Analyzing Offensive Language and Hate Speech in Political Discourse: A Case Study of German Politicians

Maximilian Weissenbacher, Udo Kruschwitz
Information Science
University of Regensburg
{maximilian.weissenbacher, udo.kruschwitz}@ur.de

Abstract

Social media platforms have become key players in political discourse. Twitter (now 'X'), for example, is used by many German politicians to communicate their views and interact with others. Due to its nature, however, social networks suffer from a number of issues such as offensive content, toxic language and hate speech. This has attracted a lot of research interest but in the context of political discourse there is a noticeable gap with no such study specifically looking at German politicians in a systematic way. We aim to help addressing this gap. We first create an annotated dataset of 1,197 Twitter posts mentioning German politicians. This is the basis to explore a number of approaches to detect hate speech and offensive language (HOF) and identify an ensemble of transformer models that achieves an F1-Macros score of 0.94. This model is then used to automatically classify two much larger, longitudinal datasets: one with 520,000 tweets posted by MPs, and the other with 2,200,000 tweets which comprise posts from the public mentioning politicians. We obtain interesting insights in regards to the distribution of hate and offensive content when looking at different independent variables.

Keywords: Social Media, Hate Speech Detection, Offensive Language, German

1. Introduction

The rise of social media has led to increased connectivity and online expression. With over half of the global population using these platforms, social media has become a vital communication medium (Braghieri et al., 2022). However, this growth has also given rise to significant challenges, particularly in controlling offensive language and hate speech due to the sheer volume of user-generated content.

To tackle the problem automated methods, including Machine Learning (ML) and Natural Language Processing (NLP), are necessary to swiftly and reliably detect harmful content while preventing post-traumatic stress in human annotators. Balancing the need to combat hate speech while preserving free speech in democratic societies is a complex challenge. An illustration of the issue's significance is the murder of Kassel's District President Walter Lübcke by a right-wing extremist, who had previously attracted attention online with spreading hate speech (Bauschke and Jäckle, 2023).

Hate speech and offensive language manifest in various forms online, leading to discussions about their precise definitions. Politicians, who are increasingly present on social media, often become targets of such content, with documented mental health consequences (Chen et al., 2012). Hate speech can have much more wide-ranging impacts on society as a whole. This has been shown in the 2019 General Election in the UK where politicians resigned due to hate speech targeted at them (Scott, 2019).

There has been some work exploring the problem area looking at English texts, however, so far there has been no systematic investigation into this using the context of German politicians (and using postings in German). Our aim is to contribute to our understanding of offensive language and hate speech in political discourse by providing an investigation that can serve as a reference point for future research looking at different political contexts. Note that the *technical* novelty is not the key contribution of the work but the exploration of a growing problem (offensive language and hate speech) in a setting that has received surprisingly little attention. As such we establish a first reference point for future investigations that go beyond the chosen setting.

This paper makes the following contributions:

- We create a dataset of tweets about German politicians[1] manually annotated to identify *hateful or offensive language* (HOF).
- We explore a variety of state-of-the-art approaches to train a classifier to detect HOF when applied to these German tweets. The best-performing classifier is used to annotate two much larger datasets[2] automatically (one comprising tweets by politicians and a second one of tweets by the general public).
- We systematically analyze how politicians

[1] our focus is on members of parliament (MPs)
[2] as well as a control dataset

TRAC-2024 Workshop, pages 60–72

and parties are targeted on Twitter.[3]

- To foster reproducibility and replicability we make all code, datasets and detailed plots available via a GitHub account[4].

2. Related Work

The focus of this work is on detecting offensive language and hate speech (Chen et al., 2012; Schmidt and Wiegand, 2017; Husain and Uzuner, 2021; Davidson et al., 2017). We use the term **Hate & Offensive Language ('HOF')** as a broader category, following Schmidt and Wiegand (2017). The task is commonly framed as supervised text classification covering both binary and multiclass cases. Traditional ML methods were shown to be effective but the performance varied with the dataset (Gitari et al., 2015; Chen et al., 2012). In recent years transformer-based models have emerged as the most promising for HOF detection (Mosbach et al., 2020; Mandl et al., 2021; Demus et al., 2022; Wolf et al., 2020).

Naturally, **datasets** for this task require manual annotation and are used for training and testing. Notable standard datasets include Davidson et al. (2017) and Waseem and Hovy (2016) for English tweets, along with datasets in other languages such as Danish and Arabic, each annotated to capture offensive language use (Chowdhury et al., 2020; Sigurbergsson and Derczynski, 2019). Several German-language datasets have been proposed including Ross et al. (2017), GermEval 2018 Datasets (Wiegand et al., 2018), HASOC 2019 (Mandl et al., 2019), HASOC 2020 (Mandl et al., 2021), and the DeTox-dataset (Demus et al., 2022). Most of these datasets have a class imbalance, e.g. sometimes as little as 12% representing hate in multi-class datasets (Founta et al., 2018). It can be argued both ways as to whether to use balanced or unbalanced datasets (Mozafari et al., 2020; Madukwe et al., 2020).

Defining offensive language and hate speech varies across datasets, especially with fine-grained annotation of multiple categories. This incompatibility issue is widespread (Fortuna et al., 2020). Also, many HOF datasets suffer from low inter-annotator agreements, showcasing the task's complexity (Ross et al., 2017; Waseem and Hovy, 2016; Struß et al., 2019). An exception is Demus et al. (2022) in fine-grained annotation for German offensive language.

Several studies delve into the role of social media in **political discourse** and analyze politicians' tweets (Antypas et al., 2023; Xia et al., 2021; Theocharis et al., 2020). Solovev and Pröllochs (2022) studied hate speech in replies to U.S. Congress politicians, observing disparities based on personal characteristics. Ben-David and Fernández (2016) investigated hate speech and covert discrimination on Facebook pages of extreme-right Spanish political parties. Fuchs and Schäfer (2021) explored misogynistic hate speech towards female Japanese politicians on Twitter, emphasizing the prevalence of negative sentiments. Agarwal et al. (2021) conducted a case study on hate speech towards UK MPs on Twitter, revealing hate concentration towards specific topics and MPs with ethnic minority backgrounds. They noted negative sentiments in cross-party conversations. Looking at German politicians on social media, Schmidt et al. (2022) performed sentiment analysis during the 2021 German Federal Election, observing a predominance of neutral and negative sentiments, with opposition parties expressing more negativity. Bauschke and Jäckle (2023) analyzed social media hate speech against German mayors, highlighting mayor reactions and their impact. Paasch-Colberg et al. (2021) mapped offensive language in German user comments on immigration, identifying a prevalence of offensive language. Jaki and De Smedt (2019) studied right-wing German hate speech on Twitter during the 2017 German Federal Election, revealing a significant portion of offensive tweets targeting the immigration policy and politicians, emphasizing the need to reduce offensive expressions online.

To conclude, this research is motivated by the ongoing need to effectively detect offensive language and hate speech on social media as well as to fully understand the general picture emerging in political discourse. In light of the detrimental impact of such posts on democratic processes and social interactions, employing advanced NLP techniques is crucial. This study aims to contribute to insights into how HOF is perceived in political discourse. Moreover, the dissemination of the annotated datasets should contribute to advancing problem-solving capabilities in this domain. The work can be seen as consisting of two parts, a technical part followed by a detailed analysis. We will first outline data acquisition and annotation before exploring different classification approaches aimed at identifying the best one to choose for the automatic classification of larger datasets which will allow us to obtain some detailed insights into the political discourse on Twitter in Germany.

3. Data Acquisition

Our work aims to get insights into how German politicians receive HOF on the social media platform Twitter. Therefore a representative dataset

[3] We will be referring to the platform as 'Twitter' in this paper.

[4] https://github.com/MaxiWeissenbacher/german_political_hatespeech_detection

	Count	Percentage
HOF	799	63.9%
NOT	359	28.7%
Not Sure	92	7.4%
Sum	1.250	100%

Table 1: Statistics of the final Annotation Dataset.

had to be acquired first. To the best of our knowledge, no public list of all German politicians with their respective Twitter accounts exists. We decided to focus on German MPs and therefore scraped this information from 'bundestag.de' (the page of the German parliament). As a result, 740 politicians were found, and 523 were identified with an active Twitter account, i.e. most politicians appear to be active on social media, in line with similar findings in the UK (Agarwal et al., 2021).
The list was then used to scrape[5] all tweets posted by politicians from 2020 until 2022, resulting in a dataframe with 521.381 tweets. We refer to this as **Politicians Dataset**. We did this to identify highly debated topics in specific months using BERTopic. Several studies (Solovev and Pröllochs, 2022; Theocharis et al., 2020) tried to find a reasonable period of time when scandals or events that are relevant for politics have happened. We did this with BERTopic (Grootendorst, 2022) and two prominent topics emerged: discussions about the withdrawal of German troops from Afghanistan in July 2021 and the start of the Russo-Ukrainian war in February 2022. Other dominant themes included elections, climate protection, and Corona vaccination discussions until September 2021, with a resurgence in winter.
We used these two prominent topics to create our HOF detection dataset for a two-month period in line with Agarwal et al. (2021), where a politician is mentioned by the public. The baseline dataset consists of tweets from February 2022 until April 2022. Also, a control-group dataset was built to generalize findings containing tweets from July 2021 until September 2021. As a result, the baseline dataset consists of 2.226.216 million tweets (1.775.251 after removing duplicates) with 160.845 different users (referred to as **Mentions Dataset**) and the control group dataset with 1.534.835 million tweets and 116.680 unique users (**Control Group Dataset**).

4. Data Annotation

To train machine learning models or to fine-tune large language models on the task of HOF detection, a subset of the created datasets has to be annotated. For the annotation, over 20 native speakers were used, all of whom were members of the University of Regensburg. Most of them were students of Information Science and were compensated in a manner related to their studies (experimental hours). We used a binary classification: HOF (hate, offensive or profane content) and NOT following existing guidelines (Wiegand et al., 2018; Mandl et al., 2019, 2021). The detailed guidelines can be found in the Github repository. If the annotators were unsure, they should classify the tweets as "Not Sure" (NS). They were asked to annotate as objectively and neutrally as possible, even if a tweet did not reflect their political opinion. The simplest method to create an annotation dataset would be to randomly sample a specific number of tweets and use them for labeling the data. However, this approach would likely result in a very small proportion of HOF tweets. To get more HOF tweets, we filtered tweets containing words from the 'https://insult.wiki' lexicon, containing more than 6000 German swear words. We further applied a sentiment model (Guhr et al., 2020) to the filtered tweets and only used tweets with a negative sentiment assuming that negative sentiment is more likely related to hate speech (Schmidt and Wiegand, 2017; Alfina et al., 2017). As a result, 86k tweets with swear words and a negative sentiment were retrieved.
To ensure good annotation quality a pilot study compared the inter-annotator agreement between five crowd-sourcing annotators[6] and five annotators in our own institution. Each group labeled 100 tweets. The annotators from Prolific were paid fairly, while the annotators from our institution could have their time counted towards study-related credits. For this, a web application on 'Streamlit' with 'AWS' was built to make the annotation process accessible online. Somewhat surprisingly, the Fleiss Kappa score of our own annotators was 0.4 higher than from the Prolific annotators with κ = 0.71. Therefore we conducted the remaining annotation in-house. Many studies (Schmidt et al., 2022; Mandl et al., 2021) rely on just three annotators with majority voting, but we decided to use five annotators per tweet to increase the quality. Five groups with five persons per group annotated 250 tweets each resulting in an annotated dataset of 1.250 tweets, each classified by five annotators (1.197 tweets with removing no-majority group tweets). The inter-annotator agreement can be interpreted as substantial (κ = 0.69). Table 1 shows the class distribution of the final annotation dataset. Some tweet examples can be found in Table 2.

5. Implementational Aspects

Before looking at the actual experiments to identify the most suitable classification approach we

[5]using the Twitter API V2 for Academic Research

[6]We used Prolific: prolific.com

Tweet	English Translation	Label
@BonengelDirk @Beatrix_vStorch @jamila_anna @KathrinAnna Dumm wie Brot und absolut unfähig! Und mehr gibt es zu diesem Abschaum von Heuchlern nicht zu sagen	@BonengelDirk @Beatrix_vStorch @jamila_anna @KathrinAnna Stupid as bread and absolutely incompetent! And there is nothing more to say about this scumbag of hypocrites	HOF
@SaraNanni @OlafScholz Leider hat sich die Außenpolitik hinsichtlich Menschenrechte nicht wirklich geändert. Weitere Kooperationen mit Diktaturen ist einfach ein No Go.	@SaraNanni @OlafScholz Unfortunately, foreign policy on human rights hasn't really changed. Further cooperation with dictatorships is simply a no go.	NOT
@Hendrixx_T6 @Jackisback110 @Nicole_Hoechst Thematisieren und Pöbeln sind zwei verschiedene Sachen. Wer hier dauernd von Diktatur, Staatsfunk oder Merkelmilizen wie Brandner redet, will nur den Pöbel auf der Strasse mobilisieren! #EkelhAfD	@Hendrixx_T6 @Jackisback110 @Nicole_Hoechst Thematizing and bullying are two different things. Anyone who keeps talking about dictatorship, state radio or Merkel militias like Brandner just wants to mobilize the rabble on the streets! #DisgustingAfD	NS

Table 2: Annotation examples.

will report some implementational aspects (more details on Github). BERT-based models were obtained from Hugging Face using Transformers (Wolf et al., 2020), fine-tuned with the Huggingface Trainer API in PyTorch. These models were programmed in JupyterLab with access to an 'NVIDIA GeForce RTX 2080 Ti' GPU.

Different models were trained on the unbalanced data (Table 1) as a pilot study to understand which models work well. In total, 16 different models were implemented, mostly BERT-based. The overall best results were achieved with the "Electra German Uncased" model and the "German Toxicity Classifier" with an F1-Macro score of 0.77. To get the optimal combination of hyperparameters we did hyperparameter optimization and found using the Optuna Grid Search framework with 20 trials worked better than a randomized search with WandB. The hyperparameter search resulted in a learning rate of 4.5e-05, 5 Epochs, a Batch Size of 8, a Weight Decay of 0.02 and 0.3 Warmup Steps. These hyperparameters were used for all models in the following approaches. We focus on F1, Precision, and Recall for evaluation and not accuracy due to data imbalance (using 5-fold cross-validation). Statistical significance is assessed with two-tailed t-tests ($p < 0.05$), and for the data analysis part we computed individual scores for every week and then applied t-tests.

6. Identifying the Best Classifier

To identify an effective classifier for our unannotated datasets, we explored various methodologies, focusing on model generalizability and performance validation. For all of the following approaches, the same test dataset was used. Addressing data imbalance was our first step, incorporating 'NOT'-Tweets from the GermEval 2018 dataset to achieve balanced class distribution. This method, avoiding over- and undersampling to prevent overfitting and data loss, significantly improved the F1-Macro score by 8% with the Electra German Uncased model.

Model: Voting	F1	Precision	Recall
Ens. 3: Soft	0.90	0.90	0.90
Ens. 3: Hard	0.94	0.94	0.94
Ens. 5: Soft	0.88	0.88	0.88
Ens. 5: Hard	0.89	0.89	0.89

Table 3: Macro Ensemble Modeling results.

Further, we expanded our dataset by combining training data from GermEval 2018, 2019, and HASOC 2019, which increased the sample size from 1,158 to 17,363. However, this led to an unbalanced class distribution (30.7% HOF) and a 2% decrease in classification performance, likely due to varied data quality and class distribution. We made sure that there is no duplicated data in the test and training datasets when using additional data.

An ensemble approach, utilizing combinations of three and five classifiers with hard and soft voting, demonstrated superior performance. Specifically, an ensemble of 'Electra German Uncased', 'German Toxicity Classifier', and 'Deepset gBERT Base' models emerged as the most effective, as summarized in Table 3.

These results illustrate a hard-voting ensemble of three systems as the best solution, achieving an F1 of 0.94. This ensemble strategy proved effective, with the model correctly predicting 153 out of 159 'HOF' test samples. All three individual models are published on the Huggingface platform.[7]

Transfer learning evaluations on GermEval 2019 and HASOC 2019 Subtask A German test datasets yielded mixed outcomes. While the model performed exceptionally well on HASOC, demonstrating successful transfer learning, it achieved modest results on GermEval 2019. This

[7] https://huggingface.co/mox/

variance underscores the complexities of transfer learning, even with consistent annotation guidelines across datasets.
Before applying the hard-voting ensemble of three classifiers to annotate the full datasets using our binary classification scheme ('HOF' or 'NOT') we conducted a sanity check. We had the model predict 100 random tweets (17 HOF, 83 NOT), and three annotators classified the same tweets. The inter-annotator agreement between model predictions and human annotations yielded a Fleiss κ score of 0.70, slightly higher than the agreement among human annotators in the final annotation. The model correctly classified 14 out of 17 'HOF' tweets, resulting in an average macro F1-Score of 0.85.

7. Analysing Political Discourse

We applied the best-performing hard-voting ensemble to automatically annotate all three datasets, i.e. 'Politicians', 'Mentions' and 'Control Group'. In case a tweet mentioned more than one politician, we duplicated the tweet.
Again we refer the interested reader to the repository for detailed information, code, plots and figures on all the analyses.

7.1. Politicians Dataset

First, we analyze the 'Politicians' dataset with 521.381 tweets.[8] As expected, the amount of HOF from MPs to MPs is relatively low, with 2.56%. We notice that the 'AfD' (far right on the political spectrum) spreads significantly and consistently more HOF over time than the other parties. For the remaining parties, the proportion of tweets posted tagged as HOF is approximately the same.
Looking at the targets of hateful and offensive language and taking gender as the independent variable, we see no significant difference between male (2.9%) and female (2.3%) MPs. Drilling down to the individual posters to identify which politician is posting the most tweets towards an MP classified as HOF we find Martin Reichardt of the 'AfD' (username: m_reichardt_afd) to be the highest ranked one. On the other hand we observe that Olaf Scholz (SPD, centre-left), Karl Lauterbach (SPD), and Christian Lindner (FDP, liberal) received the most offensive tweets from other politicians. All three are government ministers.
Here is an example tweet that was classified by the model as HOF posted by Marin Reichardt that offensively mentions Karl Lauterbach:

> *"@BMG_Bund @Karl_Lauterbach Lasst doch bitte das Pflegepersonal mit dem Geschwätz dieses inkompetenten, verwirrten Narzisten in Ruhe! #Pflegenotstand #LauterbachRuecktrittJetzt"*

Looking at the party level, we find that most HOF tweets are spread by the 'AfD' (24%) and the 'SPD' (22%). The 'CDU/CSU' (centre-right), 'FDP', and 'Bündnis 90/Die Grünen' (left) combine in a similar percentage range of 15-17%. The least HOF content was spread by 'Die Linke' (far-left).
Looking at the parties that receive the most offensive content, we see that the 'SPD' receives significantly more than the other parties with 39% of all HOF-classified tweets. The distribution of the remaining parties looks similar to those of the parties that spread HOF, with the exception of the 'AfD'. Interestingly we see that the 'AfD' receives only 5.4% of HOF-classified tweets, which is slightly above the value of 'Die Linke' with 4.7%. Network analysis showed that the 'SPD', 'Bündnis 90/Die Grünen' and the 'CSU/CDU' are tightly knit where the 'AfD' is slightly decoupled from the other parties. However, there is still interaction between all parties, which can be seen in Figure 1 (Each color represents a party: Green = 'Die Grünen'; Red = 'SPD', Yellow = 'FDP', Blue = 'AfD', Black = 'CDU/CSU', Purple = 'Die Linke').
Commonly, an MP mentions colleagues in their own party. We also observe that many HOF tweets originating from the 'SPD' are targeted again towards politicians of the same party. One reason could be that an 'SPD' politician mentions a colleague in a tweet and then offends a different person. This is where our approach of not drilling down further has its limitations as we do not aim to determine exactly the person a tweet is targeted at in cases where more than one politician is being mentioned in a tweet. We leave a detailed exploration of this for future work.

7.2. Mentions Dataset

Let us now focus on the 'Mentions' dataset, i.e. the crawl of tweets that were posted by the general public mentioning the Twitter handles of German MPs. As already indicated, the dataset consists of more than 2 million tweets from over 150 thousand different users. 456.374 of those tweets were classified as HOF (20.5%).
Figure 2 shows the distribution of HOF-classified tweets targeted at each individual political party. It can be seen that the 'AfD' receives the largest proportion of hateful or offending messages (over 30% of all tweets targeted at the party). As an illustration we also include a word cloud with the most frequent words found in offending tweets (Figure 3). The term frequency analysis shows that topics like 'nazi', 'putin' or 'fckafd' are often mentioned in HOF tweets.[9]

[8] The 'Politicians' dataset covers a time with SPD, FDP and Bündnis 90/Die Grünen forming a coalition government in Germany.

[9] Additional word clouds can be found in the project

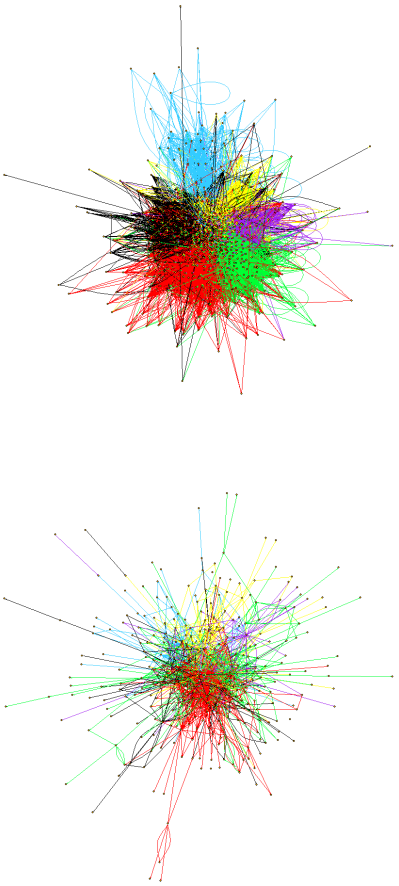

Figure 1: Top: Network Graph: "Who mentions whom?"- Bottom: Network Graph: "Who spreads HOF?".

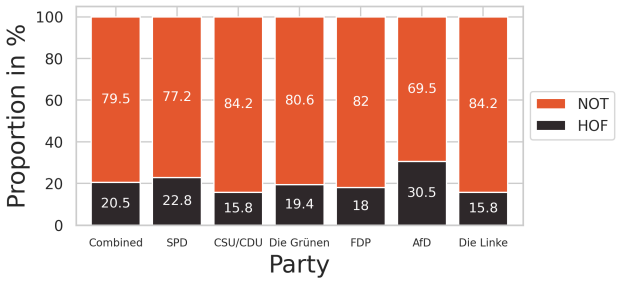

Figure 2: HOF per party (Mentions Dataset).

The 'SPD' and 'Die Grünen' are second and third in the ranked list of HOF-classified tweets targeted at the party level with 'CDU/CSU' and 'Die Linke' at the bottom. Interestingly, this pattern is in line with what the 'Control Group' dataset shows.
We also investigated whether there is a noticeable difference between Government ('SPD','Die Grünen', 'FDP') and Opposition ('CDU/CSU', 'AfD', 'Die Linke') parties, but found no significant difference in the amount of HOF content received by each group.

repository.

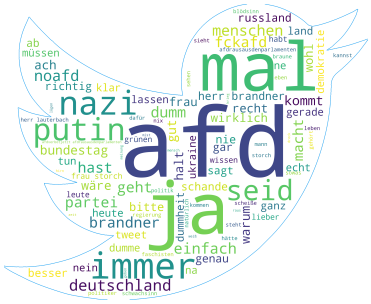

Figure 3: 'AfD' HOF word cloud (Mentions data).

We were also interested in the virality of a tweet based on its class. We found that on average, a HOF tweet has fewer likes (-1.65 likes), fewer replies (-0.28 replies), and fewer retweets (-0.27 retweets) than a NOT tweet (Average Likes: 6.24; Average Replies:0.61; Average Retweets: 0.65).
Analyzing offensive posts by gender (of the mentioned MP) we find that there is a statistically significant difference between male and female politicians with male politicians receiving more hateful and offensive content than female ones ($p = 0.04$).
Looking at a more fine-grained level of individual politicians, we notice a clear outlier. Karl Lauterbach (SPD, Minister of Health) is both mentioned the most (almost 20% of all tweets) and is also the most 'attacked' politician by far. The term frequency analysis shows that topics like 'corona' or 'impfung' (vaccination) are often mentioned when there is a tweet mentioning Karl Lauterbach.
Figure 4 displays the total counts of tweets tagged as 'HOF' and 'NOT', respectively, for the 15 most commonly mentioned MPs and it can clearly be seen how Karl Lauterbach stands out. The 'Control Group' dataset offers the same insight which is somewhat surprising because he was not yet in office as Minister of Health (the post was held by Jens Spahn at that point who was only the third-most commonly HOF-targeted MP). Nevertheless, the actual traffic targeted at Karl Lauterbach increased substantially.
There is one other interesting difference between the 'Mentions' and the 'Control Group' datasets. The percentage of HOF-classified tweets in the 'Control Group' dataset is smaller than in the 'Mentions' dataset (14.8% vs. 20.5%). This could possibly be explained because the overall sentiment in Germany was perhaps more positive right before the election.

8. Discussion

We discuss, reflect on and contextualize the three main parts of our work, i.e. **dataset creation** and **annotation**, the **modeling** part looking at identify-

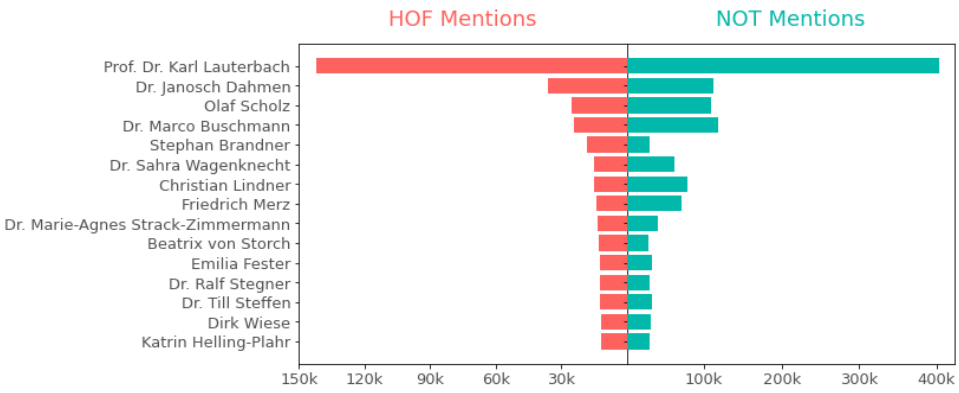

Figure 4: HOF distribution on MP level (Mentions Dataset).

ing a classifier with its experimental work, and the **data analysis** part.

8.1. Data Acquisition and Annotation

Utilizing an official German parliament Web site for scraping current MPs' Twitter usernames, we created three datasets: **Politicians**, **Mentions**, and **Control Group Dataset**. The Politician Dataset covers a broader time frame to secure sufficient data from 523 users, unlike the Mentions Dataset's 160,000 users. A limited two-month period would have been inadequate for reliable analysis. It was employed to pinpoint key topics for the Mentions Dataset's time span selection. A subset of the Mentions Dataset was used to create an annotated dataset to aid HOF detection classifier development. We found it important to use annotation guidelines that have already been used in previous works (Wiegand et al., 2018; Mandl et al., 2019), as one key problem of many publicly available datasets is that different definitions of hate speech or offensive language are being used and that they are therefore not compatible for transfer learning tasks (Fortuna et al., 2020).

We encountered challenges in creating a balanced dataset due to a relatively small proportion of messages tagged as hateful or offensive. We chose a binary classification task focusing on whether a politician is targeted by HOF in general rather than specific hate types, as the agreement between annotators decreases with a more fine-grained classification (Ross et al., 2017; Waseem and Hovy, 2016; Kwok and Wang, 2013). To handle the class imbalance, we adopted a strategy to gather more positive class samples, which may result in better generalization of ML models (Madukwe et al., 2020) but on the other hand this could lead to bias when applying to a real world scenario. We had each tweet classified by five annotators to have the most robust possible justification for the label of each tweet, and conducted the work in the spirit of the Perspectivist Manifesto[10]. Finally we achieved a substantial agreement (κ=0.69) – higher than in previous studies (Ross et al., 2017; Waseem and Hovy, 2016; Kwok and Wang, 2013; Struß et al., 2019). We highlight the efficacy of the lexicon-sentiment approach, with 63.9% of tweets classified as HOF, albeit not reflecting real-world class imbalance. We note that the data set size is clearly limited in size and scope.

8.2. Modeling

In our study, BERT-based models emerged as the most effective for classification, corroborated by existing research (Mandl et al., 2019; Wiegand et al., 2018; Demus et al., 2022). Addressing data imbalance by integrating NOT tweets from different datasets, as per consistent annotation guidelines, led to an 8% F1-Macro improvement for the Electra German Uncased model. However, preprocessing that removed social media nuances, like emojis, reduced performance. Expanding training data resulted in a 2% F1-Macro decrease due to class distribution imbalances. Our annotation approach, involving a team of five, ensured data quality and model reliability, contrasting with other methods that used fewer annotators (Struß et al., 2019). Ensemble learning further improved our model, achieving a competitive F1-Macro score of 0.94 (Zimmerman et al., 2018). Generalizability tests showed varied results, indicating future research opportunities. A sanity check with manual annotations confirmed the model's efficacy in HOF prediction, aligning with the literature (Chowdhury et al., 2020; Sigurbergsson and Derczynski, 2019) and validating our annotation quality.

8.3. Data Analysis

Analyzing the three datasets revealed challenges in identifying the exact target of a tweet when multiple individuals are mentioned. Despite this chal-

[10] http://pdai.info/

lenge, the analysis identified that 2.56% of tweets from politicians were classified as Hate and Offensive Language (HOF), with the Russo-Ukrainian war being a prominent topic. Hateful tweets were predominantly from MPs of the 'AfD', followed by the 'SPD', consistent with prior research by Ben-David and Fernández (2016) on hate dissemination by political parties, where their main finding was, that extreme-right political parties and the mainstream party in Spain spread the most hate.

Jaki and De Smedt (2019) also found that even political leaders broadcast hate speech, often used as a tactical instrument. Looking at which MP receives the most hate from other MPs, we see several leading politicians. We should however also note that some key politicians do not have a Twitter account or were not listed which means that any findings we offer can only be a partial picture.

An interesting (and worrying) finding is that 20.5% of all tweets posted by the public in which a MP is mentioned were identified as hateful or offensive. Looking at a party level, we see that unlike in the politicians' dataset, where the 'SPD' received the most hate, in this dataset 'AfD' MPs are mentioned in the most HOF-Tweets with 30.5%. This was also confirmed with the analysis of the Control Dataset. This shows that the 'AfD' spreads much hate among politicians and is less so the target while the general public (as represented on social media) tends to target the party in public discourse. This suggests that other politicians do not respond to the 'AfD's' jibes and largely ignore them. The mainstream, however, does not and mentions them most often in HOF-Tweets. This manifests in a high occurrence of words like 'nazi', 'fckafd' or 'putin'. We strongly assume that the name Putin has a negative connotation in this case since Vladimir Putin invaded Ukraine at that time.

Looking at the HOF distribution by gender we note a significant difference, with male MPs receiving more hate than female MPs. The difference was even higher in the 'Control Dataset'. This is somewhat surprising as it is in contrast to Fuchs and Schäfer (2021) with female Japanese MPs receiving more hate. However it is in line with Theocharis et al. (2020) who investigated the same issue with Members of Congress in the United States. Agarwal et al. (2021) observed that male and female MPs in the UK received equal amounts of offensive texts. A contributing factor to our finding could be the prominence of a (male) key politician (Lauterbach) in the context of the corona crisis. As a highly emotionally discussed topic it attracted a lot of offensive and hateful comments (in particular targeted at individuals such as prominent subject experts).

So the tweets aimed at a single MP do heavily influence the overall distribution of HOF tweets per gender, but this also confirms the 'pile-on' effect that was already observed by (Agarwal et al., 2021) for UK MPs, where MPs often experience a significant increase in online hate when dealing with a high volume of mentions related to a particular event or situation.

Another interesting finding of this work is that offensive and hateful tweets are less viral than non-offensive ones, with fewer likes, replies, or retweets. This contradicts the findings by Mathew et al. (2019) who observed that hate speech tweets tend to spread faster and reach a much wider audience than other content. But they also mentioned that this is mostly the case for verified accounts, and we assume that most accounts in our dataset are not verified.

One last finding worth pointing out is that the assumption by Schmidt et al. (2022) was confirmed, that the general sentiment shifts at specific events. We saw overall less HOF in the 'Control' dataset than in the 'Mentions' dataset. Reasons for this could be that sentiment right before the election was more positive than during the Ukraine war outbreak.

9. Conclusion

Our work is motivated by the fact that social media has developed into a medium of choice to communicate not just personal messages but to contribute to the political discourse with much wider-ranging impacts on society as a whole. While some studies have already investigated the role of politicians in this context we argue that there are still many open research directions. This is even more true when looking at languages other than English. We make several contributions. We provide an **annotated dataset** of 1,250 'X' posts about German MPs which are labeled as containing hateful or offensive language (HOF) or not. We also present an investigation into which **automatic classification** approaches are most promising to annotate a much larger dataset. We identify a transformer-based ensemble offering competitive performance. While our exploration into transfer learning results in variable performance, we also observe that a sanity check on our own data gives an overall satisfactory model performance. This is the basis to annotate larger datasets to conduct a more thorough analysis around the theme of using offensive and hateful tweets **targeting German politicians and parties**. Among our findings we note that male MPs experience significantly more hate than female. We see our work as a stepping stone towards more comprehensive studies in this field, and we hope that our findings will serve as a reference point for that. To foster reproducibility and comparability we also make all sources available via Github.

10. Ethical Considerations and Limitations

Whenever social media data is being processed ethical concerns naturally arise. This is particularly true if the data contains some personal information. Also bias and mitigation play a crucial role in the task of hate speech detection. In addressing bias within hate speech detection, we recognized the need to balance the dataset to counter class imbalances. For data annotation, we experimented with lexicon-based and sentiment-based approaches, with a lexicon-sentiment combination proving more effective. This method could cause bias, however without this method the size of the collection labelled as HOF tweets would be much reduced, so more annotators would have been needed to get a reliable amount of positive samples. Employing ensemble techniques, we curated a diverse model set, aiming to reduce individual model biases and enhance overall fairness. Continuous monitoring and evaluation were crucial, focusing on identifying and rectifying biased predictions.

Despite efforts for proper data collection and annotation, the dataset has limitations due to Twitter API policies restricting data publication. A retrieval script is provided in the GitHub Repository, but it requires time and a Twitter developer account with research access. Additionally, deleted users or tweets, especially HOF tweets pose challenges in reproducing the work. The study acknowledges Twitter's role as one of many social networks, focusing on political discussions. However, it only considers single tweets mentioning MPs, lacking the context of whole conversations.

Generalizing model performance remains challenging due to small test datasets in cross-validation folds. Notably, high-ranking politicians like Anna-Lena Baerbock and Robert Habeck are not included, which could potentially affect the data analysis. Robert Habeck's Twitter account is deactivated, while Anna-Lena Baerbock's username might not have been listed on the Bundestag website during scraping or due to a late-identified error.

Future work should explore large-language models' performance in annotation tasks and investigate their role in generating meaningful synthetic data to enhance model generalizability. Scrutinizing data from different timeframes and events beyond the Russo-Ukrainian war outbreak could provide deeper insights. Moreover, cross-border investigations and topic identification of HOF-tweets are promising avenues for further research.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback.

11. Bibliography

Pushkal Agarwal, Oliver Hawkins, Margarita Amaxopoulou, Noel Dempsey, Nishanth Sastry, and Edward Wood. 2021. Hate speech in political discourse: A case study of UK MPs on Twitter. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 5–16.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.

Dimosthenis Antypas, Alun Preece, and Jose Camacho-Collados. 2023. Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. *Online Social Networks and Media*, 33:100242.

Rafael Bauschke and Sebastian Jäckle. 2023. Hate speech on social media against German mayors: Extent of the phenomenon, reactions, and implications. *Policy & Internet*, 15(2):223–242.

Anat Ben-David and Ariadna Matamoros Fernández. 2016. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, 10:27.

Luca Braghieri, Ro'ee Levy, and Alexey Makarin. 2022. Social media and mental health. *American Economic Review*, 112(11):3660–3693.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for german offensive language and conversation analysis. In *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022), Association for Computational Linguistics, Online*, pages 54–61.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Tamara Fuchs and Fabian Schäfer. 2021. Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter. In *Japan forum*, volume 33, pages 553–579. Taylor & Francis.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France. European Language Resources Association.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Sylvia Jaki and Tom De Smedt. 2019. Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv preprint arXiv:1910.07518*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):1621–1622.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2021. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 29–32, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA. Association for Computing Machinery.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Sünje Paasch-Colberg, Christian Strippel, Joachim Trebbe, and Martin Emmer. 2021. From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, 9(1):171–180.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Thomas Schmidt, Jakob Fehle, Maximilian Weissenbacher, Jonathan Richter, Philipp Gottschalk, and Christian Wolff. 2022. Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 74–87.

Jennifer Scott. 2019. Women MPs say abuse forcing them from politics. *BBC News*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2019. Offensive language and hate speech detection for Danish. *arXiv preprint arXiv:1908.04531*.

Kirill Solovev and Nicolas Pröllochs. 2022. Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *Proceedings of the ACM Web Conference 2022*, pages 3656–3661.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language.

Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. The dynamics of political incivility on Twitter. *Sage Open*, 10(2):2158244020919447.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ethan Xia, Han Yue, and Hongfu Liu. 2021. Tweet sentiment analysis of the 2020 us presidential election. In *Companion proceedings of the web conference 2021*, pages 367–371.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Appendices

In the following, additional tables and plots can be seen. More plots can be found in the GitHub Repository.

Appendix A: Results for the Pilot Modeling Approach, mentioned in ”Experimental Setup”

Model	F1 (Macro)	Precision (Macro)	Recall (Macro)
Electra German Uncased	0.77	0.78	0.76
German Toxicity Classifier	0.77	0.78	0.76
DBMDZ gBERT Uncased	0.75	0.77	0.74
XLM RoBERTa T-Systems	0.74	0.76	0.73
Deepset gBERT Base	0.75	0.75	0.75
gBERT HASOC 2019	0.73	0.75	0.72
XLM RoBERTa Base	0.74	0.74	0.74
Distil gBERT Base	0.73	0.74	0.73
gBERT Cased	0.73	0.74	0.73
DBMDZ gBERT cased	0.72	0.74	0.72
Cardiff XLM RoBERTa Base	0.71	0.74	0.70
mBERT Uncased	0.68	0.70	0.68
mBERT Cased	0.68	0.68	0.69
SVM	0.63	0.69	0.62
LSTM	0.60	0.62	0.59
DeHateBERT German	0.54	0.51	0.58

Table 4: Performance comparison of the models for the pilot approach.

Appendix B: Dataset Balancing Results

Model	F1 (Macro)	Precision (Macro)	Recall (Macro)
Electra German Uncased	0.85	0.85	0.85
German Toxicity Classifier	0.84	0.85	0.84
DBMDZ gBERT Uncased	0.84	0.85	0.84
XLM RoBERTa T-Systems	0.83	0.83	0.83
Deepset gBERT Base	0.74	0.72	0.77

Table 5: Performance comparison of the models for the Balancing Approach.

Appendix C: Transfer Learning Results of GermEval 2019.
The ’Electra German Uncased’ Model from Table 4 would have ranked first. The ’3 Ensemble Hard Voting’ model with the best performance at our work only would have ranked on the 15th place.

Team	**Rank**	**Average**		
		F1	**Precision**	**Recall**
Our Electra German Uncased	1	81.10	81.12	81.08
UPB	2	76.35	77.55	76.95
UPB	3	76.35	77.55	76.95
UPB	4	76.60	77.12	76.86
TUWienKBS	5	77.15	76.45	76.80
TUWienKBS	6	77.01	76.49	76.75
3 Ensemble from Table 3 (Hard Voting)	15	71.70	77.90	69.95

Table 6: Results of GermEval 2019, with the added results from the authors.

Appendix D: Transfer Learning Results of HASOC 2019 - Subtask A
The '3 Ensemble Hard Voting' model (the best-performing model on our datasets) would have been on Rank 1 at HASOC 2019 - Subtask A.

Team	Rank	F1	
		Macro	Weighted
3 Ensemble from Table 3 (Hard Voting)	1	0.6333	0.8055
HateMonitors	2	0.6162	0.7915
LSV-UdS	3	0.6064	0.7997
Our Deepset gBERT base	4	0.6101	0.7965
Our Electra German Uncased	5	0.6070	0.7931
LSV-UdS	6	0.5948	0.7799
3Idiots	7	0.5774	0.7887
NITK-IT_NLP	8	0.5739	0.6796

Table 7: Results of HASOC 2019 - Sub Task A German, with the added results from the authors.

Appendix E: Which MP spreads or receives most hate (politicians dataset)?

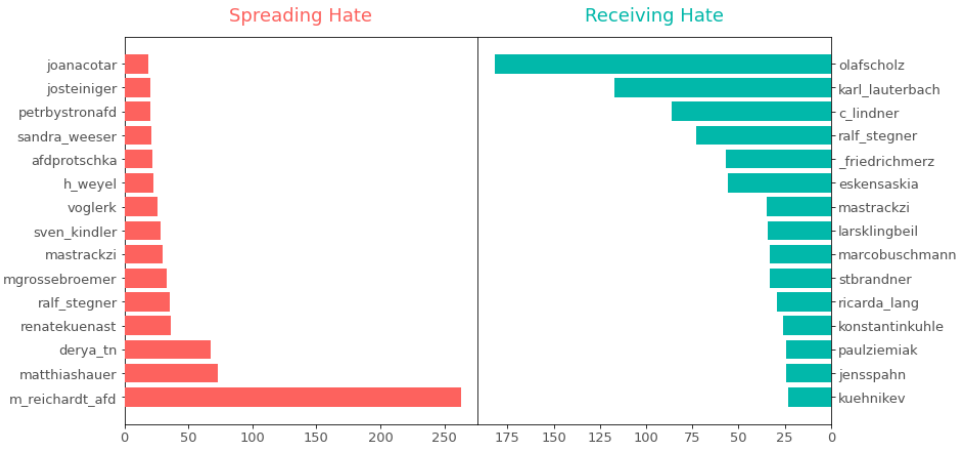

Figure 5: MPs that spread most HOF and MPs that receive most HOF by another MP.