Detecting and Reducing Youth Language Bias in Hate Speech Detection

Anonymous ACL submission

Abstract

With the increase of adolescents and children active online, it is of importance to evaluate the algorithms which are designed to protect them from physical and mental harm. This work measures the bias introduced by youth language on hate speech detection models. The research constructs a novel framework to identify language bias within trained networks. It introduces a technique to detect emerging hate phrases and evaluates the unintended bias attached to them. The research focuses specifically on slurs used in hateful speech. Therefore, three bias test sets are constructed: one for the emerging hate speech terms, one for established hate terms, and one to test for overfitting. Based on the test sets, three scientific and one commercial hate speech detection model are evaluated and compared. For evaluation, the research introduces a novel Youth Language Bias Score. Lastly, the research applies fine-tuning as a mitigation strategy for youth language bias and trains and evaluates the newly trained classifier. The research introduces a novel framework for bias detection, identifies that the language used by adolescents has influence on the performance of the classifiers in hate speech classification, and provides the first hate speech classifier specifically trained for online youth language.

1 Introduction

In the physical world, children and adolescents have the right to mature free from negative influences. In Germany and the European Union, this applies directly to the digital world.[1] At a time where the majority of children have access to the internet and their own devices (Rohleder, 2022), this right needs to be evermore protected. With the influential role social media plays in the development of children and the vast amounts of hate speech present in social media (McCarthy, 2020), the need for efficient and accurate working mechanisms to protect adolescents from online hate speech becomes clear. To handle this complex problem, artificial intelligence used for algorithmic hate speech detection is a viable option. Systems involving natural language processing are required to be algorithmically fair and fitted within different social groups (Blodgett and O'Connor, 2017).

While most models have some sort of indented bias – for example being eager to detect hate speech instead of non-hate speech – unintended biases can negatively influence the performance of the system (Dixon et al., 2018). It has been shown that different unintended biases exist (e.g.: gender, racial, topic, author bias) and have influence on the accuracy of the trained algorithms (Röttger et al., 2021). The change in time and topic of an online conversation has been shown to have a great effect on algorithmic hate speech detection (Florio et al., 2020).

This research wants to raise awareness for the understudied field of youth language (YL) as the source for bias, which reduces the performance of hate speech detection classifiers. The language used by adolescents varies compared to the language used by adults (Schwartz et al., 2013). This research goes further than establishing that lexical topical change has influence on hate speech classifiers. A novel framework for youth bias evaluation is provided. It innovatively describes the process of how to identify the bias introduced by the age group to hate speech classification and further shows how to mitigate the bias in an existing model. This new bias field is understudied due to the difficulty of obtaining age annotated data.All language is in a state of change, due to the fast-changing and widely different character of the youth language it must be differentiated between age groups. The need to protect adolescents from harmful influences makes it important to evaluate commonly used hate speech detection models.

[1] https://www.kjm-online.de/themen/jugendmedienschutz

2 Related Work

Due to issues as regulations and safety concerns, it is difficult to construct data sets in the realm of hate speech and youth language. However, within the subfield of cyberbullying, researchers have delved into relevant investigations. Notably, Sprugnoli et al. (2018) undertook a study focusing on teens, which constructed a dataset from chat conversations among Italian school students. Menini et al. (2019) devised a cyberbullying monitoring system in the United Kingdom. They pinpointed multiple high schools on Instagram, along with their students and friends. Meanwhile, Wijesiriwardene et al. (2020) established a Twitter-based multimodal dataset. This dataset concentrated on toxic interactions and involved American high school students, identified manually. Additionally, Bayzick et al. (2011) introduced a dataset which was comprised of chat conversations originating from MySpace.com. This dataset also included self-reported author age information. In more recent times, Fillies et al. (2023) amassed an English hate speech dataset from annotated Discord messages between teenagers. Age identification drew from a subset of users who volunteered their age information.

In the field of named entity recognition (NER) the main objective is to recognize named entities in text (Ling et al., 2015). Entity Linking focuses on connecting the new discovered entities to an underlaying concept (Hoffart et al., 2014). NER was first based on static vocabularies and rules, but has seen a shift towards more advanced transformer based solutions (Heist and Paulheim, 2022). Färber et al. (2016) identified different groups of emerging entities. Different approaches exist to identify entities, such as by connecting contexts and entities (Akasaki et al., 2019) or identifying emerging entities by validating that they are not reflected in a corpus (Derczynski et al., 2017) or connected knowledge base (Nakashole et al., 2013).

Utilizing pre-trained models to detect hate speech is a common practice and often yields accurate predictions. However, it's been demonstrated on multiple occasions that these models, along with other classification algorithms, can exhibit biases toward minority groups. Instances of bias tied to gender (Kurita et al., 2019) and race (Kennedy et al., 2020) have been extensively examined. Although age has been recognized as a potential source of bias in data (Hovy and Prabhumoye, 2021), its impact on pre-trained networks remains underexplored. Furthermore, different factors such as topic (Wiegand et al., 2019; Justen et al., 2022), author (Nejadgholi and Kiritchenko, 2020), and time (Justen et al., 2022) have been shown to have influence on prediction quality. To mitigate these biases, diverse approaches have emerged. Some center on specific domains or tasks by fine-tuning the models with new data (Park et al., 2018; Zhang et al., 2018; Beutel et al., 2017). New and promising approaches, as seen by Cai et al. (2022), are considering feature importance during training to mitigate bias.

For evaluating different types of bias within a trained model, or the underlying data set, different strategies were established. The most influential strategies are explained in more detail: On the one hand, following Dixon et al. (2018), one common approach is to create a positive and negative balanced test set of identity terms. In their research,Dixon et al. (2018) populated the data set with a range of chosen identity terms. By evaluating the ability of a model to classify the test set correctly, unintended bias towards identity terms can be shown. The work of Röttger et al. (2021) and Röttger et al. (2022) goes further by not just focusing on identity terms by creating a functional test covering 29 model functionalities ranging from slurs to identity terms. Röttger et al. (2021)test cases were human annotated.

3 Research Design

3.1 Definitions Hate Speech and Slurs

The definition of Founta et al. (2018) was chosen due to its inclusion of specific characteristics and reference to different linguistic styles. However, it has to be adapted to fit the following research, extending it to include individuals. This change is supported by the definition of the European Council and is necessary due to the focus on slurs, which are often based on group discrimination but directed against individuals. The definition in this work is: "Hate speech is language that attacks or diminishes, that incites violence or hate against groups [or individuals], based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur in different linguistic styles, even in subtle forms or when humor is used".

This research considers a slur as hateful. Slurs are terms that are used to insult or harm another person, and they can be based on different at-

tributes such as ethnicity or physical attributes, e.g. "Kanake" is a hatful slur used in Germany towards Turkish people. In contrast, "Turks" would be an identity term. It is acknowledged that slurs can also be reclaimed and therefore not hateful, but the context necessary to make this judgment will not be provided to the models in question, therefore making the statements always hateful.

3.2 Definition Youth Language

Bahlo et al. (2019) identified two main ways to characterize youth language. Firstly, it can be seen as a common systematic core of language that is shared between all adolescents (Bahlo et al., 2019). The systematic core identifies youth languages as the linguistic style used by a generation to differentiate themselves to other age and social groups. For this they share a common trait, such as interests, social activities, or friendships.

Secondly, youth language can be defined as group characteristic variations (Bahlo et al., 2019). Here the approach is the possibility to describe youth language as a variety of language itself, reducing youth language to three levels: linguistic structure, linguistic context, and nonlinguistic dimensions, which are defined as location, group identities, situation, and time.

To summarize, the two approaches differ in their focus: one emphasizes the variation of language present in the system while the other defines it based on the speaker's perspective, describing the language as a small subgroup specific style of conversation. Following the first approach, this research views youth language as a variation of language that can be found within the present language of adolescents, while acknowledging that there are group characteristic variations in any setting.

3.3 Bias and Fairness

Due to the inherent nature of solving a specific task on specific data, every machine learning model contains bias. Dixon et al. (2018) establish fairness as "a potential negative impact on society, and in particular when different individuals are treated differently." Dixon et al. (2018) defines their unintended bias as a "model contains unintended bias if it performs better for comments about some groups than for comments about others groups." The definition of bias is based on the concept of equalized odds by Hardt et al. (2016). This research extends on this work by abstracting and defining: a model contains unintended bias if it performs more precise for comments from some groups and about certain groups than for comments from and about other groups, groups can be based on demographic features, location, group identities, situation, and time. This new definition can be used to establish multiple biases, as done in this research regarding the age of the authors represented in the topics and linguistic features of their language.

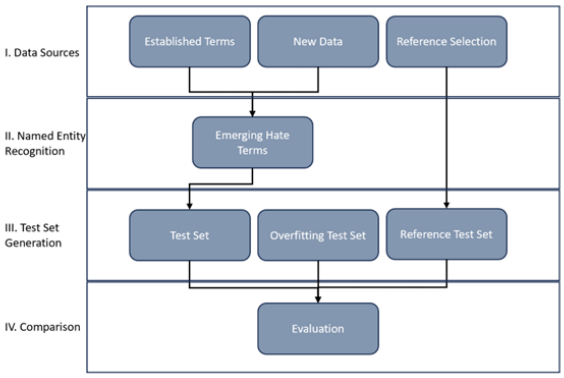

Figure 1: Framework for Bias Detection

3.4 Framework Architecture

The framework consists of a multistep process, combining multiple state-of-the-art research approaches out of the fields of emerging entity detection, functional testing of hate speech classifiers, and bias detection methods. Graphic 1 gives a general overview of the process.

The framework consists of four layers wherein each layer describes a state of the framework. The top layer is the input data state. The Framework needs three input sources. First, an annotated data set featuring hate speech language from a specific age group, in this case youth language is required. Secondly, a collection of existing and established hate speech terms is needed. Thirdly, a smaller subsection of established hate speech terms as a reference section is necessary.

In stage two, the new data is analyzed in regards to the new hate speech terms. NER is applied to all, as hate speech classified, terms excluding the already known and established terms. To further identify words that connect to hateful content, only terms are selected that are statistically more often seen in the hateful class then in the non-hateful class. This reduces examples with wrong labels assigned during annotation. These newly discovered terms then can be weighted by number of appearances. This creates a ranking within the established

terms, therefore excluding singletons (emerging words only used once and specific to an event or context). These terms do need human evaluation to explicitly choose slurs and separate them from e.g. identity terms. They also need to be validated to be hateful on their own.

After a set of emerging slurs have been established, these terms are used to build the new test set in the third layer. Here, the existing neutral templates from Röttger et al. (2021) are used (e.g. "Her colleague is a [SLUR]") and extended by the newly created templates (e.g. "I know a [SLUR] when I see one"). The same procedure is done to the established hateful slurs test set as a baseline.

The test sets are then used to evaluate the prediction quality of the chosen classifiers. In the optimal case, the model should have a similar classification performance of the new and old terms. The difference in accuracy is classified as a possible youth language bias. Lastly, in the third stage, using on non-hateful terms, the test set against overfitting is introduced and is used in the fourth step to validate a model towards possible overfit on the template cases used in the creation of both test sets.

3.5 Data Sets

The data set by Fillies et al. (2023) was selected, it provides a hate speech data set in English containing annotated discord messages between teenagers, in which age identification relied upon information that was voluntarily provided by the users. It was collected during March 2021 and June 2022 and contains anonymized hate speech youth language, consisting of 88.395 annotated chat messages. For 35.553 messages, there are age annotations provided, averaging the author age to under 20 years of age. 6,42% of the total messages were classified as hate speech. This data set is the source of emerging hate speech terms and further validates the performance of existing classifiers. In regards to the list of established terms, this research refers to a publicly available list of 1600+ popular English profanities including possible variations.[2].

The reference selection of well-established hate terms are taken from the research of Röttger et al. (2021), as they define a list of 18 slurs. These terms were then filtered to the top 10 terms (see Appendix A), all are included in the established terms list.

[2]https://github.com/surge-ai/profanity

4 Emerging Entity Detection

4.1 Methodology

To detecting emerging entities this research follows the work by Färber et al. (2016). The approach identifies entities that are not already existing in their knowledge graph. Within this research, the knowledge graph is replaced by the list of established terms, as a knowledge base. This research advances further by introducing an association score (called the Class Hate Score (CHS)). This score calculates how often a word (w) appears within a hateful context or within a non-hateful context, it classifies an entity as hateful the closer it is associated to the non-hateful class. The hate score is determined by calculating the frequency (cntnohate) of the word in the total amount of non-hateful (nohate) and then dividing it by the words frequency (cnthate) as a percentage of the total number of words in the hateful (hate) content. The higher the association with the hateful class, the closer the score is to zero. The closer the score is to 1, the more the word is associated to a non-hateful context.

$$CHS(w) = \frac{(\frac{cntnohate(w)}{nohate})}{(\frac{cnthate(w)}{hate})} \quad (1)$$

4.2 Experiments

The experiment went through seven stages and is written as a python script. After the data set was selected, the first step of combining the annotations into a binary schema of hate and no-hate followed. In the second stage, multiple rudimentary cleaning steps such as e.g. filtering out links and special characters were performed. For the third stage, nltk[3] frameworks functionality of part-of-speech tagging was utilized to detect entities. Here, only nouns (singular, plural and proper) were selected for further deliberation. In the fourth step, the detected terms were filtered for terms that were not included in the knowledge base of existing English hate speech terms. In the fifth stage, the remaining terms were weighed by the self-proposed class hate score. The last step of the experiment was a human-based evaluation differentiating between identity terms and slurs within the top 20 rankings, ordered from the lowest to the highest hate score class.

Slur	Class Hate Score	Target
'femboy'	0.00567	Men
'emmy'	0.0176	Women
'pervert'	0.0211	Sexual Orient.
'daft'	0.0211	Intelligence
'slappers'	0.0263	Women
'moron'	0.03512	Intelligence
'cuck'	0.03831	Men
'autists'	0.0421	Intelligence
'chuck'	0.0527	Men
'periods'	0.0527	Women

Table 1: Top 10 Detected Emerging Youth Language Hate Terms, their CHS, and Target

4.3 Results

In table 1, it is visible that the emerging entity detection has detected a wide range of terms, all of which are established as hate terms in certain contexts. The experiment detected two types of emerging entities following the definition Färber et al. (2016). Firstly, words such as: 'autists', 'periods','emmy' are known words, and are connected to a prior unknown hateful context. The second type are, words with unknown surface forms (according to the established terms list) which are now connected to known hateful contexts.

To validate the findings, this process was also applied to two existing datasets: firstly, Davidson et al. (2017) and secondly, Vidgen et al. (2021). It was observable that both datasets did not produce 10 emerging slurs within the first 50 detected entities, indicating that the used slurs are covered by the established hate knowledge base and underlining the different language present in the data set from Fillies et al. (2023).

5 Bias Test Set

5.1 Methodology

The created test sets are based on two existing research approaches. Firstly, Dixon et al. (2018) tests for unintended bias by creating their test set out of template sentences with an equal proportion of hateful and non-hateful statements. Secondly, Röttger et al. (2021) and Röttger et al. (2022) build functional test sets for identity terms and slurs and include a wide range of other linguistic features, such as spelling variations and negations. Within this research it has been shown that it is difficult to test for term bias, if the surrounding template structure included hatefully charged terms as part of the template, not regarding the word that is tested. For example, the sentence "I hate [INSERT]" is classified as hate speech, by some of tested algorithms, without any inserted slurs or identity term. Therefore, this research builds a test set based on neutral statements that only receive their hateful character through the inserted slur.

As basis for that, the test cases F7 and F18 from Röttger et al. (2021) are used and validated for their usability within this context. F7 are statements that express hate by using a slur and F18 are neutral statements intended for group identifiers but also suitable for slurs. Statements regarding the self, such as in "I am a [slur]", or regarding belonging to a group, are not included. Overall, 29 template sentences were taken from Röttger et al. (2021) and doubled with 31 self-designed template sentences, bringing the templates to 60 statements. Also included from Röttger et al. (2021) are the functionality tests, F25-F29, which are tests are regarding spelling variations.

Due to the hateful character of the used slurs the statements did not need to be annotated. Overall, based on the emerging entities detected, a test set containing 3600 hateful statements was created for evaluation. This test set will be referred to as the Emerging Test Set (ETS). To counter and identify possible overfitting based on sentence structure or surrounding words, a counter test see was created with the same 3600 test sentences but no hate character (see Appendix B). This test set is referred to as the Overfitting Test set (OTS). The Reference Test Set (RTS) (see Appendix A) also contains 3600 hateful statements.

5.2 Evaluation Metrics

The evaluation of the ETS is done using accuracy, following Röttger et al. (2021). This is reasonable considering that only hateful examples are given and therefore more advanced evaluation metrics are not applicable, due to them relying on distributions of False Negatives (FN), True Negatives (TN) or False Positive (FP). The same applies to the OTS and the RTS.

To evaluate their study, Dixon et al. (2018) based their metrics for the concept of fairness proposed by Hardt et al. (2016) which says that "a model is fair if false positive (FP) and false negative (FN) are equal across statements containing the terms of

[3]https://www.nltk.org/api/nltk.tag.html

interest.”.

To quantify the youth language bias, and following Dixon et al. (2018), the false negative rate (FNR) for the test sets are calculated. A model that is unbiased will have similar values across all terms, approaching the equality of odds ideal, where FNR(ETS) === FNR(RTS) for the data sets and the Reference Test Set. This work introduces the youth language score (YLS) by subtracting the FNR(ETS) from FNR(RTS). This simple measurement can be used to measure a difference in prediction quality and therefore the fulfillment of the fairness condition. The closer the YLS to zero the smaller the assumed bias is.

$$YLS = FNR(ETS) - FNR(RTS) \quad (2)$$

6 Evaluation of Existing Classifiers

6.1 Classifiers

The research evaluated three different available research classifiers and one commercial model. All models are based on the BERT model Architecture introduced by Devlin et al. (2019), and have state-of-the art performance.

The first model is the R4 Target model published by Vidgen et al. (2021)[4] and is based on a RoBERTa model architecture introduced by Liu et al. (2019). They initially trained their model on 11 different data sets created between 2016-2020 and generated new cases to improve the classifier. The classifier was tested for bias using the HATECHECK framework by Röttger et al. (2021).

The second model HateExplain was published by Mathew 2020[5] and the data was extracted from Gab and Twitter from January 2019 to June 2020 based on keywords.

The third research model IMSyPP[6] is trained on YouTube comments collected between January 2020 to May 2020[7] (Ljubešić et al., 2021).

The commercial model is Google Jigsaw’s Perspective[8]. It is trained on data from Wikipedia and The New York Times[9]. The research used the provided feature “IDENTITY_ATTACK”, which identified “Negative or hateful comments targeting someone because of their identity.”

[4]https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target

[5]https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain-rationale-two

[6]https://huggingface.co/IMSyPP/hate_speech_en

[7]https://www.clarin.si/repository/xmlui/handle/11356/1454

[8]https://perspectiveapi.com/how-it-works/

[9]https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US

Classifier	ETS	OTS	RTS
HateExplain	0.141	0.978	0.297
R4 Target	0.577	0.519	0.862
IMSyPP	0.418	0.759	0.487
Jigsaw	0.007	0.998	0.277

Table 2: Accuracy of Models on Test Data Sets

Classifier	ETS	OTS	RTS
HateExplain	0.35	1.0	0.681
R4 Target	0.275	0.928	0.972
IMSyPP	0.597	0.761	0.761
Jigsaw	0.003	0.997	0.431

Table 3: Comparison of Performance on Test Cases (Slurs), without Spelling Variations

6.2 Results

After using the provided model architectures and pretrained models themselves to classifying all statements in the three created data sets, it can be observed that the accuracy of all algorithms decreases when applied to detected emerging youth language terms, as seen in table 2. The best performance on the Emerging Test Set was archived by the R4 Target model. In regards to the overfitting test the models Jigsaw and HateExplain had high scores. For the Reference Test Set, the R4 Target also produced the highest accuracy. To further break down the results and understand the archived accuracy, table 3 displays the accuracy of each model regarding the slurs, disregarding spelling variations. IMSyPP performs the best in identifying singular slurs and plural in the ETS and RTS while HateExplain performs the best on the OTS. Table 4 displays the accuracy regarding the test cases with spelling variation. Here R4 Target is the best performing model for the ETS and RTS and worst for the OTS.

The tables in Appendix C display the perfor-

Classifier	ETS	OTS	RTS
HateExplain	0,099	0,969	0,223
R4 Target	0,634	0,440	0,848
IMSyPP	0,385	0,748	0,428
Jigsaw	0,008	0,999	0,249

Table 4: Comparison of Performance on Test Cases Regarding Spelling Variations

Classifier	FNR ETS	FNR RTS	YLS
HateExplain	0,859	0,703	0,156
R4 Target	0,423	0,138	0,286
IMSyPP	0,582	0,513	0,069
Jigsaw	0,993	0,723	0,271

Table 5: False Negative Rates and YLS for the Models

mance of the models regarding the individual terms. The false negative rate for each classifier has been calculated, see 5. The smallest YLS and FNR can be found within the IMSyPP classifier.

6.3 Discussion of Results

The table 2 indicates that the R4 Target model performs the best on both ETS and RTS. But as stated by Vidgen et al. (2021), to train the R4 Target model, the HATECHECK framework by Röttger et al. (2021) was used to evaluate the performance. This has direct influence on the performance of the model in this framework, which is why the overfitting test set was introduced. It can clearly be seen that while looking at table 2 R4 Target is getting the highest scores in the ETS and RTS, though it scores the lowest on the OTS. This is an indication that the BERT based model is learning to identify the test set structure instead of the underlying slurs. This is underlined when separating the statements into a group containing just statements with slurs (in singular and plural), see 3, from the statements containing the different spelling errors, see 4. Here it becomes visible that IMSyPP is the most accurate model. Similarly, if the results are evaluated on the level of individual slurs (see Appendix C), it is evident that R4 Target model is familiar with all of the slurs out of the RTS but is missing a wide variety of youth language terms. It is worthy to note that all models, besides R4 Target, had fundamental problems in adapting to the proposed spelling errors, while HateExplain shows the biggest difference in performance. The lowest performance is seen in the Jigsaw model, which can be rooted in the hate speech definition or the set threshold. It can be seen that all models perform better on the RTS indicating that the detection of youth language in emerging hate speech is of importance. Overall, based on the youth language score, and taking the overfitting evaluation into consideration, it can be said that IMSyPP is the best suited model for detecting hate specific slurs in youth language. It shows that a youth language bias exists in different well-established classifiers. The finding that IMSyPP is the best performing model can be explained by the collection time and source of the training data, which indicates that time and source of the training data is more important than the size of the machine learning models regarding generalizability and youth language hate speech detection.

7 Youth Language Classifier

7.1 Model Setup

As section 6.3 found, model IMSyPP has the best performance and the smallest youth language bias. Fine-tuning is a proven and established method for bias mitigation (Park et al., 2018; Zhang et al., 2018; Beutel et al., 2017) and relies on providing the model with new data. This research therefore takes the dataset introduced in chapter 3.5 by Fillies et al. (2023) to include the youth language into the model, hence mitigating the bias. The model IMSyPP was provided for research over the platform Hugging Face. In the first step, the 9 labels of the data set from (Fillies et al., 2023) are matched to the four classes used in the IMSyPP model. The matching can be found in Appendix D.

After the matching, all hateful classes were selected and combined with randomly chosen but equally distributed non-hateful statements, which created a new subset containing 11342 messages wherein 50% were hateful and 50% were not hateful. The ration is similar to the ratio of the corpus used for training the IMSyPP model. In The data was then shuffled and tokenized using the tokenizer provided by IMSyPP. A 90% train and 10% test split was chosen and, utilizing PyTorch, the new data set used to fine-tune the pre-trained IMSyPP model. The hyperparameter were kept at the default, using a Learning rate of 5e-5, the AdamW optimizer, 3 epochs, 500 learning steps and batch size of 8.

7.2 Results

The accuracy of the fine-tuned model increased on the ETS (from Acc: 0.418 to Acc:0.613) and on the RTS (from Acc: 0.487 to Acc:0.668). A decrease in performance was visible in the OTS (from Acc: 0.759 to Acc:0.664). On a word level, as seen in Appendix E, the new model now identifies 7 out of the 10 emerging slurs, instead of 6 out of the 10 before and even increased its prediction quality on the RTS. Even though a difference in accuracy is still visible, the new youth language score decreased

from 0.069 to 0.055. It is visible, see Appendix F, that the new model increased its capability to detect both plural and singular slurs and is still challenged by the spelling variations.

7.3 Discussion of Results

The new model did not eliminate age bias but is archiving an increased performance by recognizing more emerging slang terms and having a lower youth language score. It is expected to struggle with the identification of different spellings, primarily because no new data related to this aspect has been introduced in the fine-tuning. The decrease of performance in the OTS is a validation for its existence and could be an indicator of a slight tendency of over detecting hateful content. Overall, it has shown that fine-tuning the model decreases the age bias. The provided model is the first model focused on detecting hate speech within online youth language.

8 Conclusion and Further Research

This research introduces the topic of age bias to algorithmic hate speech classification. It provides a novel framework for detection and evaluation of the phenomena. It shows that the age group of the authors of an online text is influential on the performance of hate speech classifiers. It shows that time and source of the data is more important than size of the machine learning models regarding generalizability. A multistep architecture is proposed based on multiple data inputs, an emerging entity detection, human in the loop evaluation and a newly introduced class hate score indicating the importance a hate-term has in a binary hate speech data set. Three test sets were created and used for evaluation. The research further proposes a youth language score to measure the unintended age bias towards hateful slurs contained in classification algorithms. A separate test set is introduced to identify overfitting of functional tested models. The research evaluates three scientific and one commercial model, fine-tuning the best performing model to mitigate the existing bias. It is shown that age bias can be mitigated for the given model using a fine-tuning approach. The provided model itself needs to be further optimized regarding its hyperparameter. It is of interest to extend the developed framework to more than just the top ten emerging hateful terms, test different NER approaches, as well as how fine-tuning the other evaluated models changes their performance in the framework. Further development would be to evaluate if the model is applicable to different languages, different terms of interest and different age groups. This research opens the debate to understand the influence of youth bias in hate speech classification and builds a framework to quantify and analyze the bias. It provides the first hate speech detection model specifically trained for the use of detecting hate speech within online youth language.

9 Limitations

Multiple points can be raised in connection to the developed framework. It can be argued that the data set (Fillies et al., 2023) is not able to represent youth language as a whole. Therefore, the framework also only detects the age biased represented within this specific subgroup. The raised concerns are in line with the first definition of youth language, pointing out that the results of this research are not generalizable to other youth group settings. But even under this definition of youth language, the proposed framework still holds value as it is applicable to identify this bias within these developing subgroups. Following the second definition, which views youth language as a generational construct, the validity of the approach and results hold.

Secondly, the framework is used for the top ten slurs and reference slurs missing a wide variety of other emerging hate terms. As the framework was able to identify a bias in the most common emerging hate terms, it is a reasonable assumption that it is also applicable for even less established terms. This is a point for further research.

Thirdly, the framework is based on the terms acting as expressions of hate themselves. This is not applicable for all hate speech definitions. Therefore, systems like Jigsaw are misrepresented in the framework. If the hate speech definition of a model does not consider slurs as hate speech, the researcher also has the choice to choose the emerging entity terms that would be considered hate speech under their definition. It is evident that most systems consider certain slurs as hate speech, while missing certain other terms, this needs to be an active decision and not a passive phenomenon.

Acknowledgements

This research was supported by the Citizens, Equality, Rights and Values (CERV) Programme under Grand Agreement No. 101049342.

References

Satoshi Akasaki, Naoki Yoshinaga, and Masashi Toyoda. 2019. Early discovery of emerging entities in microblogs.

Nils Bahlo, Tabea Becker, Zeynep Kalkavan-Aydın, Netaya Lotze, Konstanze Marx, Christian Schwarz, and Yazgül Şimşek. 2019. *Jugendsprache*, 1 edition. J.B. Metzler.

Jennifer Bayzick, April Kontostathis, and Lynne Edwards. 2011. Detecting the presence of cyberbullying using computer software.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *ArXiv*, abs/1707.00075.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english.

Yi Cai, Arthur Zimek, Gerhard Wunder, and Eirini Ntoutsi. 2022. Power of explanations: Towards automatic debiasing in hate speech detection.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Michael Färber, Achim Rettinger, and Boulos Asmar. 2016. On emerging entity detection. In *20th International Conference on Knowledge Engineering and Knowledge Management - Volume 10024*, EKAW 2016, page 223–238, Berlin, Heidelberg. Springer-Verlag.

Jan Fillies, Silvio Peikert, and Adrian Paschke. 2023. Hateful messages: A conversational data set of hate speech produced by adolescents on discord.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences (Switzerland)*, 10.

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).

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning.

Nicolas Heist and Heiko Paulheim. 2022. Transformer-based subject entity detection in wikipedia listings.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 385–396, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Lennart Justen, Kilian Müller, Marco Niemann, and Jörg Becker. 2022. No time like the present: Effects of language change on automated comment moderation. In *2022 IEEE 24th Conference on Business Informatics (CBI)*, volume 01, pages 40–49.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nikola Ljubešić, Igor Mozetič, Matteo Cinelli, and Petra Kralj Novak. 2021. English YouTube hate speech corpus. Slovenian language resource repository CLARIN.SI.

Niall McCarthy. 2020. Facebook removes record number of hate speech posts [infographic].

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy. Association for Computational Linguistics.

Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. 2013. Fine-grained semantic typing of emerging entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1497, Sofia, Bulgaria. Association for Computational Linguistics.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Bernhard Rohleder. 2022. Kinder-& jugendstudie 2022.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the germeval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings*, page 427–439, Berlin, Heidelberg. Springer-Verlag.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

A Established Hate Words Röttgen 2021

For the Reference Test Set (RTS) containing the popular hate terms, the first ten slurs from Röttger et al. (2021) are used: "bitch, slut, whore, tranny, shemale, faggot, fag, queer, nigger, coon". The ten reference terms are included in the hate knowledge base.

B Overfitting Test Set

In the Overfitting Test Set, all statements are non-hateful. The no-hate terms are: “guy, friend, partner, buddy, girl, boy, man, women, classmate, sweetheart”.

C Prediction Accuracy for each Word and each model

Word YL	Perc.	Word Ref.	Perc.
femboy	0.778	whore	0.778
slapper	0.458	faggot	1.0
chuck	0.069	shemale	0.597
period	0.0	fag	0.819
autist	0.3066	nigger	1.0
emmy	0.0	coon	0.944
daft	0.0	slut	0.0
moron	0.792	queer	0.0
pervert	0.306	bitch	0.611
cuck	0.847	tranny	0.917

Table 6: Model: HateExplain

Word YL	Perc.	Word Ref.	Perc.
femboy	0.694	whore	1.0
slapper	0.514	faggot	0.931
chuck	0.639	shemale	0.472
period	0.153	fag	0.694
autist	0.306	nigger	0.667
emmy	0.736	coon	0.528
daft	0.236	slut	0.944
moron	1.0	queer	0.533
pervert	0.889	bitch	1.0
cuck	0.722	tranny	0.417

Table 7: Model: R4 Target

Word YL	Perc.	Word Ref.	Perc.
femboy	0.694	whore	1.0
slapper	0.514	faggot	0.931
chuck	0.639	shemale	0.472
period	0.153	fag	0.694
autist	0.306	nigger	0.667
emmy	0.736	coon	0.528
daft	0.236	slut	0.944
moron	1.0	queer	0.533
pervert	0.889	bitch	1.0
cuck	0.722	tranny	0.417

Table 8: Model: IMSyPP

Word YL	Perc.	Word Ref.	Perc.
femboy	0.0	whore	0.0
slapper	0.0	faggot	1.0
chuck	0.0	shemale	0.389
period	0.0	fag	0.958
autist	0.028	nigger	1.0
emmy	0.0	coon	0.0
daft	0.0	slut	0.0
moron	0.0	queer	0.317
pervert	0.0	bitch	0.0
cuck	0.0	tranny	0.583

Table 9: Model: Jigsaw

D Matching of Labels

Fillies et al.	IMSyPP
0 (No Hate), 8 (Skip)	0 acceptable
1 (Negative Stereotyping), 4 (Equation), 5 (Norm. of Exi. Dis.), 6 (Disguise as Irony)	1 inappropriate
2 (Dehumanization), 7 (Harmful Slander)	2 offensive
3 (Violence and Killing)	3 violent

Table 10: Matching of labels

E Word Level Prediction of the fine-tuned IMSyPP

Youth Lang.	Perc.	Reference	Perc.
'femboy'	1	'whore'	1
'slapper'	0.889	'faggot'	0.944
'chuck'	0.944	'shemale'	0.75
'period'	0	'fag'	1
'autist'	0.278	'nigger'	1
'emmy'	1	'coon'	1
'daft'	1	'slut'	0.542
'moron'	0.986	'queer'	0.825
'pervert'	0.986	'bitch'	1
'cuck'	0.139	'tranny'	0.917

Table 11: Performance on Individual Words

F Fine-Tuned Model Perfoamnce on Slur, Slurs Plural and Spelling Test Cases

Classifier	ETS	OTS	RTS
Slur	0,922	0,831	0,733
Slur Plural	0,858	0,8	0,746
Spelling	0,627	0,629	0,595

Table 12: Fine-Tuned Model Performance on each Slur Test case and Spelling Test Cases