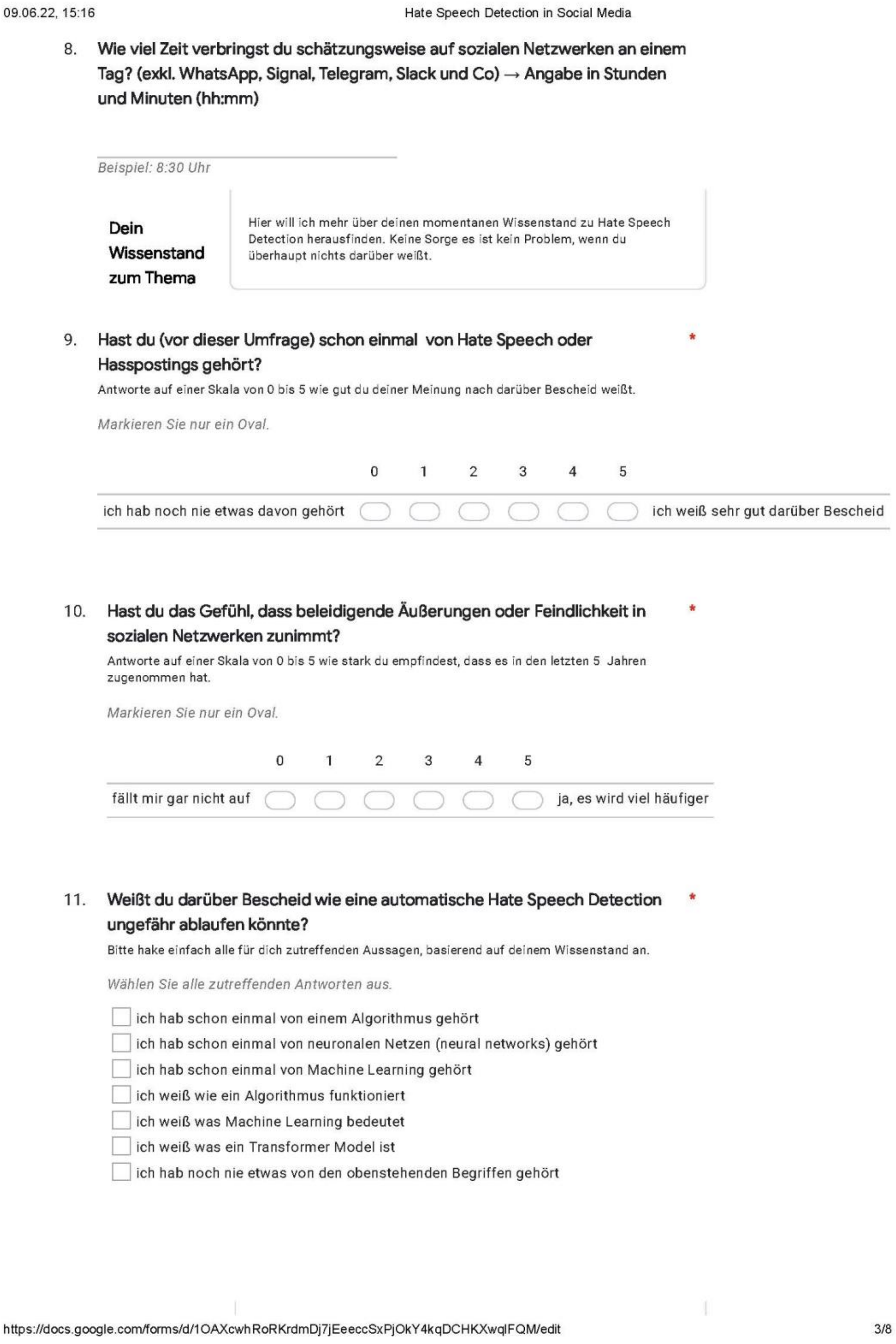
Hier geht es um ein paar kurze Fragen zu deiner allgemeinen Social Media Nutzung.

7. **Welche sozialen Netzwerke benutzt du?**

Bitte etwaige fehlende Netzwerke, die du gerne benutzt ergänzen.

Wählen Sie alle zutreffenden Antworten aus.

□ Instagram
□ Facebook
□ Twitter
□ TikTok
□ Snapchat
□ Reddit
□ Sonstiges:



8. **Wie viel Zeit verbringst du schätzungsweise auf sozialen Netzwerken an einem Tag? (exkl. WhatsApp, Signal, Telegram, Slack und Co) → Angabe in Stunden und Minuten (hh:mm)**

____

Beispiel: 8:30 Uhr

Dein Wissenstand zum Thema

Hier will ich mehr über deinen momentanen Wissenstand zu Hate Speech Detection herausfinden. Keine Sorge es ist kein Problem, wenn du überhaupt nichts darüber weißt.

9. **Hast du (vor dieser Umfrage) schon einmal von Hate Speech oder Hasspostings gehört?** *

Antworte auf einer Skala von 0 bis 5 wie gut du deiner Meinung nach darüber Bescheid weißt.

Markieren Sie nur ein Oval.

	0	1	2	3	4	5	
ich hab noch nie etwas davon gehört	○	○	○	○	○	○	ich weiß sehr gut darüber Bescheid

10. **Hast du das Gefühl, dass beleidigende Äußerungen oder Feindlichkeit in sozialen Netzwerken zunimmt?** *

Antworte auf einer Skala von 0 bis 5 wie stark du empfindest, dass es in den letzten 5 Jahren zugenommen hat.

Markieren Sie nur ein Oval.

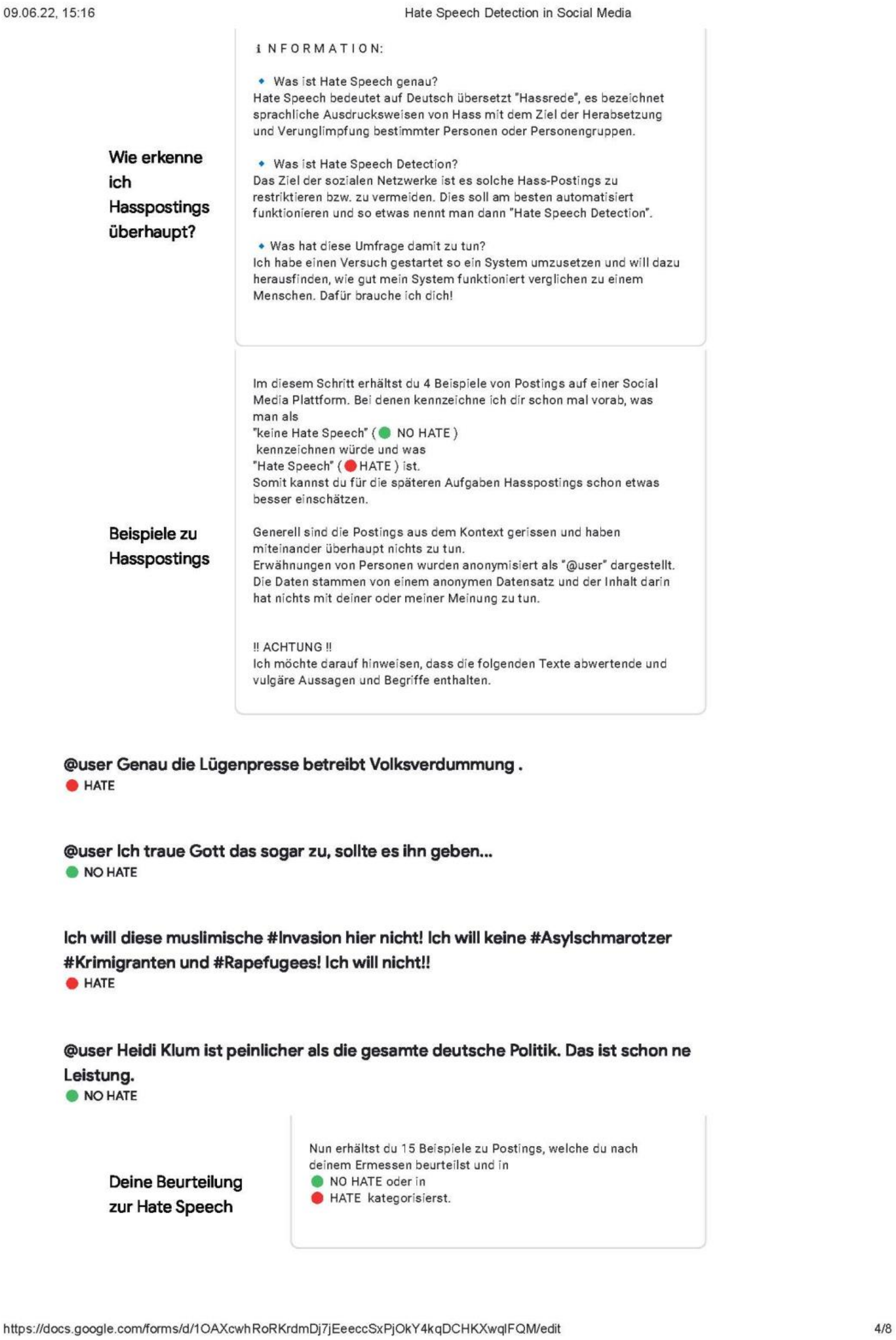
	0	1	2	3	4	5	
fällt mir gar nicht auf	○	○	○	○	○	○	ja, es wird viel häufiger

11. **Weißt du darüber Bescheid wie eine automatische Hate Speech Detection ungefähr ablaufen könnte?** *

Bitte hake einfach alle für dich zutreffenden Aussagen, basierend auf deinem Wissenstand an.

Wählen Sie alle zutreffenden Antworten aus.

□ ich hab schon einmal von einem Algorithmus gehört
□ ich hab schon einmal von neuronalen Netzen (neural networks) gehört
□ ich hab schon einmal von Machine Learning gehört
□ ich weiß wie ein Algorithmus funktioniert
□ ich weiß was Machine Learning bedeutet
□ ich weiß was ein Transformer Model ist
□ ich hab noch nie etwas von den obenstehenden Begriffen gehört



09.06.22, 15:16 Hate Speech Detection in Social Media

Wie erkenne ich Hasspostings überhaupt?

ℹ INFORMATION:

🔹 Was ist Hate Speech genau?
Hate Speech bedeutet auf Deutsch übersetzt "Hassrede", es bezeichnet sprachliche Ausdrucksweisen von Hass mit dem Ziel der Herabsetzung und Verunglimpfung bestimmter Personen oder Personengruppen.

🔹 Was ist Hate Speech Detection?
Das Ziel der sozialen Netzwerke ist es solche Hass-Postings zu restriktieren bzw. zu vermeiden. Dies soll am besten automatisiert funktionieren und so etwas nennt man dann "Hate Speech Detection".

🔹 Was hat diese Umfrage damit zu tun?
Ich habe einen Versuch gestartet so ein System umzusetzen und will dazu herausfinden, wie gut mein System funktioniert verglichen zu einem Menschen. Dafür brauche ich dich!

Beispiele zu Hasspostings

Im diesem Schritt erhältst du 4 Beispiele von Postings auf einer Social Media Plattform. Bei denen kennzeichne ich dir schon mal vorab, was man als
"keine Hate Speech" (🟢 NO HATE)
kennzeichnen würde und was
"Hate Speech" (🔴HATE) ist.
Somit kannst du für die späteren Aufgaben Hasspostings schon etwas besser einschätzen.

Generell sind die Postings aus dem Kontext gerissen und haben miteinander überhaupt nichts zu tun.
Erwähnungen von Personen wurden anonymisiert als "@user" dargestellt.
Die Daten stammen von einem anonymen Datensatz und der Inhalt darin hat nichts mit deiner oder meiner Meinung zu tun.

‼ ACHTUNG ‼
Ich möchte darauf hinweisen, dass die folgenden Texte abwertende und vulgäre Aussagen und Begriffe enthalten.

@user Genau die Lügenpresse betreibt Volksverdummung .
🔴 HATE

@user Ich traue Gott das sogar zu, sollte es ihn geben...
🟢 NO HATE

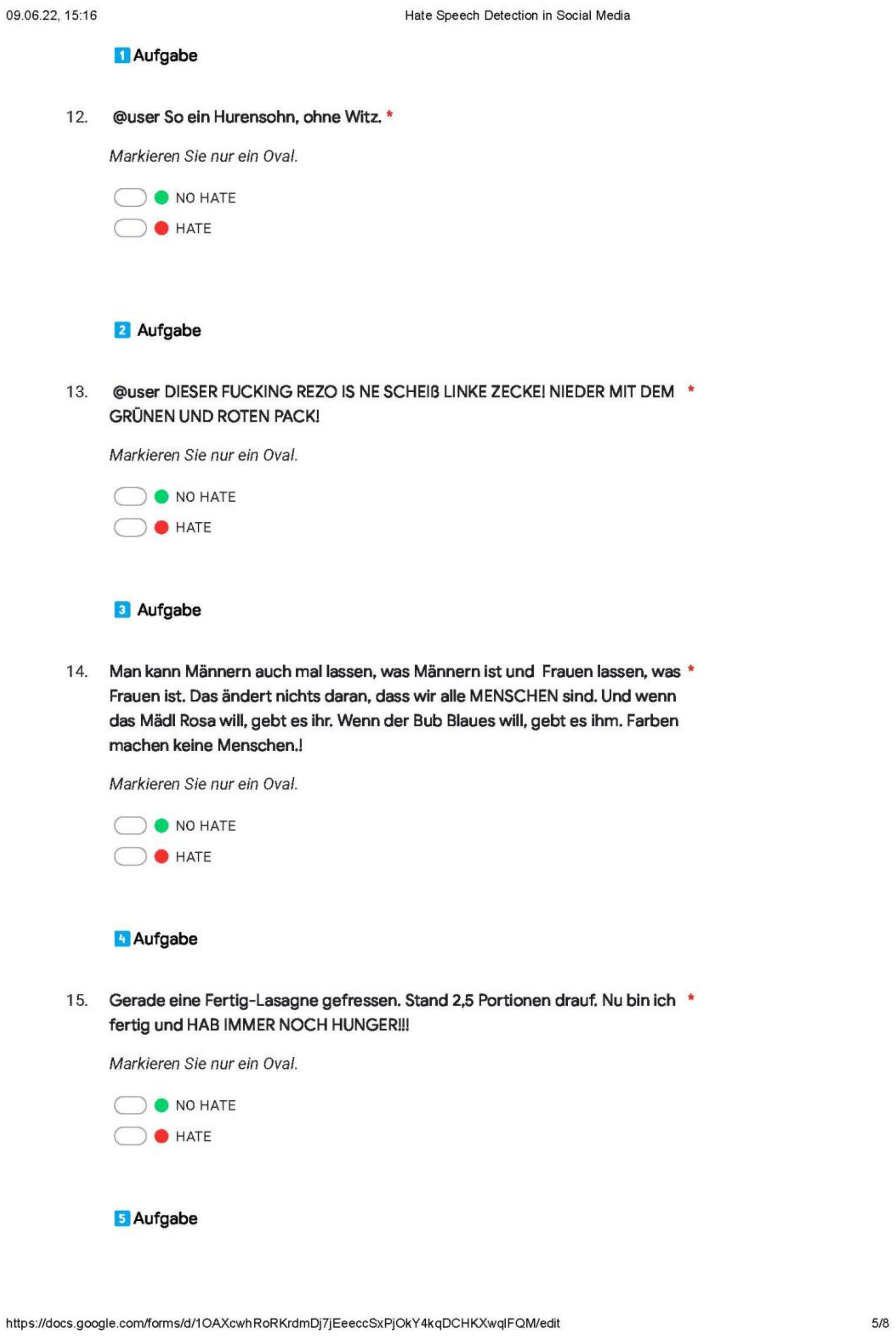
Ich will diese muslimische #Invasion hier nicht! Ich will keine #Asylschmarotzer #Krimigranten und #Rapefugees! Ich will nicht!!
🔴 HATE

@user Heidi Klum ist peinlicher als die gesamte deutsche Politik. Das ist schon ne Leistung.
🟢 NO HATE

Deine Beurteilung zur Hate Speech

Nun erhältst du 15 Beispiele zu Postings, welche du nach deinem Ermessen beurteilst und in
🟢 NO HATE oder in
🔴 HATE kategorisierst.

https://docs.google.com/forms/d/1OAXcwhRoRKrdmDj7jEeeccSxPjOkY4kqDCHKXwqIFQM/edit 4/8



1 Aufgabe

12. @user So ein Hurensohn, ohne Witz. *

Markieren Sie nur ein Oval.

○ ● NO HATE

○ ● HATE

2 Aufgabe

13. @user DIESER FUCKING REZO IS NE SCHEIß LINKE ZECKE! NIEDER MIT DEM GRÜNEN UND ROTEN PACK! *

Markieren Sie nur ein Oval.

○ ● NO HATE

○ ● HATE

3 Aufgabe

14. Man kann Männern auch mal lassen, was Männern ist und Frauen lassen, was Frauen ist. Das ändert nichts daran, dass wir alle MENSCHEN sind. Und wenn das Mädl Rosa will, gebt es ihr. Wenn der Bub Blaues will, gebt es ihm. Farben machen keine Menschen.! *

Markieren Sie nur ein Oval.

○ ● NO HATE

○ ● HATE

4 Aufgabe

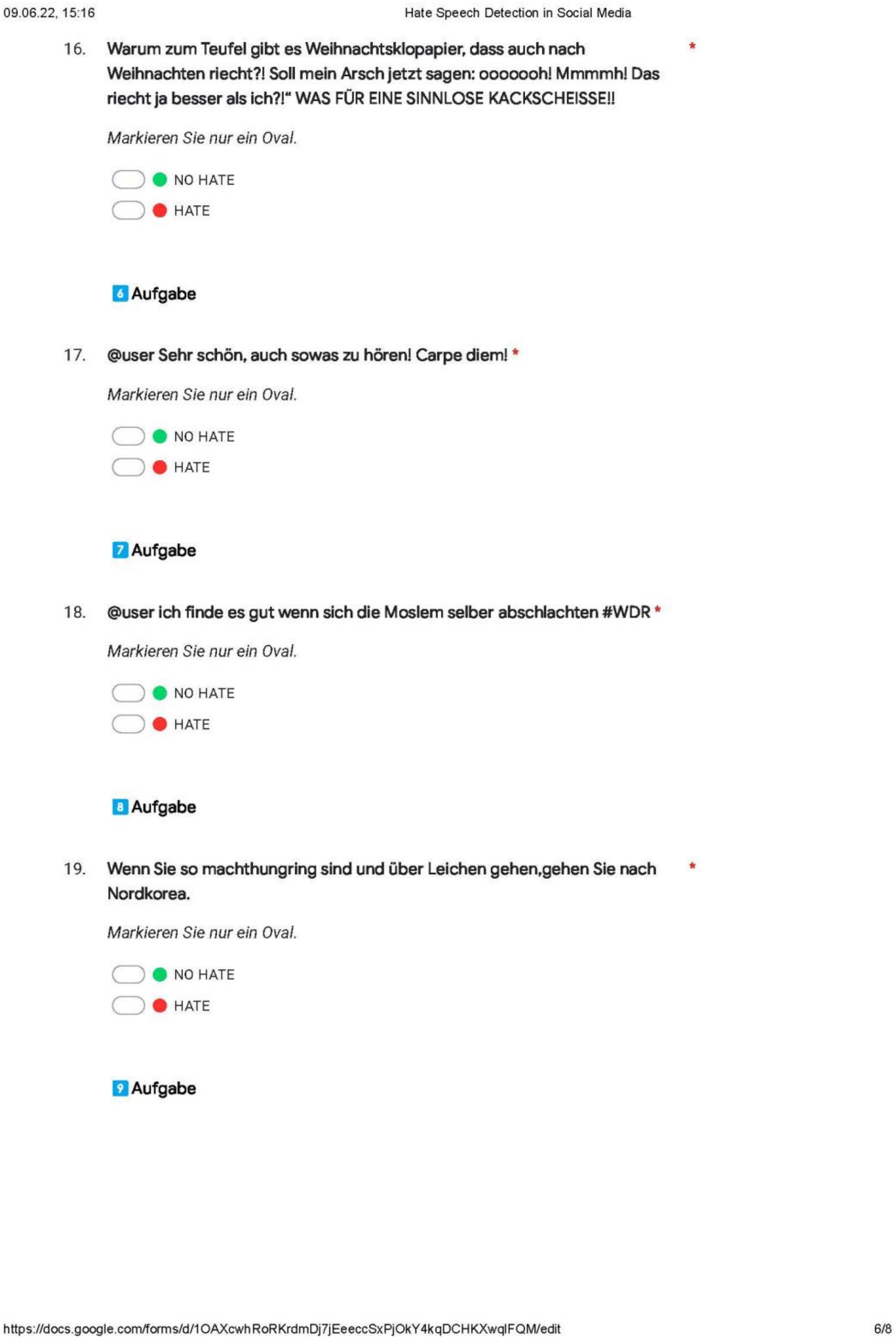
15. Gerade eine Fertig-Lasagne gefressen. Stand 2,5 Portionen drauf. Nu bin ich fertig und HAB IMMER NOCH HUNGER!!! *

Markieren Sie nur ein Oval.

○ ● NO HATE

○ ● HATE

5 Aufgabe



09.06.22, 15:16 Hate Speech Detection in Social Media

16. **Warum zum Teufel gibt es Weihnachtsklopapier, dass auch nach Weihnachten riecht?! Soll mein Arsch jetzt sagen: ooooooh! Mmmmh! Das riecht ja besser als ich?!" WAS FÜR EINE SINNLOSE KACKSCHEISSE!!** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

6 Aufgabe

17. **@user Sehr schön, auch sowas zu hören! Carpe diem!** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

7 Aufgabe

18. **@user ich finde es gut wenn sich die Moslem selber abschlachten #WDR** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

8 Aufgabe

19. **Wenn Sie so machthungring sind und über Leichen gehen,gehen Sie nach Nordkorea.** *

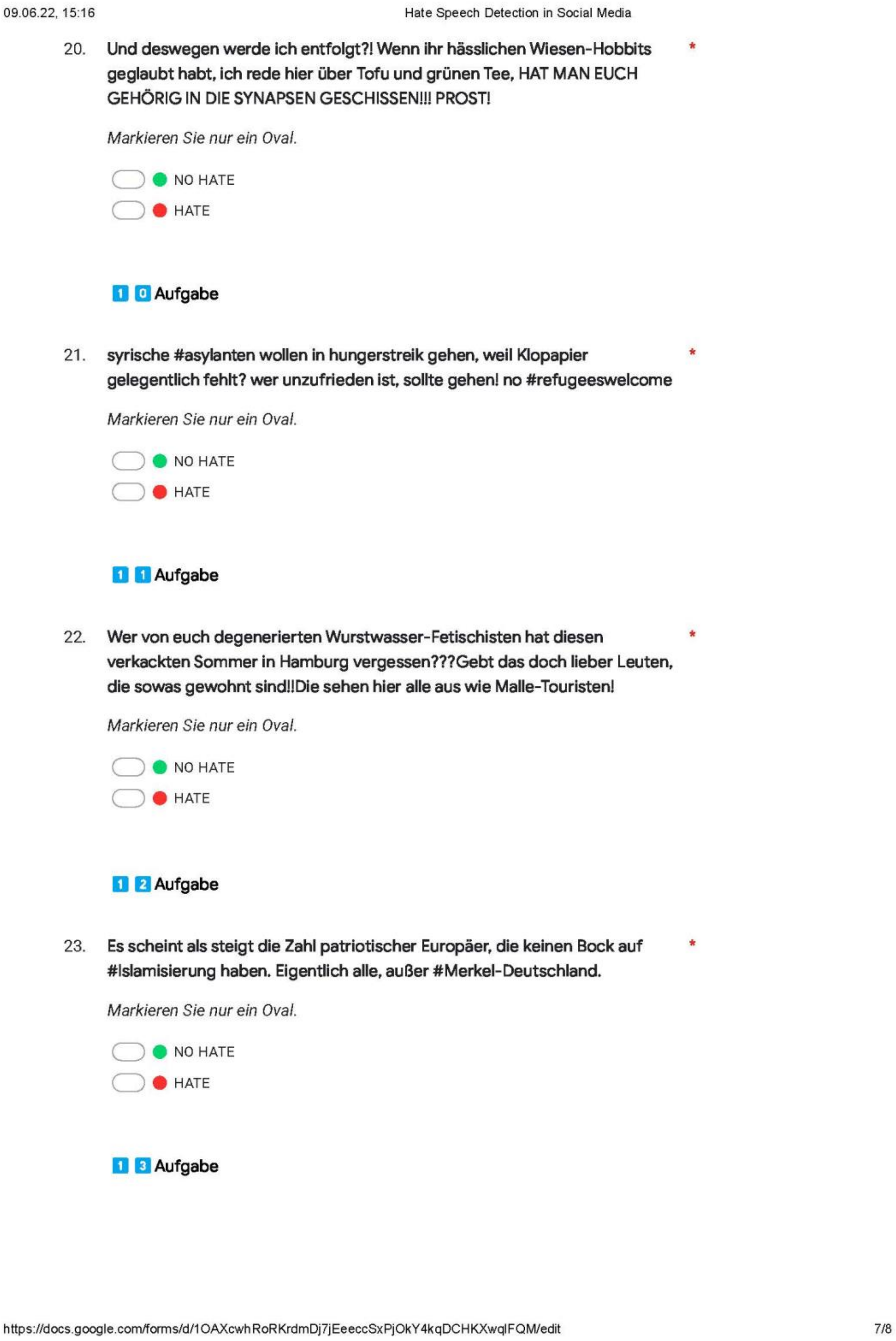
Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

9 Aufgabe

https://docs.google.com/forms/d/1OAXcwhRoRKrdmDj7jEeeccSxPjOkY4kqDCHKXwqIFQM/edit 6/8



09.06.22, 15:16 Hate Speech Detection in Social Media

20. **Und deswegen werde ich entfolgt?! Wenn ihr hässlichen Wiesen-Hobbits geglaubt habt, ich rede hier über Tofu und grünen Tee, HAT MAN EUCH GEHÖRIG IN DIE SYNAPSEN GESCHISSEN!!! PROST!** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE
○ 🔴 HATE

1️⃣ 0️⃣ Aufgabe

21. **syrische #asylanten wollen in hungerstreik gehen, weil Klopapier gelegentlich fehlt? wer unzufrieden ist, sollte gehen! no #refugeeswelcome** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE
○ 🔴 HATE

1️⃣ 1️⃣ Aufgabe

22. **Wer von euch degenerierten Wurstwasser-Fetischisten hat diesen verkackten Sommer in Hamburg vergessen???Gebt das doch lieber Leuten, die sowas gewohnt sind!!Die sehen hier alle aus wie Malle-Touristen!** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE
○ 🔴 HATE

1️⃣ 2️⃣ Aufgabe

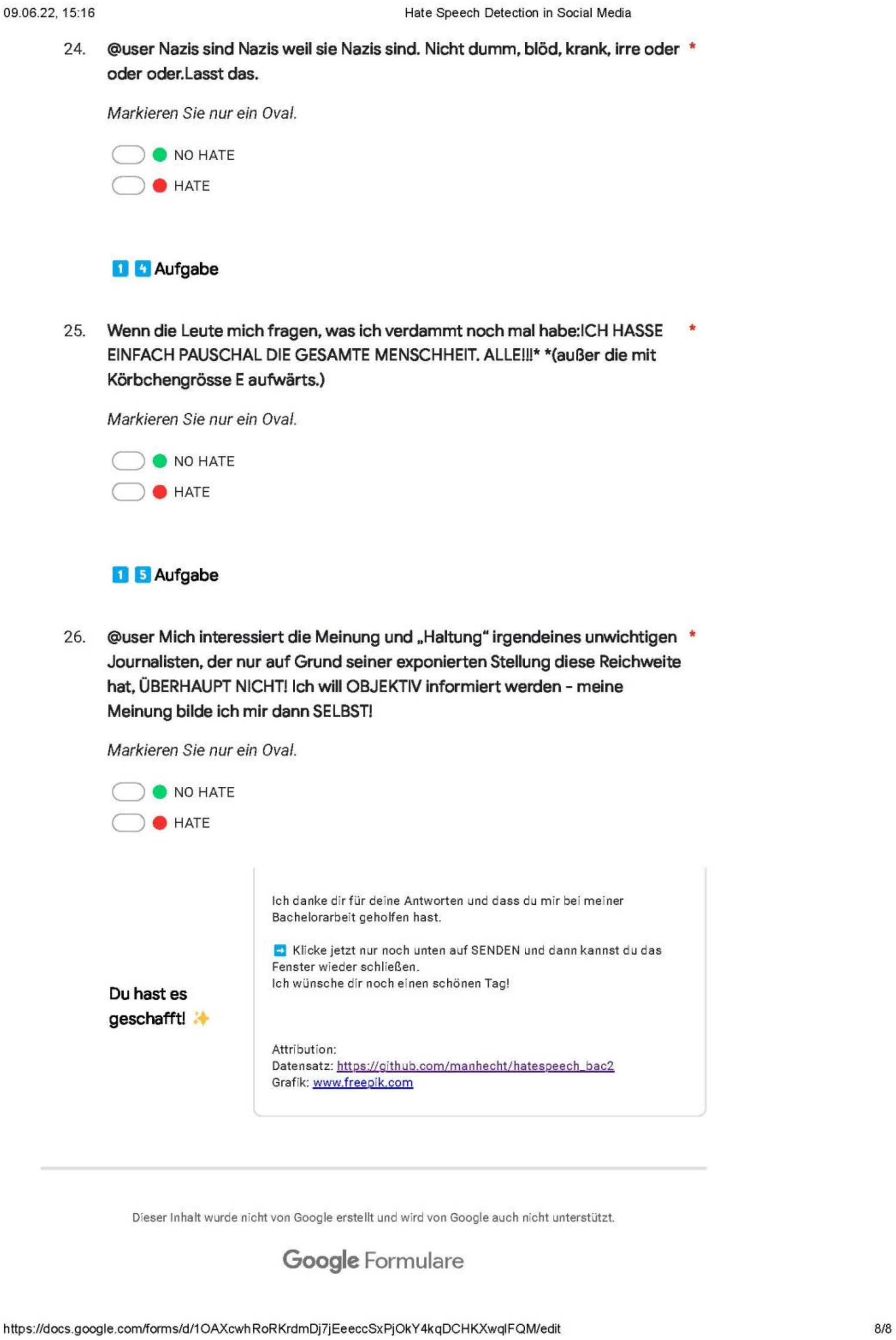
23. **Es scheint als steigt die Zahl patriotischer Europäer, die keinen Bock auf #Islamisierung haben. Eigentlich alle, außer #Merkel-Deutschland.** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE
○ 🔴 HATE

1️⃣ 3️⃣ Aufgabe

https://docs.google.com/forms/d/1OAXcwhRoRKrdmDj7jEeeccSxPjOkY4kqDCHKXwqlFQM/edit 7/8



24. **@user Nazis sind Nazis weil sie Nazis sind. Nicht dumm, blöd, krank, irre oder oder oder.Lasst das.** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

1 4 Aufgabe

25. **Wenn die Leute mich fragen, was ich verdammt noch mal habe:ICH HASSE EINFACH PAUSCHAL DIE GESAMTE MENSCHHEIT. ALLE!!!* *(außer die mit Körbchengrösse E aufwärts.)** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

1 5 Aufgabe

26. **@user Mich interessiert die Meinung und „Haltung" irgendeines unwichtigen Journalisten, der nur auf Grund seiner exponierten Stellung diese Reichweite hat, ÜBERHAUPT NICHT! Ich will OBJEKTIV informiert werden - meine Meinung bilde ich mir dann SELBST!** *

Markieren Sie nur ein Oval.

○ 🟢 NO HATE

○ 🔴 HATE

Du hast es geschafft! ✨

Ich danke dir für deine Antworten und dass du mir bei meiner Bachelorarbeit geholfen hast.

➡ Klicke jetzt nur noch unten auf SENDEN und dann kannst du das Fenster wieder schließen.
Ich wünsche dir noch einen schönen Tag!

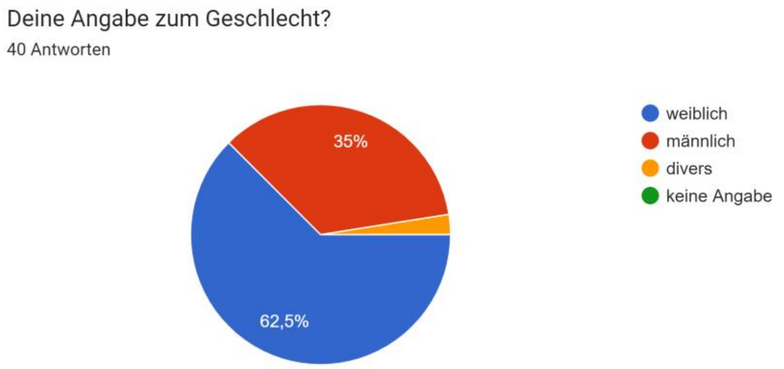
Attribution:
Datensatz: https://github.com/manhecht/hatespeech_bac2
Grafik: www.freepik.com

Dieser Inhalt wurde nicht von Google erstellt und wird von Google auch nicht unterstützt.

Google Formulare

B. Questionnaire Results

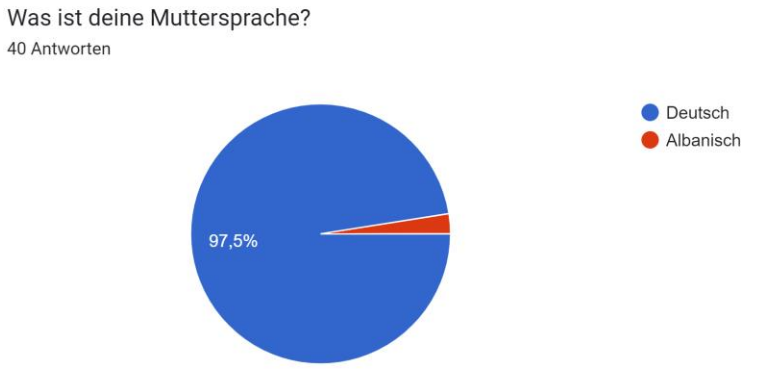
Question 1



Deine Angabe zum Geschlecht?

40 Antworten

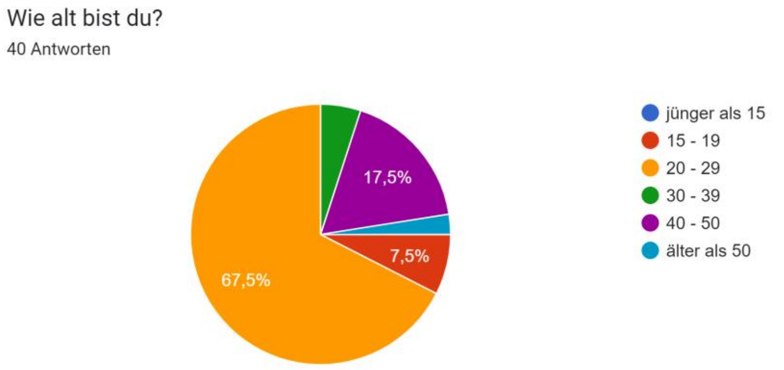
Question 2



Was ist deine Muttersprache?

40 Antworten

Question 3



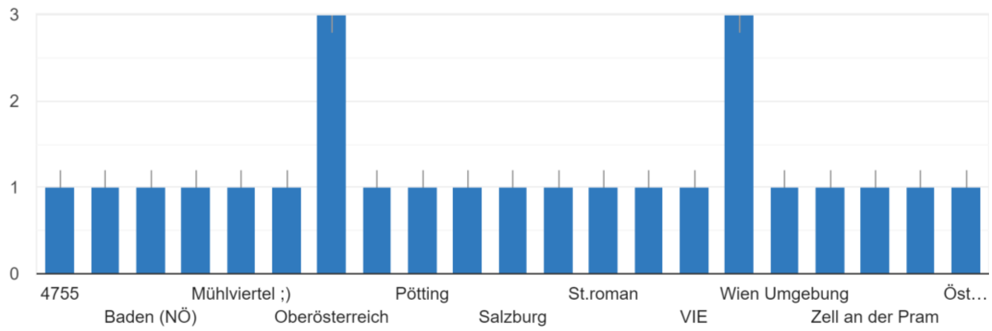
Wie alt bist du?

40 Antworten

Question 4

Wo wohnst du? (optional)

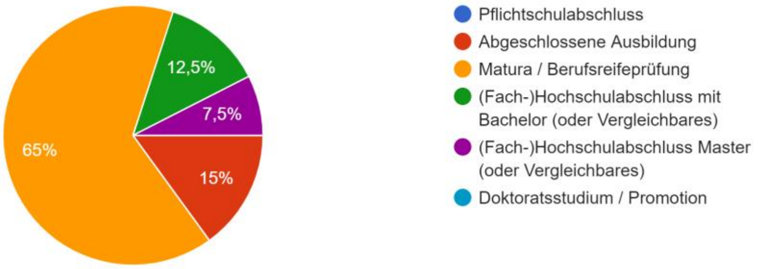
25 Antworten



Question 5

Dein höchster Bildungsabschluss momentan?

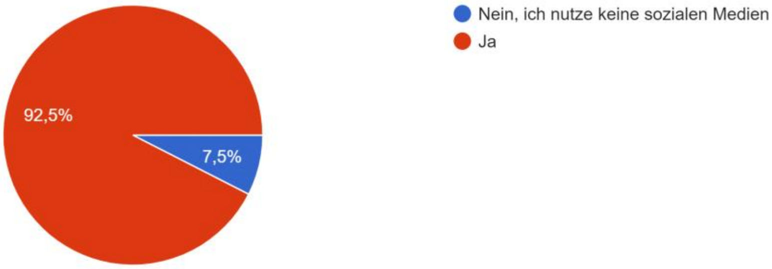
40 Antworten



Question 6

Bist du auf Social Media Kanälen?

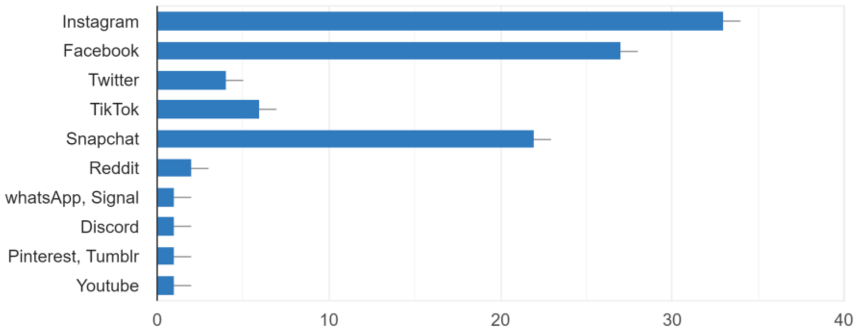
40 Antworten



Question 7

Welche sozialen Netzwerke benutzt du?

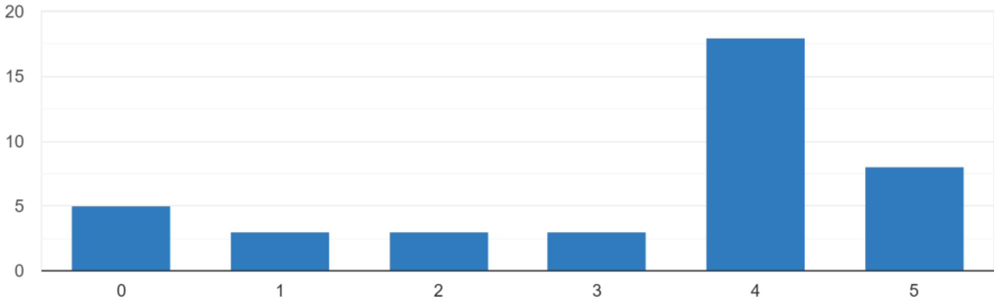
37 Antworten



Question 8

Hast du (vor dieser Umfrage) schon einmal von Hate Speech oder Hasspostings gehört?

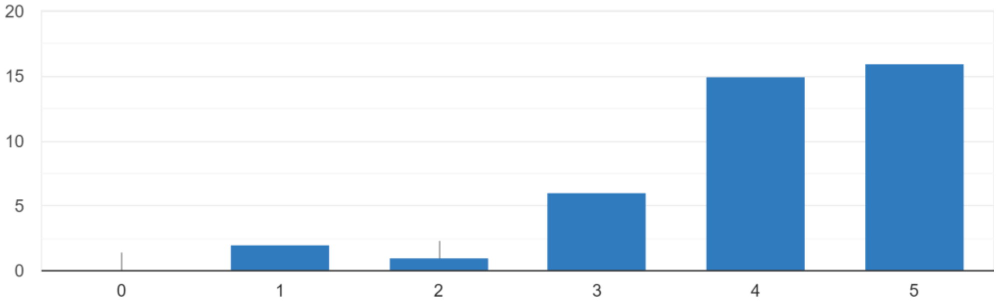
40 Antworten



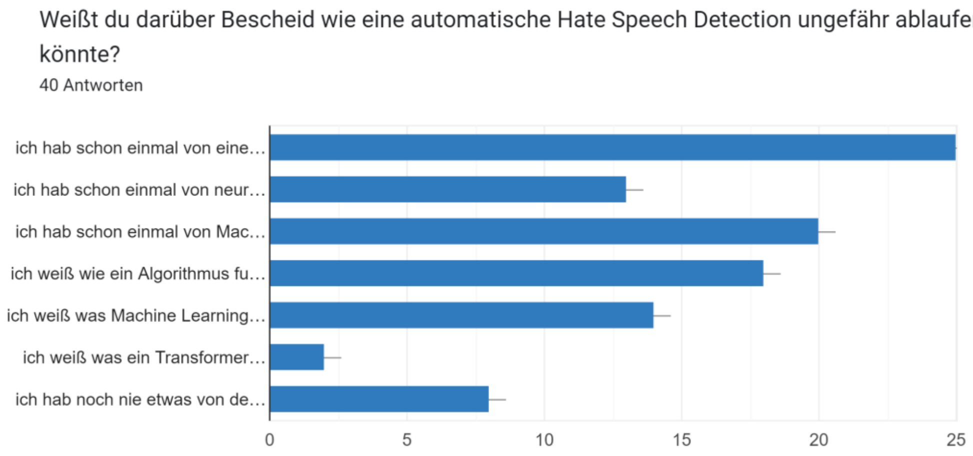
Question 9

Hast du das Gefühl, dass beleidigende Äußerungen oder Feindlichkeit in sozialen Netzwerken zunimmt?

40 Antworten



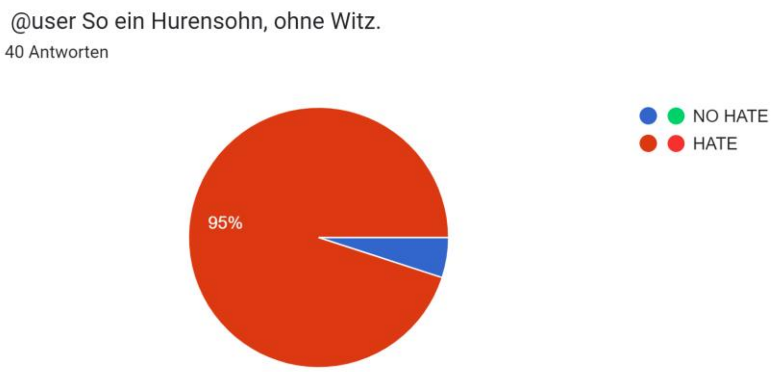
Question 10



Weißt du darüber Bescheid wie eine automatische Hate Speech Detection ungefähr ablaufen könnte?

40 Antworten

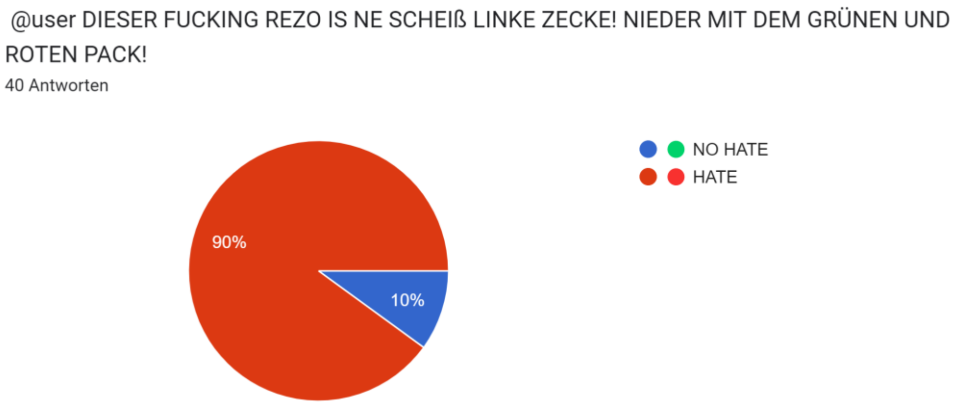
Question 11



@user So ein Hurensohn, ohne Witz.

40 Antworten

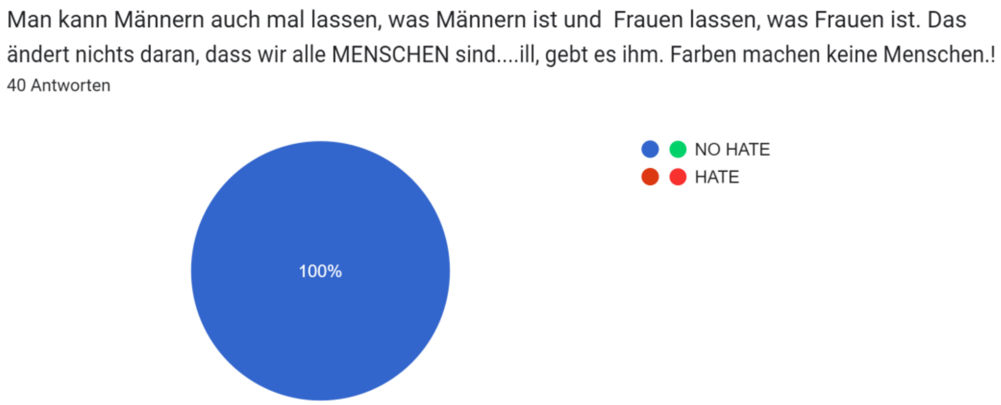
Question 12



@user DIESER FUCKING REZO IS NE SCHEIß LINKE ZECKE! NIEDER MIT DEM GRÜNEN UND ROTEN PACK!

40 Antworten

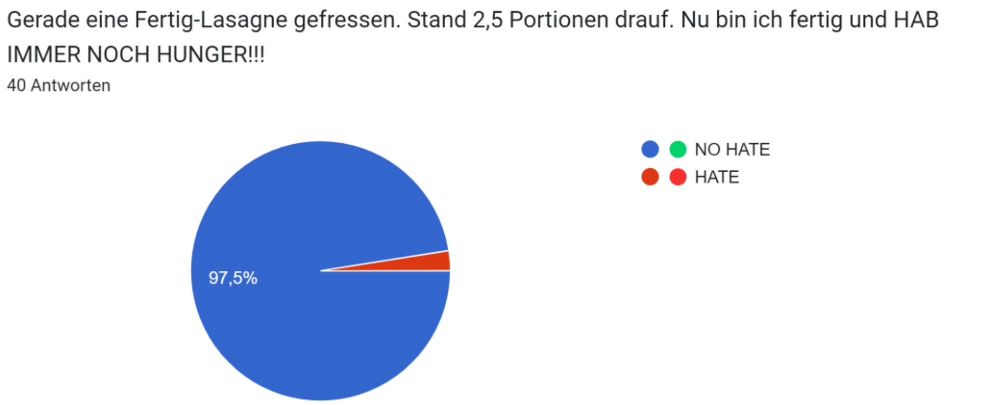
Question 13



Man kann Männern auch mal lassen, was Männern ist und Frauen lassen, was Frauen ist. Das ändert nichts daran, dass wir alle MENSCHEN sind....ill, gebt es ihm. Farben machen keine Menschen.!

40 Antworten

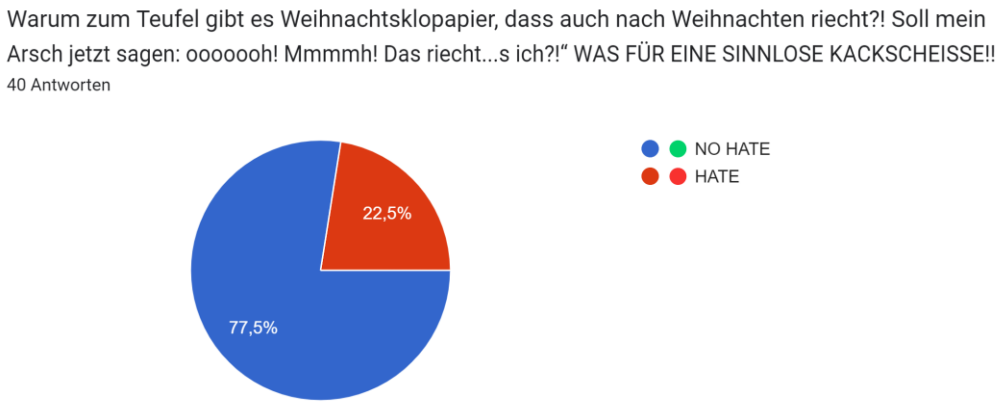
Question 14



Gerade eine Fertig-Lasagne gefressen. Stand 2,5 Portionen drauf. Nu bin ich fertig und HAB IMMER NOCH HUNGER!!!

40 Antworten

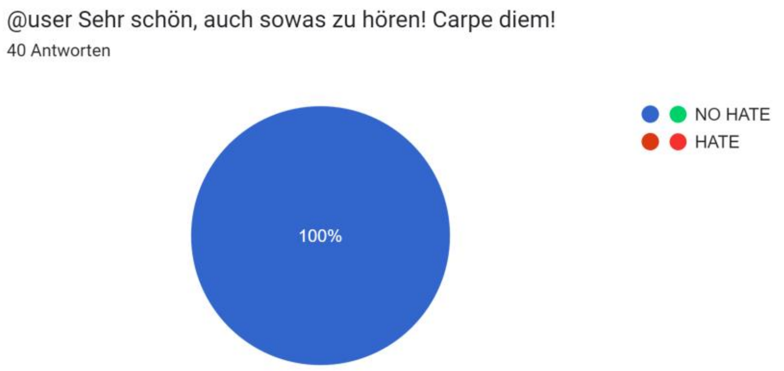
Question 15



Warum zum Teufel gibt es Weihnachtsklopapier, dass auch nach Weihnachten riecht?! Soll mein Arsch jetzt sagen: ooooooh! Mmmmh! Das riecht...s ich?!“ WAS FÜR EINE SINNLOSE KACKSCHEISSE!!

40 Antworten

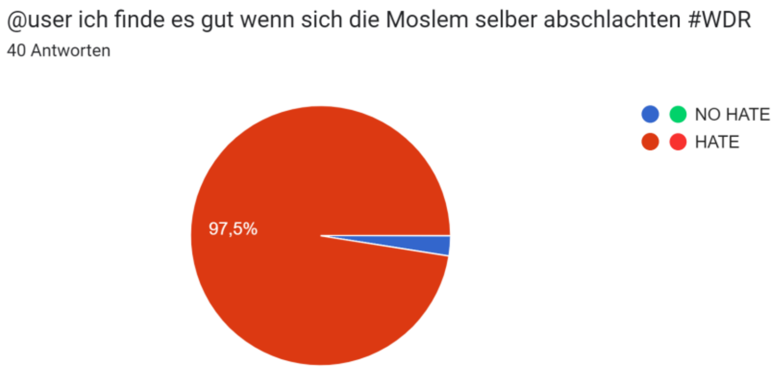
Question 16



@user Sehr schön, auch sowas zu hören! Carpe diem!

40 Antworten

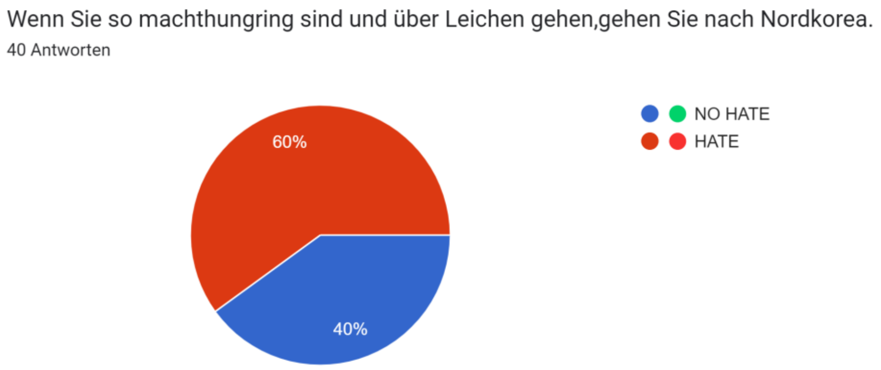
Question 17



@user ich finde es gut wenn sich die Moslem selber abschlachten #WDR

40 Antworten

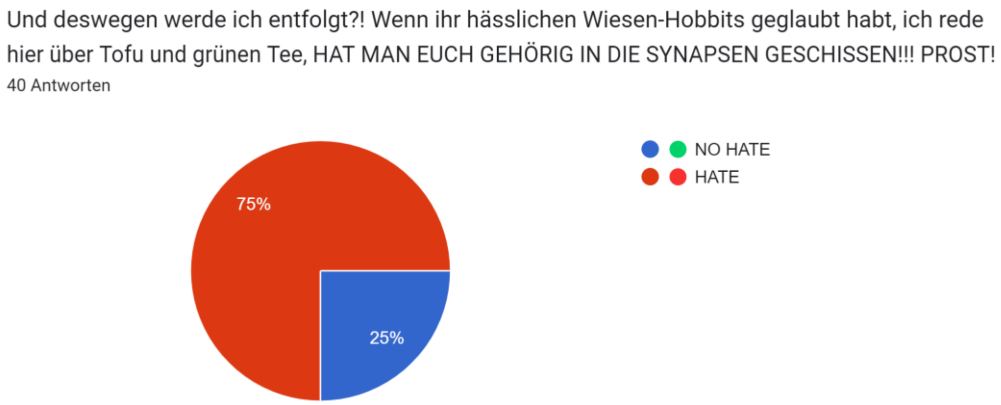
Question 18



Wenn Sie so machthungring sind und über Leichen gehen,gehen Sie nach Nordkorea.

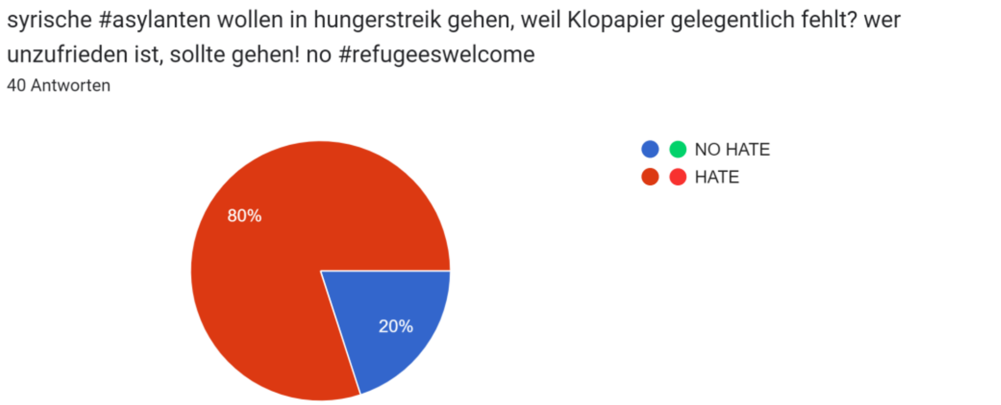
40 Antworten

Question 19


Und deswegen werde ich entfolgt?! Wenn ihr hässlichen Wiesen-Hobbits geglaubt habt, ich rede hier über Tofu und grünen Tee, HAT MAN EUCH GEHÖRIG IN DIE SYNAPSEN GESCHISSEN!!! PROST!

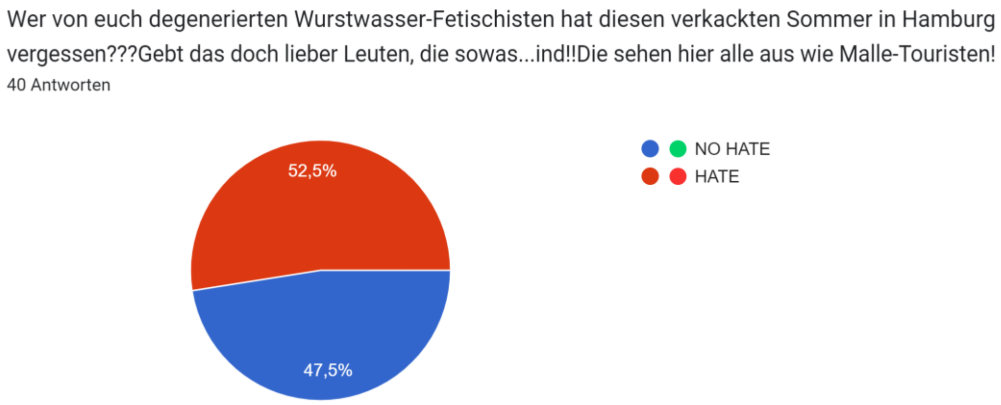
40 Antworten

Question 20


syrische #asylanten wollen in hungerstreik gehen, weil Klopapier gelegentlich fehlt? wer unzufrieden ist, sollte gehen! no #refugeeswelcome

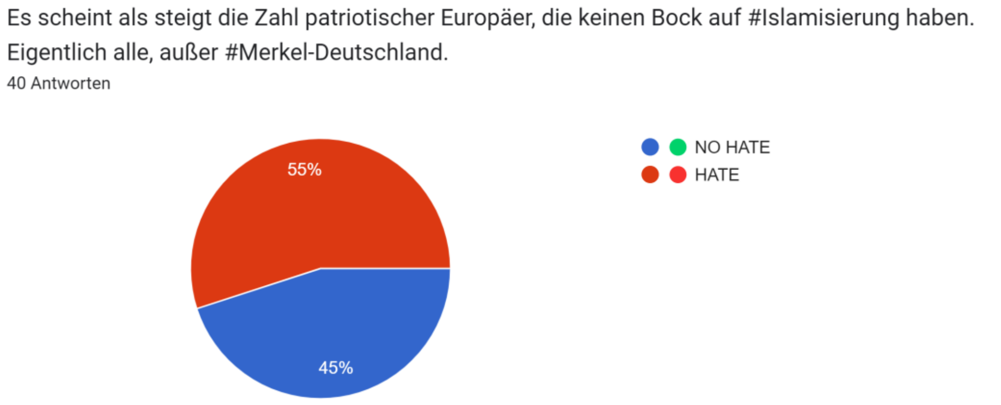
40 Antworten

Question 21


Wer von euch degenerierten Wurstwasser-Fetischisten hat diesen verkackten Sommer in Hamburg vergessen???Gebt das doch lieber Leuten, die sowas...ind!!Die sehen hier alle aus wie Malle-Touristen!

40 Antworten

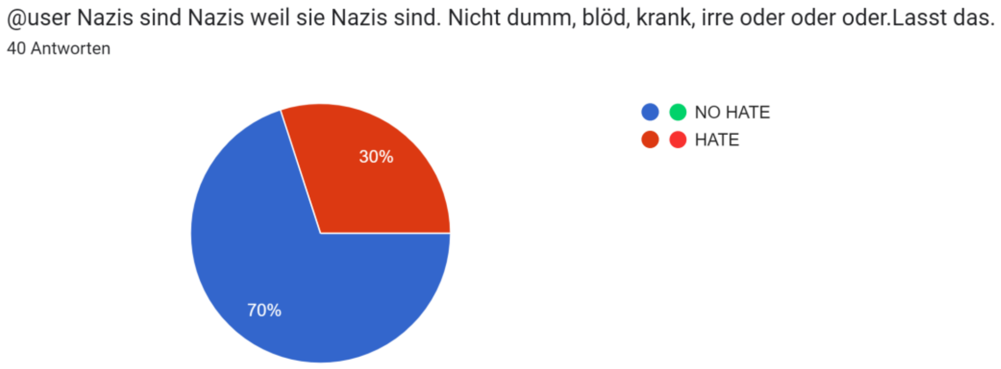
Question 22



Es scheint als steigt die Zahl patriotischer Europäer, die keinen Bock auf #Islamisierung haben. Eigentlich alle, außer #Merkel-Deutschland.

40 Antworten

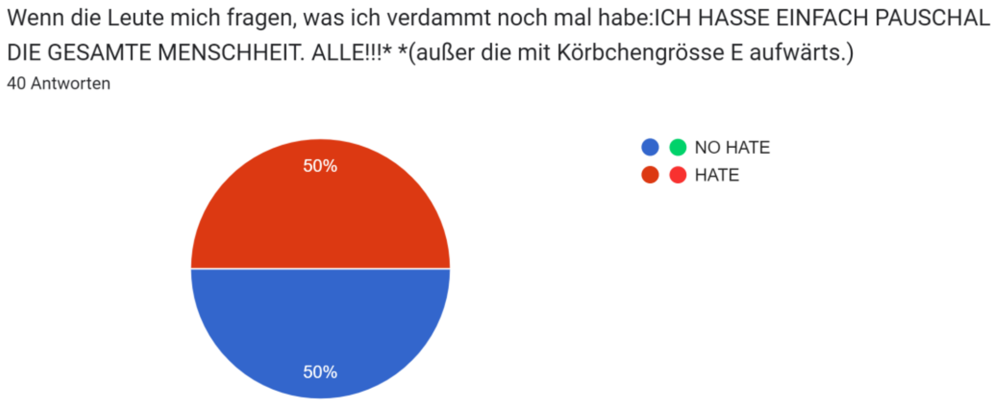
Question 23



@user Nazis sind Nazis weil sie Nazis sind. Nicht dumm, blöd, krank, irre oder oder oder.Lasst das.

40 Antworten

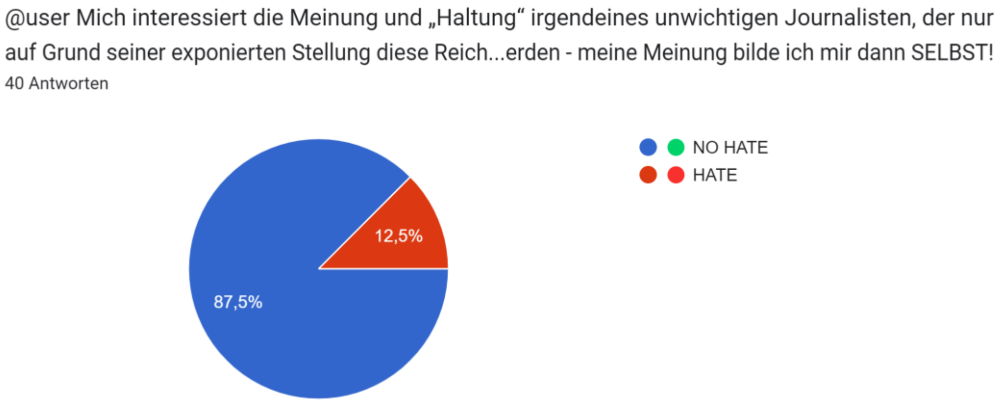
Question 24



Wenn die Leute mich fragen, was ich verdammt noch mal habe:ICH HASSE EINFACH PAUSCHAL DIE GESAMTE MENSCHHEIT. ALLE!!!* *(außer die mit Körbchengrösse E aufwärts.)

40 Antworten

Question 25



@user Mich interessiert die Meinung und „Haltung“ irgendeines unwichtigen Journalisten, der nur auf Grund seiner exponierten Stellung diese Reich...erden - meine Meinung bilde ich mir dann SELBST!

40 Antworten

C. Code

Listing 1: Code for calculation of survey metrics

```
import numpy as np
import pandas as pd
from sklearn import metrics

from src.utils import read_survey_CSV

SURVEY_CSV_PATH = "../../survey_results.csv"
df = pd.read_csv(SURVEY_CSV_PATH, sep=",")

df = df.iloc[:, 12:]
df = df.transpose(copy=True).reset_index()
df.rename(columns={'index': 'text'}, inplace=True)
df.replace(to_replace=['🟢 NO HATE', '🔴 HATE'], value=[0, 1],
                      inplace=True)

survey_df = read_survey_CSV()

column_list = ["tp", "fp", "fn", "tn", "accuracy", "f1_score"]
survey_results_df = pd.DataFrame(columns=column_list)

for i in range(1, len(df.columns)):
    single_result_df = pd.concat([survey_df['input_text'],
                                  survey_df['target_text'],
                                  df.iloc[:, i]], axis=1,
                                  keys=['text', 'y_true', 'y_pred'])

    y_true, y_pred = single_result_df['y_true'],
                     single_result_df['y_pred']

    tn, fp, fn, tp = metrics.confusion_matrix(y_true, y_pred).ravel()
    accuracy = metrics.accuracy_score(y_true, y_pred)
    f1_score = metrics.f1_score(y_true, y_pred)

    single_result_row = pd.DataFrame(data=np.array([[tp, fp, fn, tn,
                        accuracy, f1_score]]), columns=column_list)
    survey_results_df = pd.concat([survey_results_df,
                        single_result_row], ignore_index=True)

survey_results_df.to_csv('../../survey_metrics.csv')
```