Impact of Politically Biased Data on Hate Speech Classification

Maximilian Wich
TU Munich,
Department of Informatics,
Germany
maximilian.wich@tum.de

Jan Bauer
TU Munich,
Department of Informatics,
Germany
jan.bauer@tum.de

Georg Groh
TU Munich,
Department of Informatics,
Germany
grohg@in.tum.de

Abstract

One challenge that social media platforms are facing nowadays is hate speech. Hence, automatic hate speech detection has been increasingly researched in recent years — in particular with the rise of deep learning. A problem of these models is their vulnerability to undesirable bias in training data. We investigate the impact of political bias on hate speech classification by constructing three politically-biased data sets (left-wing, right-wing, politically neutral) and compare the performance of classifiers trained on them. We show that (1) political bias negatively impairs the performance of hate speech classifiers and (2) an explainable machine learning model can help to visualize such bias within the training data. The results show that political bias in training data has an impact on hate speech classification and can become a serious issue.

1 Introduction

Social media platforms, such as Twitter and Facebook, have gained more and more popularity in recent years. One reason is their promise of free speech, which also obviously has its drawbacks. With the rise of social media, hate speech has spread on these platforms as well (Duggan, 2017). But hate speech is not a pure online problem because online hate speech can be accompanied by offline crime (Williams et al., 2020).

Due to the enormous amounts of posts and comments produced by the billions of users every day, it is impossible to monitor these platforms manually. Advances in machine learning (ML), however, show that this technology can help to detect hate speech — currently with limited accuracy (Davidson et al., 2017; Schmidt and Wiegand, 2017).

There are many challenges that must be addressed when building a hate speech classifier. First of all, an undesirable bias in training data can cause models to produce unfair or incorrect results, such as racial discrimination (Hildebrandt, 2019). This phenomenon is already addressed by the research community. Researchers have examined methods to identify and mitigate different forms of bias, such as racial bias or annotator bias (Geva et al., 2019; Davidson et al., 2019; Sap et al., 2019). But it has not been solved yet; on the contrary, more research is needed Vidgen et al. (2019). Secondly, most of the classifiers miss a certain degree of transparency or explainability to appear trustworthy and credible. Especially in the context of hate speech detection, there is a demand for such a feature Vidgen et al. (2019); Niemann (2019). The reason is the value-based nature of hate speech classification, meaning that perceiving something as hate depends on individual and social values and social values are non-uniform across groups and societies. Therefore, it should be transparent to the users what the underlying values of a classifier are. The demand for transparency and explainability is also closely connected to bias because it can help to uncover the bias.

In the paper, we deal with both problems. We investigate a particular form of bias — political bias — and use an explainable AI method to visualize this bias. To our best knowledge, political bias has not been addressed in hate speech detection, yet. But it could be a severe issue. As an example, a moderator of a social media platform uses a system that prioritizes comments based on their hatefulness to efficiently process them. If this system had a political bias, i.e. it favors a political orientation, it would impair the political debate on the platform. That is why we want to examine this phenomenon by addressing the following two research questions:

RQ1 What is the effect of politically biased data sets on the performance of hate speech classi-

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 54–64

https://doi.org/10.18653/v1/P17

fiers?

RQ2 Can explainable hate speech classification models be used to visualize a potential undesirable bias within a model?

We contribute to answering these two questions by conducting an experiment in which we construct politically biased data sets, train classifiers with them, compare their performance, and use interpretable ML techniques to visualize the differences.

In the paper, we use hate speech as an overarching term and define it as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleby (2000, p.1277), as cited in Schmidt and Wiegand (2017)).

2 Related Work

2.1 Biased Training Data and Models

A challenge that hate speech detection is facing is an undesirable bias in training data (Hildebrandt, 2019). In contrast to the inductive bias — the form of bias required by an algorithm to learn patterns (Hildebrandt, 2019) — such a bias can impair the generalizability of a hate speech detection model (Wiegand et al., 2019; Geva et al., 2019) or can lead to unfair models (e.g., discriminating minorities) (Dixon et al., 2018).

There are different forms of bias. A data set, for example, could have a topic bias or an author bias, meaning that many documents are produced by a small number of authors (Wiegand et al., 2019). Both forms impair the generalizability of a classifier trained on such a biased data set (Wiegand et al., 2019). Another form of bias that has a negative impact on the generalizability of classifiers is annotator bias Geva et al. (2019). In the context of hate speech detection, it is caused by the vagueness of the term hate speech, aggravating reliable annotations (Ross et al., 2017). Waseem (2016), for example, compared expert and amateur annotators — the latter ones are often used to label large data sets. They showed that classifiers trained on annotations from experts perform better. Binns et al. (2017) investigated whether there is a performance difference between classifiers trained on data labeled by males and females. Wojatzki et al. (2018) showed that less extreme cases of sexist speech (a form of hate speech) are differently perceived by women and men. Al Kuwatly et al. (2020) were not able to confirm the gender bias with their experiments, but they discovered bias caused by annotators' age, educational background, and the type of their first language. Another form that is related to annotator bias is racial bias. Davidson et al. (2019) and Sap et al. (2019) examined this phenomenon and found that widely-used hate speech data sets contain a racial bias penalizing the African American English dialect. One reason is that this dialect is overrepresented in the abusive or hateful class (Davidson et al., 2019). A second reason is the insensitivity of the annotators to this dialect (Sap et al., 2019). To address the second problem, Sap et al. (2019) suggested providing annotators with information about the dialect of a document during the labeling process. This can reduce racial bias. Furthermore, Dixon et al. (2018) and Borkan et al. (2019) develop metrics to measure undesirable bias and to mitigate it. To our best knowledge, no one, however, has investigated the impact of political bias on hate speech detection so far.

2.2 Explainable AI

Explainable Artificial Intelligence (XAI) is a relatively new field. That is why we can find only a limited number of research applying XAI methods in hate speech detection. Wang (2018) used an XAI method from computer vision to explain predictions of a neural network-based hate speech classification model. The explanation was visualized by coloring the words depending on their relevance for the classification. Švec et al. (2018) built an explainable hate speech classifier for Slovak, which highlights the relevant part of a comment to support the moderation process. Vijayaraghavan et al. (2019) developed a multi-model classification model for hate speech that uses social-cultural features besides text. To explain the relevance of the different features, they used an attention-based approach. (Risch et al., 2020) compared different transparent and explainable models. All approaches have in common that they apply local explainability, meaning they explain not the entire model (global explanation) but single instances. We do the same because there is a lack of global explainability approaches for text classification.

3 Methodology

Our approach for the experiment is to train hate speech classifiers with three different politically bi-

ased data sets and then to compare the performance of these classifiers, as depicted in Figure 1. To do so, we use an existing Twitter hate speech corpus with binary labels (offensive, non-offensive), extract the offensive records, and combine them with three data sets each (politically left-wing, politically right-wing, politically neutral) implicitly labeled as non-offensive. Subsequently, classifiers are trained with these data sets and their F1 scores are compared. Additionally, we apply SHAP to explain predictions of all three models and to compare the explanations. Our code is available on GitHub[1].

3.1 Topic Modeling

In order to answer our research questions, we need to ensure that the data sets are constructed in a fair and comparable way. Therefore, we use an existing Twitter hate speech corpus with binary labels (offensive, non-offensive) that consists of two data sets as a starting point - GermEval Shared Task on the Identification of Offensive Language 2018 (Wiegand et al., 2018) and GermEval Task 2, 2019 shared task on the identification of offensive language (Struß et al., 2019). Combining both is possible because the same annotation guidelines were applied. Thus, in effect, we are starting with one combined German Twitter hate speech data set. In the experiment, we replace only the non-offensive records of the original data set with politically biased data for each group. To ensure that the new non-offensive records with a political bias are topically comparable to the original ones, we use a topic model. The topic model itself is created based on the original non-offensive records of the corpus. Then, we use this topic model to obtain the same topic distribution in the new data set with political bias. By doing so, we assure the new data sets' homogeneity and topical comparability. The topic model has a second purpose besides assembling our versions of the data set. The keywords generated from each topic serve as the basis of the data collection process for the politically neutral new elements of the data set. More details can be found in the next subsection.

For creating the topic model, we use the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003). A downside of LDA, however, is that it works well for longer documents (Cheng et al., 2014; Quan et al., 2015). But our corpus consists of Tweets that have a maximum length of 280 characters. Therefore, we apply the pooling approach based on hashtags to generate larger documents, as proposed by Alvarez-Melis and Savesk (2016) and Mehrotra et al. (2013).

For finding an appropriate number of topics, we use the normalized pointwise mutual information (NPMI) as the optimization metric to measure topic coherence (Lau et al., 2014). The optimal number of topics with ten keywords each (most probable non-stop words for a topic) is calculated in a 5-fold cross-validation. Before generating the topic model, we remove all non-alphabetic characters, stop words, words shorter than three characters, and all words that appear less than five times in the corpus during the preprocessing. Additionally, we replace user names that contain political party names by the party name, remove all other user names, and apply Porter stemming to particular words[2] (Porter et al., 1980). Only documents (created by hashtag pooling) that contain at least five words are used for the topic modeling algorithm.

3.2 Data Collection

After topic modeling of the non-offensive part from the original data set (without augmentations), we collect three data sets from Twitter: one from a (radical) left-wing subnetwork, one from a (radical) right-wing subnetwork, and a politically neutral one serving as the baseline. All data was retrieved via the Twitter API. The gathering process for these three biased data sets is the following:

1. Identifying seed profiles: First of all, it is necessary to select for each subnetwork seed profiles that serve as the entry point to the subnetworks. For this purpose, the following six profile categories are defined that have to be covered by the selected profiles: politician, political youth organization, young politician, extremist group, profile associated with extremist groups, and ideologized news website. In the category politician, we select two profiles for each subnetwork — one female and one male. The politicians have similar positions in their parties, and their genders are balanced. For the category political youth organization, we took the official Twitter profiles from the political youth organizations of the parties that the politicians from the previous category are a member of. In the cate-

[1] https://github.com/mawic/political-bias-hate-speech

[2] *Frauen, Männer, Linke, Rechte, Deutschland, Nazi, Jude, Flüchtling, Grüne*

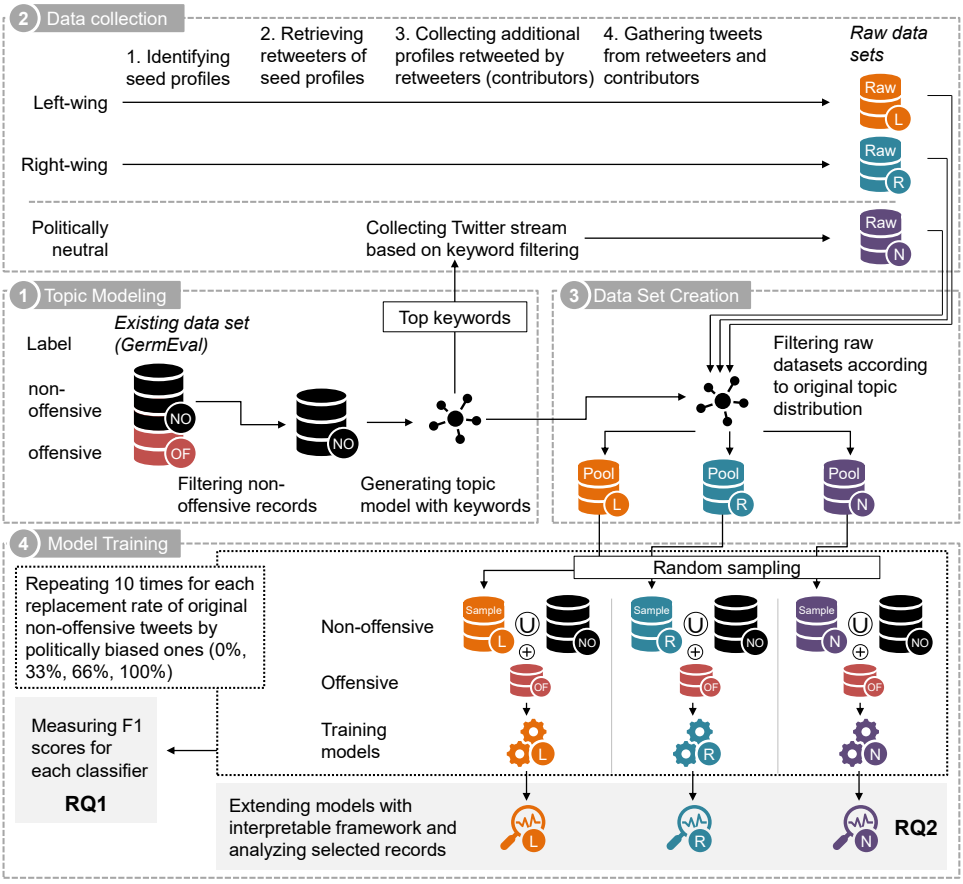


Figure 1: Methodological approach visualized

gory young politician, we selected one profile of a member from the executive board of each political youth organization. For the extremist group, we use official classifications of official security agencies to identify one account of such a group for each subnetwork. Concerning the category profile associated with extremist groups, we select two accounts that associate with an extremist group according to their statements. The statements come from the description of the Twitter account and from an interview in a newspaper. In regards to the ideologized news website, we again rely on the official classifications of a federal agency to choose the Twitter accounts of two news websites. We ensure for all categories that the numbers of followers of the corresponding Twitter accounts are comparable. The seven profiles for each subnetwork are identified based on explorative research.

2. Retrieving retweeters of seed profiles: After identifying the seven seed Twitter profiles for each political orientation as described in the previous paragraph, we are interested in the profiles that retweet these seed profiles. Our assumption in this context is that retweeting expresses agreement concerning political ideology, as shown by Conover et al. (2011a), Conover et al. (2011b), and Shahrezaye et al. (2019). Therefore, the retweets of the latest 2,000 tweets from every seed profile are retrieved - or the maximum number of available tweets, if the user has not tweeted more. Unfortunately, the Twitter API provides only the latest 100 retweeters of one tweet. But this is not a problem because we do not attempt to crawl the entire subnetwork. We only want to have tweets that are representative of each subnetwork. After collecting these retweets, we select those of their authors (retweeters) that retweeted at least four of the seven seed profiles. We do this because we want to avoid adding profiles that retweeted the seed profiles but are not clearly part of the ideological subnetwork. Additionally, we remove retweeters that appear in both subnetworks to exclude left-wing accounts retweeting right-wing tweets or vice versa. Moreover, we eliminate verified profiles. The motivation of deleting verified profiles is that these profiles are ran by public persons or institutions and Twitter has proved their authenticity. This transparency might influence the language the users use for this

profile.

3. Collecting additional profiles retweeted by retweeters (contributors): Step 3 aims to gather the profiles (contributors) that are also retweeted by the retweeters of the seed profiles. Therefore, we retrieve the user timelines of the selected retweeters (output of step 2) to get their other retweets. From these timelines, we select those profiles that have been retweeted by at least 7.5% of the retweeters. This threshold is pragmatically chosen — in absolute numbers 7.5% means more than 33 (left-wing) and 131 (right-wing) retweeters. The reason for setting a threshold is the same one as in step 2. Besides that, profiles appearing on both sides and verified ones are also deleted.

4. Gathering tweets from retweeters and contributors: Additionally to the gathered user timelines from step 3, we collect the latest 2,000 tweets from the selected contributors (step 3), if they are available. Furthermore, the profiles of selected retweeters (step 2) and selected contributors (step 3) are monitored via the Twitter Stream API for a few weeks to collect additional tweets.

The politically neutral data set is collected by using the Twitter Stream API. It allows us to stream a real-time sample of tweets. To make sure to get relevant tweets, we filtered the stream by inputting the keywords from the topic model we have developed. Since the output of the Stream API is a sample of all publicly available tweets (Twitter Inc., 2020), we can assume that the gathered data is not politically biased. The result of the data collection process is a set of three raw data sets - one with a left-wing bias, one with a right-wing bias, and one politically neutral.

3.3 Data Set Creation

Having the topic model and the three raw data sets, we can construct the pool data sets that exhibit the same topic distribution as the original non-offensive data set. They serve as pools for non-offensive training data that the model training samples from, described in the next sub-section. Our assumption to label the politically biased tweets as non-offensive is the following: Since the tweets are available within the subnetwork, they conform to the norms of the subnetwork, meaning the tweets are no hate speech for its members. Otherwise, members of the subnetwork could have reported these tweets, leading to a deletion in case of hate speech. The availability of a tweet, however, does not imply that they conform to the norms of the medium. A tweet that complies with the norms of the subnetwork, but violates the ones of the medium could be only distributed within the subnetwork and does not appear in the feed of other users. Consequently, it would not be reported and still be available.

We compose the pool data sets according to the following procedure for each politically biased data set: In step 1, the generated topic model assigns every tweet in the raw data sets a topic, which is the one with the highest probability. In step 2, we select so many tweets from each topic that the following conditions are satisfied: Firstly, the size of the new data is about five times the size of the non-offensive part from the GermEval corpus. Secondly, tweets with a higher topic probability are chosen with higher priority. Thirdly, the relative topic distribution of the new data set is equal to the one of the non-offensive part from the GermEval corpus. The reason for the increased size of the three new data sets (the three pool data sets) is that we have enough data to perform several iterations in the phase Model Training in order to contribute to statistical validity.

3.4 Model Training

In the phase Model Training, we train hate speech classifiers with the constructed data sets to compare performance differences and to measure the impact on the F1 score (RQ1). Furthermore, we make use of the ML interpretability framework SHAP to explain generated predictions and visualize differences in the models (RQ2).

Concerning the RQ1, the following procedure is applied. The basis is the original training corpus consisting of the union of the two GermEval data sets. For each political orientation, we iteratively replace the non-offensive tweets with the ones from the politically biased data sets (33%, 66%, 100%). The tweets from the politically biased data sets are labeled as non-offensive.

For each subnetwork (left-wing, right-wing, politically neutral) and each replacement rate (33%, 66%, 100%), ten data sets are generated by sampling from the non-offensive part of the original data set and the respective politically biased pool data set and leaving the offensive part of the original data set untouched. We then use these data sets to train classifiers with 3-fold cross-validation. This iterative approach produces multiple observa-

tion points, making the results more representative — for each subnetwork and each replacement rate we get $n = 30$ F1 scores. To answer RQ1, we statistically test the hypotheses, (a) whether the F1 scores produced by the politically biased classifiers are significantly different and (b) whether the right-wing and/or left-wing classifier performs worse than the politically neutral one. If both hypotheses hold, we can conclude that political bias in training data impairs the detection of hate speech. The reason is that the politically neutral one is our baseline due to the missing political bias, while the other two have a distinct bias each. Depending on the results, we might go one step further and might infer that one political orientation diminishes hate speech classification more substantially than the other one. For this, we use the two-sided Kolmogorov-Smirnov test (Selvamuthu and Das, 2018). The null hypothesis is that the three distributions of F1 scores from three sets of classifiers are the same. The significance level is $p < 0.01$. If the null hypothesis is rejected, which confirms (a), we will compare the average F1 scores of each distribution with each other to answer (b).

The classifier consists of a non-pre-trained embedding layer with dimension 50, a bidirectional LSTM comprising 64 units, and one fully connected layer of the size 16. The output is a sigmoid function classifying tweets as offensive or not. We used Adam optimization with an initial learning rate of 0.001 and binary cross-entropy as a loss function. We applied padding to each tweet with a maximal token length of 30. As a post-processing step, we replaced each out-of-vocabulary token occurring in the test fold with an <unk> token to overcome bias and data leaking from the test data into the training data.

In regards to RQ2, we apply the following procedure. We select one classifier from each subnetwork that is trained with an entirely replaced non-offensive data set. To explain the generated predictions, we apply the DeepExplainer from the SHAP framework for each classifier (Lundberg and Lee, 2017). After feeding DeepExplainer with tweets from the original corpus ($n = 1000$) to build a baseline, we can use it to explain the predictions of the classifiers. An explanation consists of SHAP values for every word. The SHAP values "attribute to each feature the change in the expected model prediction when conditioning on that feature" (Lundberg and Lee, 2017, p. 5). Comparing the SHAP values from the three different classifiers for a selected word in a tweet indicates how relevant a word is for a prediction w.r.t. to a specific class (e.g., offensive, non-offensive). Figure 3a shows how these values are visualized. This indication, in turn, can reveal a bias in the training data. Therefore, we randomly select two tweets from the test set that are incorrectly classified by the left-wing, respectively right-wing classifier and compare their predictions to answer RQ2.

4 Results

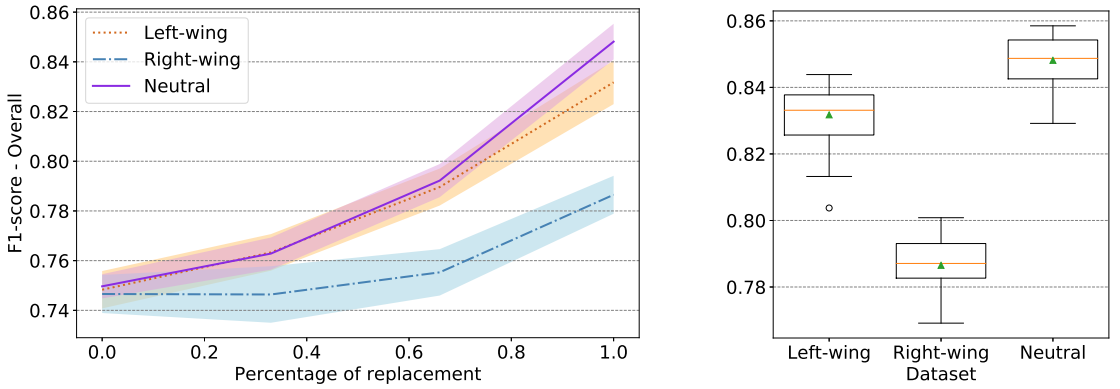
4.1 Data

The two GermEval data sets are the basis of the experiment. In total, they contain 15,567 German tweets - 10,420 labeled as non-offensive and 5,147 as offensive. The data for the (radical) left-wing subnetwork, the (radical) right-wing one, and the neutral one was collected via the Twitter API between 29.01.2020 and 19.02.2020. We gathered 6,494,304 tweets from timelines and 2,423,593 ones from the stream for the left-wing and right-wing subnetwork. On average, 1,026 tweets ($median = 869; \sigma^2 = 890.48$) are collected from 3,168 accounts. For the neutral subnetwork, we streamed 23,754,616 tweets. After removing retweets, duplicates, tweets with less than three tokens, and non-German tweets, we obtain 1,007,810 tweets for the left-wing raw data set, 1,620,492 for the right-wing raw data set, and 1,537,793 for the neutral raw data set. 52,100 tweets of each raw data set are selected for the data pools according to the topic model and the topic distribution. The input for the 3-fold cross-validation of the model training consists of the 5,147 offensive tweets from GermEval and 10,420 non-offensive ones from GermEval or the collected data depending on the replacement rate.

4.2 Results

All three classifiers show significantly ($p < 0.01$) different F1 scores. The one with the worst performance is the one trained with the right-wing data set (78.7%), followed by the one trained with the left-wing data set (83.1%) and the politically neutral one (84.8%).

Figure 2a shows how the F1 scores change depending on the replacement rate. The lines are the average F1 scores of the three classifiers, and the areas around them are the standard deviation of the multiple training iterations. At first glance,



(a) F1 scores according to replacement rate

(b) F1 scores at 100% replacement rate

Figure 2: F1 scores of the three classifier subnetworks

the political biases in the data seem to increase the performance due to the improvement of the F1 scores. This trend, however, is misleading. The reason for the increase is that the two classes, offensive and non-offensive, vary strongly with the growing replacement rate, making it easier for the classifiers to distinguish between the classes. More relevant to our research question, however, are the different steepnesses of the curves and the emerging gaps between them. These differences reveal that it is harder for a classifier trained with a politically biased data set to identify hate speech - particularly in the case of a right-wing data set. While the neutral and left-wing curves are nearly congruent and only diverge at a 100% replacement rate, the gap between these two and the right-wing curve already occurs at 33% and increases. Figure 2b visualizes the statistical distribution of the measured F1 scores at a 100% replacement rate as box plots. The Kolmogorov-Smirnov test confirms the interpretation of the charts. The distributions of the left-wing and politically neutral data set are not significantly different until 100% replacement rate — at 100% $p = 8.25 \times 10^{-12}$. In contrast to that, the distribution of the right-wing data set already differs from the other two at 33% replacement rate — at 33% left- and right-wing data set $p = 2.50 \times 10^{-7}$, right-wing and neutral data set $p = 6.53 \times 10^{-9}$ and at 100% left- and right-wing data set $p = 1.69 \times 10^{-17}$, right-wing and neutral data set: $p = 1.69 \times 10^{-17}$. Thus, we can say that political bias in a training data set negatively impairs the performance of a hate speech classifier, answering RQ1.

To answer RQ2, we randomly pick two offensive tweets that were differently classified by the three interpretable classifiers. Subsequently, we compare the explanations of the predictions from three different classifiers. These explanations consist of SHAP values for every token that is fed into the classifier. They indicate the relevance of the tokens for the prediction. Please note: not all words of a tweet are input for the classifier because some are removed during preprocessing (e.g., stop words). A simple way to visualize the SHAP values is depicted in Figure 3a. The model output value is the predicted class probability of the classifier. In our case, it is the probability of how offensive a tweet is. The words to the left shown in red (left of the box with the predicted probability) are responsible for pushing the probability towards 1 (offensive), the ones to the right shown in blue (right of the box) towards 0 (non-offensive). The longer the bars above the words are, the more relevant the words are for the predictions. Words with a score lower than 0.05 are not displayed.

Figure 3a shows the result of the three interpretable classifiers for the following offensive tweet: @<*user*>@<*user*> *Natürlich sagen alle Gutmenschen 'Ja', weil sie wissen, dass es dazu nicht kommen wird.* (@<*user*>@<*user*> *Of course, all do-gooders say "yes", because they know that it won't happen.*)

The left-wing and neutral classifiers predict the tweet as offensive (0.54, respectively 0.53), while the right-one considers it non-offensive (0.09). The decisive factor here is the word *Gutmenschen*. *Gutmensch* is German and describes a person "who is, or wants to be, squeaky clean with respect to morality or political correctness" (PONS, 2020). The word's SHAP value for the right-wing classifier is 0.09, for the left-wing one 0.45, and for the

neutral one 0.36. It is not surprising if we look at the word frequencies in the three different data sets. While the word *Gutmensch* and related ones (e.g., plural) occur 38 times in the left-wing data set and 39 times in the neutral one, we can find it 54 times in the right-wing one. Since mostly (radical) right-wing people use the term *Gutmensch* to vilify political opponents (Hanisch and Jäger, 2011; Auer, 2002), we can argue that differences between the SHAP values can indicate a political bias of a classifier.

Another example of a tweet that one politically biased classifier misclassifies is the following one (see Figure 3b): @<*user*>@<*user*> *Hätte das Volk das recht den Kanzler direkt zu wählen, wäre Merkel lange Geschichte.* (*If the people had the right to elect the chancellor directly, Merkel would have been history a long time ago.*)

The right-wing (0.10) and neutral classifiers (0.35) correctly classify the tweet as non-offensive, but not the left-wing one (0.96). All three have in common that the words *Volk* (German for people) and *Merkel* (last name of the German chancellor) favoring the classification as offensive, but with varying relevance. For the right-wing classifier, both terms have the lowest SHAP values (*Volk*: 0.05, *Merkel*: 0.04); for the neutral classifier, the scores are 0.34 (*Volk*) and 0.16 (*Merkel*); for the left-wing classifier, they are 0.14 (*Volk*) and 0.31 (*Merkel*). The low values of the right-wing classifier can be explained with relative high word frequency of both terms in the non-offensive training set. Another interesting aspect is that the term *Kanzler* (chancellor) increases the probability of being classified as offensive only in the case of a left-wing classifier (SHAP value: 0.08). We can trace it back to the fact that the term does not appear in the non-offensive part of the left-wing data set, causing the classifier to associate it with hate speech. This example also shows how a political bias in training data can cause misleading classifications due to a different vocabulary.

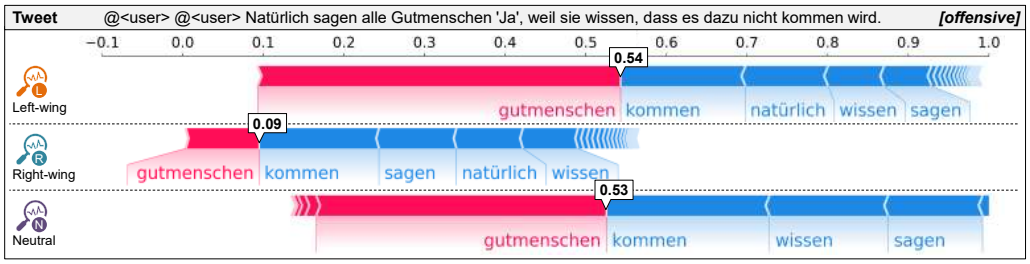
5 Discussion

The experiment shows that the politically biased classifiers (left- and right-wing) perform worse than the politically neutral one, and consequently that political bias in training data can lead to an impairment of hate speech detection (RQ1). In this context, it is relevant to consider only the gaps between the F1 classifiers' scores at 100% replacement rate. The gaps reflect the performance decrease of the politically biased classifiers. The rise of the F1 scores with an increasing replacement rate is caused by the fact that the new non-offensive tweets are less similar to the offensive ones of the original data set.

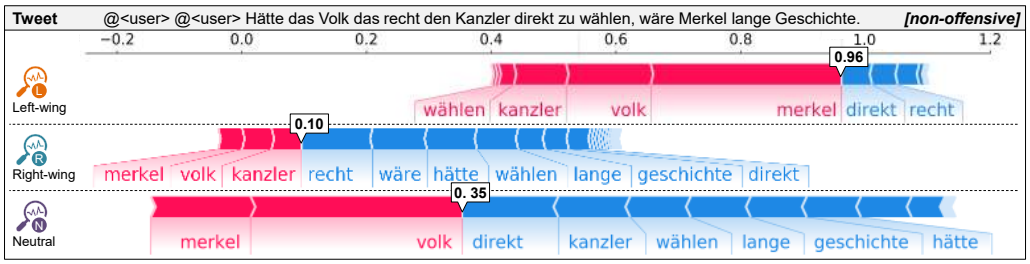
The results also indicate that a right-wing bias impairs the performance more strongly than a left-wing bias. This hypothesis, however, cannot be confirmed with the experiment because we do not have enough details about the composition of the offensive tweets. It could be that right-wing hate speech is overrepresented in the offensive part. The effect would be that the right-wing classifier has more difficulties to distinguish between offensive and non-offensive than the left-wing one even if both data sets are equally hateful. The reason is that the vocabulary of the right-wing data set is more coherent. Therefore, this hypothesis can neither be confirmed nor rejected by our experiment.

Concerning RQ2, we show that explainable ML models can help to identify and to visualize a political bias in training data. The two analyzed tweets provide interesting insights. The downside of the approach is that these frameworks (in our case SHAP) can only provide local explanations, meaning only single inputs are explained, not the entire model. It is, however, conceivable that the local explanations are applied to the entire data set, and the results are aggregated and processed in a way to identify and visualize bias. Summing up, this part of the experiment can be seen rather as a proof-of-concept and lays the foundation for future research.

Regarding the overall approach of the experiment, one may criticize that we only simulate a political bias by constructing politically biased data sets and that this does not reflect the reality. We agree that we simulate political bias within data due to the lack of such data sets. Nevertheless, we claim the relevance and validity of our results due to the following reasons: Firstly, the offensive data part is the same for all classifiers. Consequently, the varying performances are caused by non-offensive tweets with political bias. Therefore, the fact that the offensive tweets were annotated by annotators and the non-offensive tweets were indirectly labeled is less relevant. Furthermore, any issues with the offensive tweets' annotation quality do not play a role because all classifiers are trained and tested on the same offensive tweets. Secondly, we con-



(a) Tweet wrongly classified by right-wing classifier



(b) Tweet wrongly classified by left-wing classifier

Figure 3: SHAP values for the two selected tweets

struct the baseline in the same way as the left- and right-wing data set instead of using the original data set as the baseline. This compensates confounding factors (e.g., different time, authors). Thirdly, we use a sophisticated topic-modeling-based approach to construct the data sets to ensure the new data sets' topic coherence.

6 Conclusion

We showed that political bias in training data can impair hate speech classification. Furthermore, we found an indication that the degree of impairment might depend on the political orientation of bias. But we were not able to confirm this. Additionally, we provide a proof-of-concept of visualizing such a bias with explainable ML models. The results can help to build unbiased data sets or to debias them. Researchers that collect hate speech to construct new data sets, for example, should be aware of this form of bias and take our findings into account in order not to favor or impair a political orientation (e.g., politically balanced set of sources). Our approach can be applied to identify bias with XAI in existing data sets or during data collection. With these insights, researchers can debias a data set by, for example, adjusting the distribution of data. Another idea that is fundamentally different from debiasing is to use these findings to build politically branded hate speech filters that are marked as those. Users of a social media platform, for example, could choose between such filters depending on their preferences. Of course, obvious hate speech would be filtered by all classifiers. But the classifiers would treat comments in the gray area of hate speech depending on the group's norms and values.

A limitation of this research is that we simulate the political bias and construct synthetic data sets with offensive tweets annotated by humans and non-offensive tweets that are only implicitly labeled. It would be better to have a data set annotated by different political orientations to investigate the impact of political bias. But such an annotating process is very challenging. Another limitation is that the GermEval data and our gathered data are from different periods. We, however, compensate this through our topic modeling-based data creation.

Nevertheless, political bias in hate speech data is a phenomenon that researchers should be aware of and that should be investigated further. All in all, we hope that this paper contributes helpful insights to the hate speech research and the fight against hate speech.

Acknowledgments

This paper is based on a joined work in the context of Jan Bauer's master's thesis (Bauer, 2020). This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proc. 4th Workshop on Online Abuse and Harms*.

David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *10th Intl. AAAI Conf. Weblogs and Social Media*.

Katrin Auer. 2002. Political Correctness – Ideologischer Code, Feindbild und Stigmawort der Rechten. *Österreichische Zeitschrift für Politikwissenschaft*, 31(3):291–303.

Jan Bauer. 2020. Political bias in hate speech classification. Master's thesis, Technical Univesiy of Munich. Advised and supervised by Maximilian Wich and Georg Groh.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proc. 28th WWW Conf.*, pages 491–500.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011a. Predicting the political alignment of twitter users. In *2011 IEEE 3rd Intl. Conf. Privacy, Security, Risk, and Trust and 2011 IEEE 3rd Intl. Conf. Social Computing*, pages 192–199.

Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011b. Political polarization on twitter. In *5th Intl. AAAI Conf. Weblogs and Social Media*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. 11th ICWSM Conf.*

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proc. 2018 AAAI/ACM Conf. AI, Ethics, and Society*, pages 67–73.

Maeve Duggan. 2017. *Online harassment 2017*. Pew Research Center.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 1161–1166.

Astrid Hanisch and Margarete Jäger. 2011. Das Stigma "Gutmensch". *Duisburger Institut für Sprach-und Sozialforschung*, 22.

Mireille Hildebrandt. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, 20(1):83–121.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892.

Marco Niemann. 2019. Abusiveness is non-binary: Five shades of gray in german online newscomments. In *IEEE 21st Conference Business Informatics*, pages 11–20.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

PONS. 2020. Gutmensch - Deutsch-Englisch Übersetzung — PONS.

Martin F Porter et al. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proc. Workshop on Trolling, Aggression and Cyberbullying (TRAC@LREC)*, pages 137–143.

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*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. 57th ACL Conf.*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. 5th Intl. Workshop on Natural Language Processing for Social Media*, pages 1–10.

Dharmaraja Selvamuthu and Dipayan Das. 2018. *Introduction to statistical methods, design of experiments and statistical quality control*. Springer.

Morteza Shahrezaye, Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. 2019. Estimating the political orientation of twitter users in homophilic networks. In *AAAI Spring Symposium: Interpretable AI for Well-being*.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proc. 15th KONVENS*, pages 354–365.

Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2018. Improving Moderation of Online Discussions via Interpretable Neural Models. In *Proc. 2nd Workshop on Abusive Language Online*, pages 60–65.

Twitter Inc. 2020. Sample stream - Twitter Developers. https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample.

Bertie Vidgen, Rebekah Tromble, Alex Harris, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proc. 3rd Workshop on Abusive Language Online*, pages 80–93.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2019. Interpretable Multi-Modal Hate Speech Detection. In *Intl. Conf. Machine Learning AI for Social Good Workshop*.

Cindy Wang. 2018. Interpreting neural network hate speech classifiers. In *Proc. 2nd Workshop on Abusive Language Online*, pages 86–92.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proc. 1st Workshop on NLP and Computational Social Science*, pages 138–142.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–608.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proc. 14th KONVENS*.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proc. 14th KONVENS*.