On the Interaction of Identity Hate Classification and Data Bias

Donnie Parent, Nina Georgiades, Charvi Mishra,
Khaled Mohammed, Sandra Kübler
Indiana University
{daparent,ngeorgia,chmish,mohammek,skuebler}@iu.edu

Abstract

Hate speech detection is a task where machine learning models tend to be limited by the biases introduced by the dataset. We use two existing datasets of hate speech towards identity groups, the one by Wiegand et al. (2022) and a reannotated subset of the data in AbuseEval (Caselli et al., 2020). Since the data by Wiegand et al. (2022) were collected using one syntactic pattern, there exists a possible syntactic bias in this dataset. We test whether there exists such a bias by using a more syntactically general dataset for testing. Our findings show that classifiers trained on the dataset with the syntactic bias and tested on a less constrained dataset suffer from a loss in performance in the order of 20 points. Further experiments show that this drop can only be partly attributed to a shift in identity groups between datasets.

1 Introduction

Hate speech detection is an area of NLP that has gained importance over the last few years. However, while our knowledge of how to approach the problem computationally increases, we find that many of the existing datasets come with unintended limitations and biases. These datasets are known to have biases towards particular targets of the abusive language or the hateful language used. These biases tend to affect implicit hate, or "instances where a comment/post alludes to stereotypes or other negative attributes and associates them with a particular group of individuals, especially as it relates to attributes related to a group or individual's identity (e.g., ethnic heritage, nationality, gender, sexual orientation, religion, disabilities, body shape)" (Lopez and Kübler, 2025), even more strongly, as there are uncountable ways of creating implicit hate.

Creating datasets for hate speech detection is challenging, since hate speech is a rare phenomenon, and thus difficult to find. Implicit hate, since it is indirect and does not use easily identifiable terms such as slurs, is even more difficult to find and to detect automatically. To better detect implicit hate, Wiegand et al. (2021b) argue in favor of individual models for specific subtypes of implicit hate. They have created datasets and models for hateful comparisons (Wiegand et al., 2021a), hate towards identity groups (Wiegand et al., 2022), and euphemistic abuse (Wiegand et al., 2023).

In our current work, our main focus is implicit hate directed at identity groups. Examples of such hateful language are shown in (1) (copied from (Wiegand et al., 2022)).

(1) a. Jews succumb to cultural degeneracy.
b. Gay people are contaminating our planet.
c. Women fabricate menopausal symptoms.

The dataset by Wiegand et al. (2022) was created by searching for tweets where an identity group subject was followed by a negative polarity verb. Selected tweets were then annotated for hateful content. Since all examples exhibit very similar syntactic structures, our question concerns the generalizability of this type of data. In other words, does a model trained on data limited to one syntactic pattern generalize to data with more diverse syntactic patterns? Or does this syntactic bias result in lower performance?

Note that our intention is not to criticize the method chosen by Wiegand et al. (2022) to create their dataset. We are aware that implicit hate is very difficult to find because of its very nature, and that we need specific search strategies to find such examples. However, every decision that we make with regard to search strategies has an effect on the dataset and classifiers, and we need to understand these effects. Our current work is intended to pro-

Proceedings of Recent Advances in Natural Language Processing,pages 900–906
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-103

vide a closer look at the consequences of one such decision.

The structure of the remainder of the paper is as follows: Section 2 explains our research questions, and section 3 describes related work. In section 4, we explain our methodology, and in section 5, we discuss our results. We conclude in section 6.

Offensive Content Warning: This report contains some examples of hateful content. This is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Please be aware that this content could be offensive and cause you distress.

2 Research Questions

Our main research question concerns the syntactic pattern that was used to create the dataset by Wiegand et al. (2022): Can a classifier trained on a syntactically constrained dataset generalize successfully to other data that do not follow this pattern? If this is the case, that would mean that there is no syntactic bias, and Wiegand et al. have succeeded in their goal of creating a highly reusable dataset. If not, the syntactic pattern constitutes a bias. We also investigate whether the bias we find is due to the choice of negative examples or due to differences in the identity groups covered in the data.

3 Related Work

Work on hate speech towards identity groups is in its early stages: Yoder et al. (2022) investigate how hate speech varies by identity. Sachdeva et al. (2022) create a multi-label classifier to identify identity groups targeted by hate speech. They show that the model trained on the Measuring Hate Speech corpus generalizes well two other datasets. Jin et al. (2025) investigate hate speech detection for identity groups. They show that such detection models assign a higher hate score based on the mention of specific target identities. They also show that such models are more accurate when there is strong intensity in terms of the stereotype. In related work, Zueva et al. (2020) address unintended bias towards protected identity groups in Russian by creating artificial, non-toxic sentences about such identity groups to use as training data.

Bias may be introduced into hate speech detection in different ways. The first bias concerns sampling (Wiegand et al., 2019; Razo and Kübler, 2020); In determining how to sample data to increase the percentage of hate speech, the data can be biased towards specific users or topics. Other types of bias are introduced in the annotation process: Sap et al. (2022) show that annotators are biased towards judging African American English as hate speech, and Lopez Long et al. (2021) show that especially untrained annotators have a tendency to confuse casual profanity and argumentative language with hate speech. Finally, Lopez and Kübler (2025) show that utterances need to be analyzed in their context to determine whether they contain hate speech.

Our current work uses the definition of implicit hate by Lopez and Kübler (2025): "instances where a comment/post alludes to stereotypes or other negative attributes and associates them with a particular group of individuals, especially as it relates to attributes related to a group or individual's identity (e.g., ethnic heritage, nationality, gender, sexual orientation, religion, disabilities, body shape)." This definition has a particular focus on identity and the intersection between implicit hate and identity hate.

4 Methodology

4.1 Datasets

Our work focuses on two datasets, the dataset by Wiegand et al. (2022) and a subset of AbuseEval (Caselli et al., 2020), where all tweets that were deemed implicitly abusive by Caselli et al. were annotated whether they are abusive towards identity groups (Parent et al., 2025). We use the Wiegand dataset for training and testing (in domain), and the subset of AbuseEval, henceforth Caselli-Ident, for out-of-domain testing on more general data.

Wiegand et al. (2022) created an abusive language dataset (using both English and German) sentences by sampling stereotypical Twitter posts. The creation of this dataset follows the assumption by Wiegand et al. (2021b) that all subtypes of hate (e.g., racism, sexism) must be addressed separately rather than all together. This dataset contains implicit hate towards the following groups: gay people (using the terms "gay people" and "lesbians"), Jews, Muslims, and women. Notably, no nicknames or slurs for any of the identity groups were used in the data gathering, only the listed "names" of the identity groups. They searched for posts where the target group is the subject of a negative-polar verb. This search strategy leads to

data that are syntactically similar.

Some examples of hate speech from the Wiegand dataset are as follows:

(2) ALL women blab on social media.

(3) Muslims bungle everything they attempt to the detriment of everybody.

(4) jews are disrupting the world peace.

(5) Gay people are diluting any sanctity.

(6) lesbians dehumanize other LGBTQ people.

Caselli et al. (2020) used the OffensEval dataset OLID (Zampieri et al., 2019) and reannotated it with more detailed information. OLID is an English dataset collected from Twitter, which is annotated for whether the post is offensive, whether the post has a target, and what that target is. Caselli et al. annotated each post as Explicit, Implicit, or Not Abusive. The resulting dataset is called AbuseEval.

The focus of our research is implicit hate towards identity groups. Since we are interested in whether the search strategy employed by Wiegand et al. (2022) to create their identity hate dataset results in an unintended bias, we needed a dataset annotated for hate towards identity groups that is syntactically more diverse. In order to minimize genre effects, the data should also originate from Twitter. Since no such dataset existed, we used AbuseEval and extracted all posts labeled as Implicit. We then annotated this subset for identity hate. We annotated the posts for the type of hate speech (identity hate, non-identity related hate, or not abusive). If the post was identified as identity hate, annotators decided whether the identity group was referenced implicitly or explicitly. Finally, annotators extracted the target of the abusive from the post. For more details about the dataset and the annotation process, refer to (Parent et al., 2025).

The annotated dataset is freely available at https://github.com/donnieparent/X490-hate-speech/tree/main. We have provided two versions, one with all five annotations per example, and one with the a single label, determined by the majority vote.

Since our annotations showed that about half of the tweets labeled as implicit hate by Caselli's annotation were not considered identity abuse by our annotators, the dataset resulting from our annotations was similarly balanced as compared to the Wiegand dataset. Table 1 shows the distribution of classes in the three datasets we use.

Dataset	Ident.	Not ident.	Not Abuse
Wiegand	57.0	0	43.0
Caselli-Ident	49.5	49.3	1.3
Caselli	32.9	32.7	34.4

Table 1: Distribution (in %) of identity hate, not identity hate, and non-hate in the datasets.

4.2 Machine Learning Models

We tested three different classifiers: Support Vector Machines (SVM), Random Forest, and Naive Bayes, all in the scikit-learn (Pedregosa et al., 2011) implementation. We avoided using LLMs in order to avoid unintended effects of data leakage since we cannot guarantee that the dataset we use for testing is not part of the training data for the LLM.

We performed parameter optimization on the Wiegand data, using a five-fold cross validation grid search to determine the optimal parameters for each of the three classifiers. Going forward, we used the best-performing parameters for all experiments.

For the experiments to determine the generalizability of the Wiegand dataset, we consider our baseline to be the model performance on the Wiegand dataset. We perform a 5-fold cross validation on the entire dataset. Next, we train on 80% of the Wiegand dataset, to ensure the same training set size as in the 5-fold CV, and test on the Caselli-Ident dataset. To determine whether the negative examples in Caselli-Ident are misleading (since they were originally considered hate speech), we replace these negative examples with 403 non-abusive posts from the original Caselli dataset. This new dataset then serves as test set. Finally, we attempted training on not just the Wiegand dataset alone, but a combination of Wiegand data and a small part of Caselli-Ident. This new training set was comprised of 80% of Wiegand data and 20% of Caselli-Ident.

4.3 Evaluation

We mainly report macro-averaged F1. Since the datasets are close to balanced, there is little difference between micro-averaged and macro-averaged F1, and precision and recall are equally balanced.

Test data	Wiegand			Caselli-Ident			Caselli		
Classifier	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
SVM	74.78	74.14	74.27	58.88	56.38	53.21	56.15	54.64	51.80
Random Forest	73.70	70.53	70.62	54.59	53.66	51.30	45.39	46.05	44.13
Naive Bayes	75.46	73.32	73.58	58.64	58.63	58.63	59.90	59.87	59.86

Table 2: Classification results (macro-averaged) when training and testing in-domain (Wiegand data), out-of-domain (Caselli-Ident) and out-of-domain (using non-abusive examples from the Caselli data).

5 Results

5.1 Comparing Wiegand and Caselli-Ident Test Data

Our main research question concerns the potential bias in the Wiegand dataset, based on the sampling of stereotypical sentences with similar syntactic patterns. To investigate this, we trained the classifiers on Wiegand data and tested on Wiegand and on Caselli-Ident data. The results are shown in Table 2.

First, all three models performed well on the Wiegand test set, with SVM resulting in the highest F1 score of 74.27, followed by Naive Bayes (73.58), and Random Forest with an F1 score of 70.62. This shows that the classifiers are capable of addressing the problem successfully.

When testing the models on Caselli-Ident, the models' performance noticeably suffers, with F1 deteriorating by 15-20 percent points. Naive Bayes reaches the highest F1 score of 58.63, and Random Forest continues to perform the worst with an F1 score of 51.30. These results show that there are significant differences between the two datasets, even though both are sampled from Twitter, and both focus on hate speech towards identity groups.

One reason why the results are lower on the Caselli-Ident set may be found in the fact that the negative examples in this dataset were originally annotated as hate speech by Caselli et al. and contain a mix of non-hate and hate that is not directed towards identity groups. This means that these examples are not clear-cut, easy examples of non-hate. Therefore, they are likely more difficult to distinguish from hate speech than the non-hate examples in the original Caselli dataset. For this reason, we repeated the experiment after having replaced the negative examples by an equal number of non-hate examples from the original Caselli dataset.

The results are reported in the third column of Table 2. When testing on this dataset, we would expect an improved performance. However, the SVM and Random Forest models show an additional drop in F1. Naive Bayes performs slightly better, its F1 score increasing by slightly more than a percent point to 59.86. This shows that Naive Bayes is more robust towards out-of-domain data. However, overall the examples in this test set are further away from the Wiegand set than the negative examples in Caselli-Ident. Thus while the latter may be more difficult to distinguish from the hate examples, they are more in-domain than the negative examples from the original Caselli set.

5.2 Do the Identity Groups Matter?

In the Wiegand dataset, the identity hate is directed at a specific subset of identity groups, including gay people, Jews, Muslims, and women. In Caselli-Ident, there is no restriction on the identity groups. This leads to the question whether the classifiers learn hate speech directed specifically at the chosen identity groups but are not able to generalize to other identity groups. For this reason we use the target annotations in the Caselli-Ident dataset and separate the hate examples into two groups: one group with target groups covered in the Wiegand dataset, and one group with other targets. Then we evaluate separately on both groups. The results are shown in Table 3.

On the Wiegand groups, the SVM has an F1 score of 58.29, Random Forest's F1 score is 52.35, and Naive Bayes has the highest F1 score of 65.30. This means that the Random Forest model only improves by about one percent point over the results on all categories. In contrast, SVM and Naive Bayes improve by 5 and 7 percent points respectively. This shows that the groups covered by Wiegand et al. (2022) are clearly easier to classify correctly, however the results are still significantly below the results when testing on Wiegand data.

For all other identity groups, the F1 scores range between 50.76 and 57.35, i.e., they are consistently about one percent point lower that on the whole dataset. As such, these results show that while the discrepancy in identity groups contributes to the

	Wiegand groups			Other groups		
Classifier	Prec	Rec	F1	Prec	Rec	F1
SVM	61.57	58.50	58.29	58.35	56.05	52.28
Random Forest	54.69	53.35	52.35	54.18	53.42	50.76
Naive Bayes	65.24	66.30	65.30	57.38	57.39	57.35

Table 3: Evaluation of classification results on Caselli-Ident hate examples split into groups covered by Wiegand et al. (2022) and other groups.

	OOD			add in-domain data		
Classifier	Prec	Rec	F1	Prec	Rec	F1
SVM	58.88	56.38	53.21	65.60	65.02	64.54
Random Forest	54.59	53.66	51.30	75.50	72.41	71.23
Naive Bayes	58.64	58.63	58.63	71.39	71.40	71.36

Table 4: Comparing out of domain data (repeated from Table 2) and adding in-domain training data.

difficulty in classifying the Caselli-Ident dataset, they are not the main source of difficulty.

5.3 Adding In-Domain Data

For our final research question, we approached the domain differences between the two datasets by adding a small amount of Caselli-Ident data to the Wiegand training set. Having access to some data from the test domain during training should help the classifiers to cover this domain better.

Thus, we trained the models on a set made up of 80% Wiegand and 20% Caselli-Ident. The test set was the subset of Caselli-Ident that was not used in the training. The results of this experiment are shown in Table 4.

By adding Caselli-Ident data to the training set, the models performed noticeably better: Random Forest and Naive Bayes reached F1 scores of 71.23 and 71.36 respectively, which is an improvement of about 20 percent points for Random Forest and 13 percent points for Naive Bayes. The SVM reached the lowest F1 score of 64.54, which is still 11 percent points higher than the out-of-domain result. Note that Random Forest marginally surpasses the results when testing on Wiegand data, and Naive Bayes reaches similar results. Only the SVM cannot profit as much from the added target domain data.

These results show that it is possible to reach high results on the Caselli-Ident dataset, thus confirming our hypothesis that the search strategy in the Wiegand data set created a syntactic bias in the dataset.

6 Conclusion and Future Work

We have investigated the question whether a data collection of hate speech toward identity groups based on specific syntactic patters affects a classifier's ability to identify such hate in other syntactic patterns. Our results show that this is the case: Results decrease by 15-20 percent points when tested out of domain on the Caselli-Ident dataset, which corresponds to more diverse syntactic patterns. We have also shown that this is not due to our choice of negative examples, nor is it due to the wide range of identity groups. In contrast, adding in-domain data improved results considerably—showing that the syntactically diverse dataset is not inherently more difficult.

Looking ahead, we plan to investigate the divergence in identity groups more closely, with the goal of developing more robust methods that can better handle unseen identity groups. We also plan on investigating fairness of such machine learning methods to determine if similar utterances are classified consistently across identity groups.

7 Limitations

The quality of hate speech detection models always depends on the quality of the annotations. The data we use has been collected from three different datasets, and has thus been annotated by three different groups of annotators. There is a strong probability that the definition of the phenomenon in question, annotation guidelines, and boundaries between categories vary between those datasets, which introduces a discrepancy between the annotations that can affect classifier accuracy.

Acknowledgments

We are grateful to Annika Shankwitz, who helped with many parts of the project.

References

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France.

Yiping Jin, Leo Wanner, and Aneesh Moideen Koya. 2025. What the #?*!: Disentangling hate across target identities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 199–221, Albuquerque, New Mexico. Association for Computational Linguistics.

Holly Lopez and Sandra Kübler. 2025. Context in abusive language detection: On the interdependence of context and annotation of user comments. *Discourse, Context & Media*, 63:100848.

Holly Lopez Long, Alexandra O'Neill, and Sandra Kübler. 2021. On the interaction between annotation quality and classifier performance in abusive language detection. In *Proceedings of the Conference on Recent Advances in NLP (RANLP)*, Online.

Donnie Parent, Nina Georgiades, Charvi Mishra, Khaled Mohammed, and Sandra Kübler. 2025. Annotating hate speech towards identity groups. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Dante Razo and Sandra Kübler. 2020. Investigating sampling bias in abusive language detection. In *Proceedings of the 4th Workshop on Online Abuse and Harms (WOAH)*, pages 70–78, Online.

Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, and Chris Kennedy. 2022. Targeted identity group prediction in hate speech corpora. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 231–244, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, WA.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612, Seattle, United States.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: The problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How hate speech varies by target identity: A computational analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, MN.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian

hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69, Online. Association for Computational Linguistics.