JOURNAL OF
EMERGING INVESTIGATORS

Article

Battling cultural bias within hate speech detection: An experimental correlation analysis

Shlok Bhattacharya[1], Jean Kanzinger[1]
[1] Chagrin Falls High School, Cleveland, Ohio

SUMMARY

Hate speech detection systems have become essential in the advancing digital world. They limit the dissemination of hateful and offensive language online. However, the machine learning algorithms that provide the basis for these systems struggle to identify hate speech versus clean speech within a cultural context, allowing the growth of cultural bias. Though previous methods had aimed to mitigate the cultural bias of a machine learning model, we attempted to find a new understanding with regard to cultural bias. This study sought to determine a correlation between increasing the amount of cultural speech used to train the machine learning model and the model's cultural bias when classifying hate speech and clean speech. Additionally, we hypothesized that increasing the cultural weight of a training dataset would mitigate the cultural bias. To test this hypothesis, we created a unique method named Categorial Weighted Training (CaWT), derived from multiple other methods of previous researchers, to identify a correlation. CaWT involved the creation of multiple culturally-weighted training datasets and training a machine learning algorithm against them. From this, the results illustrated minimal correlation between the cultural weight of a training dataset and the model's performance on cultural speech. However, a significant negative correlation exists between the cultural weight and the model's performance on non-cultural speech. This implies that increasing the cultural weight does not affect the model's cultural bias but decreases the model's performance on non-cultural speech, suggesting that a lower cultural weight is ideal within the limitations of our research.

INTRODUCTION

The past few decades have introduced new technologies that allow for instantaneous global communication among people of various cultures. The expansion of social media networks and the Internet enables unprecedented access and dissemination of information (1). Considering the importance of global communication, it has proven essential to maintain a user-friendly environment to refine this communication network.

Unfortunately, the presence of hateful language grows with the rise of user-generated content (2). Additionally, the option to remain anonymous promotes the dissemination of hateful content in the Internet (1). Furthermore, this hateful content may consequently cause mental illnesses such as depression, even pushing people toward suicide, highlighting the need for an improved system to mitigate hate speech (HS) – language that some may find offensive – and promote clean speech – language that is not considered offensive – on the Internet (3). As a result, social media platforms attempt to combat this rising problem with HS detection software. These platforms utilize artificial intelligence with the aim of identifying social media posts that contain HS. However, the classification process does not adapt well to the speech of different cultures and their specific distinctions between HS and clean speech. For example, certain slang may be considered clean when said within a culture, however, it may be considered HS when those outside that culture use it. This distinction between HS and clean speech proves to be essential when mitigating the apparent cultural bias in current HS detection.

HS detection software is created using supervised machine learning (ML) models which involves the process of training a model on previous data to predict future classifications (4). Like humans, computers require information in the form of a training dataset to learn patterns from, akin to a study guide. However, they are still vulnerable to bias. If a system is presented with a biased training dataset, the model may falsely flag content by certain groups as hateful or may not flag hateful content by some groups as HS (5,6). These biases originate from various sources of the ML process but are all related to the training datasets. Any bias present is taught to the ML algorithm, and its performance reflects these biases. For instance, due to the current method of training HS detection algorithms, phrases in African American English (AAE) were twice as likely to be labeled as more hateful than phrases in other dialects despite it being considered clean by AAE speakers (7). This illustrates that the idea of HS has strong cultural implications, so for individuals, depending on one's cultural views, a phrase may be interpreted as hateful, however current models struggle to consider this cultural context in their classifications (8).

There have been attempts to mitigate this cultural bias within ML models. For example, Ji Ho Park and her colleagues at Hong Kong University of Science and Technology proposed a method called bias fine-tuning. This method first trains a model on a large dataset without bias and then fine-tunes it on a smaller, more biased dataset (5). The intuition is that the model is generalized to the data from the first training dataset and then is focused on a specific aspect of the data from the second training dataset (5). This model was used to mitigate gender biases in HS detection by 90-98% (5). Another procedure was biased topic sampling used by Dante Razo and Sandra Kübler from Indiana University, Bloomington. They trained an ML model on a training dataset which only

DOI: https://doi.org/10.59720/23-156

contained topics known to incite abuse such as politics and religion (9).

Our research addresses the extent of the impact that a culturally-weighted training dataset (CWTD) has on the ability of a support vector machine (SVM) model to mitigate cultural bias in the context of HS detection. We chose an SVM model because SVMs are capable of classifying data into given categories such as hate/clean and cultural/non-cultural speech. Specifically, our unique proposed method, called Categorical Weighted Training (CaWT), was inspired by both bias fine-tuning and biased topic sampling. We hypothesized that increasing the cultural weight (CW) will have a negative relationship with cultural bias, which would be desirable since it indicates a decrease in cultural bias. However, we also suspected that increasing CW would decrease the SVM performance on non-cultural speech. The final results both affirmed and refuted parts of this hypothesis. We also found that CW has a negative relationship with SVM performance on non-cultural speech. However, we found that increasing the CW of a training dataset had no noticeable effect on cultural bias in HS detection. This implies that, overall, within the range of 70% to 95% CW, the SVM performance is optimized at 70% CW.

RESULTS

Classification Measurements

We used the F1-score to measure cultural bias where a greater F1-score on cultural speech indicated a lower presence of cultural bias. Precision and recall were elements of the F1-score. Within a given class, precision quantified the proportion of false positives, where a low precision score indicated more false positives than true positives. On the other hand, recall quantified the proportion of false negatives, where a low recall score indicated more false negatives than true positives.

Classification Model Performance

The model's overall performance for Hate Cultural Speech (HCS) increased slightly as CW increased since the slope of the trendline was slightly positive at 0.0962 and had a strong coefficient of determination, or R^2, of 0.776 (**Figure 1**). Furthermore, the model was less likely to classify HCS as one of the other three classifications at greater values of CW due to the positive relationship of recall with CW (**Figure 1**). However, the SVM model made more false HCS classifications with an increase in CW, indicated by the negative slope of precision (**Figure 1**). Since the R^2 values for precision and recall were greater than 0.9, these scores for HCS maintained a strong correlation with CW (**Figure 1**).

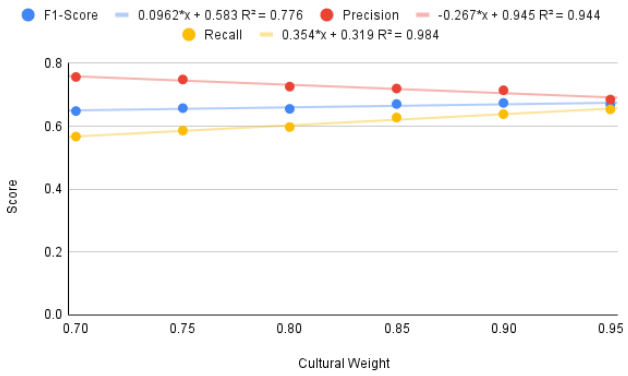

Figure 1: Hate Cultural Speech (HCS). Scores vs. Cultural Weight (CW). Scatterplot with best-fit line illustrating the correlation between the scores – F1-score, precision, recall – relative to HCS and the CW of the training dataset. Each data point is the average score of the sub-experiments within the specific parent experiments. R^2 is the coefficient of determination and represents the correlation of the scores to the CW.

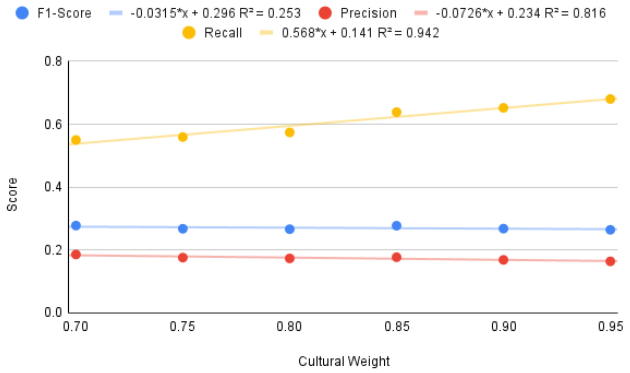

Figure 2: Clean Cultural Speech (CCS). Scores vs. Cultural Weight (CW). Scatterplot with best-fit line illustrating the correlation between the scores – F1-score, precision, recall – relative to CCS and the CW of the training dataset. Each data point is the average score of the sub-experiments within the specific parent experiments. R^2 is the coefficient of determination and represents the correlation of the scores to the CW.

For Clean Cultural Speech (CCS), the model's performance on this classification remained largely unchanged due to the F1-score's near-zero slope and low R^2 value of 0.253 (**Figure 2**). Regardless, the strong correlation, with R^2 of 0.816, but slight negative slope of the precision indicated that the SVM model made more false CCS classifications when the CW was increased (**Figure 2**). Since the precision remained below 0.2, under 20% of the model's CCS predictions were correct throughout the experiments (**Figure 2**). Additionally, the SVM model was less likely to classify CCS as one of the other three classifications. Specifically, the slope and R^2 of the recall were 0.568 and 0.942, respectively, concluding a strong, positive relationship between CW and the recall for CCS (**Figure 2**).

In terms of Hate Non-Cultural Speech (HNCS), the model's performance slightly decreased due to its negative slope (**Figure 3**). However, the R^2 value of the F1-score was 0.339, which indicated a weaker correlation to CW (**Figure 3**). The yellow trendline, representing the recall score, indicated that as CW increased, the model was more likely to classify HNCS as another classification with greater CW values for the training dataset (**Figure 3**). Since the R^2 was 0.894 for recall, this was a strong correlation (**Figure 3**). On the other hand, the precision score increased slightly with CW while maintaining a strong correlation, demonstrating that of all the classified HNCS, fewer were falsely classified as CW increased (**Figure 3**).

The model performance on Clean Non-Cultural Speech (CNCS) decreased since the F1-score experienced a negative relationship with CW (**Figure 4**). Similarly, the recall decreased as well, suggesting that the model falsely

DOI: https://doi.org/10.59720/23-156

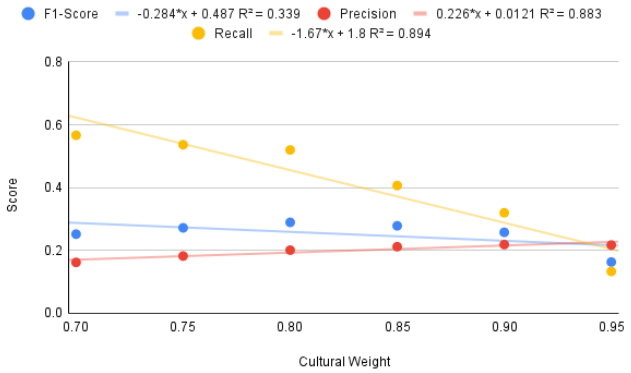

Figure 3: Hate Non-Cultural Speech (HNCS). Scores vs. Cultural Weight (CW). Scatterplot with best-fit line illustrating the correlation between the scores – F1-score, precision, recall – relative to HNCS and the CW of the training dataset. Each data point is the average score of the sub-experiments within the specific parent experiments. R^2 is the coefficient of determination and represents the correlation of the scores to the CW.

classified more CNCS phrases in other categories as the CW increased (**Figure 4**). Conversely, precision increased so the model's false CNCS classifications decreased as CW increased (**Figure 4**). All these associations were extremely strong with R^2 values all above 0.9 (**Figure 4**).

DISCUSSION

Upon the analysis of the presented data, it is evident that a correlation between CW and SVM performance exists. The F1-scores of both the cultural speech categories remained relatively unchanged as CW increased due to their near-zero slopes. However, the F1-scores of the non-cultural speech categories illustrated a more dramatic decrease as CW increased. In other words, increasing CW does not impact the model's overall performance when classifying cultural speech, but decreases its overall performance when classifying non-cultural speech. Thus, increasing CW does not necessarily reduce cultural bias due to this relationship

Though the F1-scores illustrate a general picture regarding how well the SVM model categorizes certain speech, the precision and recall scores create a more in-depth image that specifies how the model thinks when it categorizes phrases into the four categories.

Together, the scores indicate that as the CW increases, the model will decreasingly misclassify the cultural speech as non-cultural speech due to their being more cultural speech to train on. Conversely, since the precision for the cultural speech categories decreases with the increase in CW, the model will increasingly misclassify the non-cultural phrases as cultural speech due to their being less non-cultural speech to train on.

Furthermore, the model became more likely to classify non-cultural phrases as cultural speech due to non-cultural speech being the minority class. This means that the model was not well trained on non-cultural speech and was therefore misclassified more often as cultural speech. On the other hand, the model was less likely to classify cultural phrases as non-cultural speech when trained with a dataset of a high CW due to cultural speech being the majority class. Specifically, the model was more comfortable classifying cultural speech because it was more present than non-cultural speech in the training dataset.

Overall, the model's resulting recall and precision scores grew in opposite directions for both cultural speech and non-cultural speech, which provides specific details regarding the false negatives and false positives of each performance. The diverging precision-recall curves occur due to the precision-recall tradeoff (10). In other words, this concept illustrates that as one of the scores grows, the other score is offset as a result of the imbalanced dataset our procedure utilizes (10). In this case, since cultural speech is the majority class and non-cultural speech is the minority class in the training datasets, the precision-recall curves must diverge due to this imbalance. Additionally, since cultural speech is the majority class, the model is already adequately trained on it so adding more cultural speech does not necessarily improve the model performance on this category. For this reason, the F1 curve remains constant. This refutes our initial hypothesis that increasing the CW of a training dataset will correlate to an increase in SVM performance, thus a decrease in cultural bias. Despite this, we correctly predicted that the CW of a training dataset will have a negative relationship to the SVM performance on non-cultural speech. This occurs because non-cultural speech is the minority class so as its percentage decreases when CW increases, the model has less non-cultural phrases to train on thus the performance on non-cultural speech will decrease.

This new understanding regarding the relationship between a training dataset's CW and cultural bias provides social media companies with the necessary information for improved HS detection systems. Specifically, this study, within its limitations, finds it best that companies including X, Meta Platforms, and Alphabet should train SVM models on datasets that have about 70% CW when using CaWT. Since CW has a minimal relationship with the model performance on cultural speech, a lower CW does not correlate to cultural bias. However, a lower CW does correlate with a greater model performance on non-cultural speech. Thus, to improve HS detection software using CaWT, social media platforms should utilize a training dataset where around 70% of it

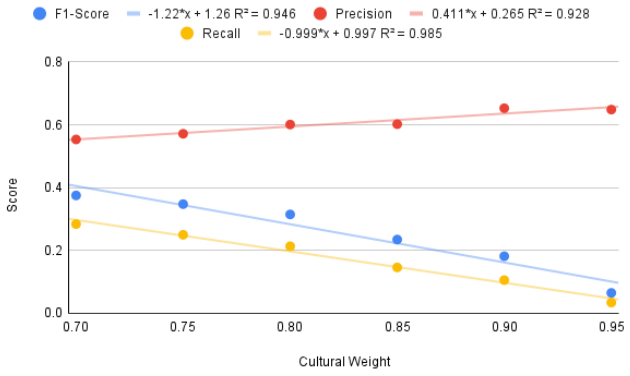

Figure 4: Clean Non-Cultural Speech (CNCS). Scores vs. Cultural Weight (CW). Scatterplot with best-fit line illustrating the correlation between the scores – F1-score, precision, recall – relative to CNCS and the CW of the training dataset. Each data point is the average score of the sub-experiments within the specific parent experiments. R^2 is the coefficient of determination and represents the correlation of the scores to the CW.

DOI: https://doi.org/10.59720/23-156

contains cultural speech.

Though our research concludes an intriguing correlation between the CW of a training dataset and the four different speech classifications, there exist limitations that future researchers may address to further this study on CaWT. For example, HateXplain only contained 640 HNCS phrases which limited the size of the sub-experiment training datasets to 3,640 phrases in order to accommodate large CWs. However, compared to other dataset sizes, this size is relatively small. For example, HateXplain contains 15,383 total phrases and is intended to be used as one training dataset. Additionally, the F1 for HCS remains around 0.65, while for CCS, it is approximately 0.27 (Figure 1, Figure 2). Despite equal phrase counts in both classes, the significant difference in their F1-scores indicates data insufficiency in the training dataset. Although 3,640 phrases proved sufficient to determine whether the correlation between CW and model performance existed, future researchers should utilize larger datasets that contain more than 10,000 phrases for more accurate and higher-performance results. Furthermore, researchers could expand the scope of the research rather than limiting themselves to a CW range of 70% to 95%. The results of such a study would be able to provide a more general understanding of the effects of increasing CW on CaWT.

Additionally, with a larger dataset size, researchers could address the relatively large 5% CW increment size used in this research. A larger dataset may allow the increment size to be 3% or less. This will allow researchers to collect more data on the CW percentages that were not used in this research, resulting in a more accurate trendline. Ideally, researchers should find or create a large dataset that allows an increment size of 1%, to accommodate more specific proportions, while ensuring that the number of clean speech phrases is equivalent to the number of HS phrases.

Furthermore, other researchers should consider increasing the number of sub-experiments beyond 10. Increasing the number of sub-experiments allows for a more accurate average to represent the CW percentage of the parent experiments. However, running one sub-experiment for a training dataset of size 3,640 on a testing dataset of size 910 takes around 45 seconds. For this reason, future researchers should create an algorithm that runs these sub-experiments a given number of times to optimize the time efficiency of the researcher, while allowing more sub-experiments to represent the data.

Also, HateXplain does not provide information regarding the contributors to the posts. More specifically, the users and targets of the posts are unknown to maintain anonymity. However, this ethical limitation, though reasonable, does not allow research regarding online interactions among those of different cultures or those of the same culture. Researchers should aim to train the SVM model on phrases with differing cultural contexts. For example, to better HS detection, an SVM model should understand the difference between a certain phrase said within the commenter's culture than when that phrase is directed toward a different culture. If future researchers address this limitation, HS detection may prove more reliable in a social media environment where much diversity exists.

While CaWT targets HS detection, it can be extended to other fields using the same concept of weighing the training dataset. CaWT opens new possibilities for future research. Anywhere SVM models seem to output bias, CaWT can be applied to that scenario with its results investigated.

This research concludes that CW and the performance of an SVM model on cultural speech are weakly correlated, thus having no effect on cultural bias. However, the performance on non-cultural speech is negatively related to CW. Furthermore, this research explores in-depth the reasoning behind the model's F1-score by explaining the precision and recall scores of each classification. These scores provide insights into how the model thinks while classifying in terms of false positives and false negatives. This knowledge may prove useful to create an SVM model specific to HS detection. In short, CaWT gives future researchers and social media companies new knowledge on how CW impacts performance on individual classifications, which may improve the currently applied HS detection software on social media platforms.

MATERIALS AND METHODS

Score Descriptions

The SVM experiments were measured using three scores: F1-score, precision, and recall. The F1-score considers the false positives and false negatives in its calculation; thus, it yields a refined story regarding the model's performance on imbalanced datasets. Furthermore, the F1-score is calculated using the following formula:

$$2\left(\frac{Precision \cdot Recall}{Precision + Recall}\right).$$

From the formula, the precision and recall scores are elements within the F1-score. Precision calculates the accuracy of all the positive predictions of a class:

$$\frac{True\ Positives}{True\ Positives + False\ Positives} = \frac{True\ Positives}{Positive\ Predictions}.$$

Recall calculates the accuracy of the actual positives of a class:

$$\frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{True\ Positives}{Actual\ Positives}.$$

Though the F1-score provides information regarding the overall SVM performance on a classification, precision and recall provide more detail regarding the model's classifications in terms of false positives and false negatives.

Materials

We operated an SVM model which is an ML model capable of classifying data and is widely used in text classification settings. We conducted SVM experiments using supervised learning in Jupyter Notebook, meaning that we presented the model with a training dataset that contained annotations telling it what each data phrase should be categorized as. The model then learns from the training dataset and can predict the classifications of the testing phrases in the testing dataset.

We also utilized HateXplain, an annotated dataset published by professors at the Indian Institute of Technology in Kharagpur and Universität Hamburg. To create HateXplain, the researchers asked annotators to label social media posts as hate, offensive, or normal speech and then label the target community of the post such as African, Indian, Islam, Jewish, or even no target (11). Additionally, to ensure the accuracy

DOI: https://doi.org/10.59720/23-156

of the annotations, three annotators voted on each type-of-speech label where the disputed phrases were omitted from HateXplain (11). We classified all offensive speech classifications as HS due to the scope of this research focusing on the target of the phrases rather than the specific speech classification. Furthermore, we labeled a phrase as cultural speech if the majority of annotators agreed that the speech targeted a certain group. Overall, the entire HateXplain training dataset consisted of 8,492 HCS, 640 HNCS, 1,735 CCS, and 4,516 CNCS. Since HateXplain contained 15,383 carefully annotated phrases, we utilized this dataset for the experiments.

All experiments were conducted in Jupyter Notebook with a Dell Inspiron 16 laptop, equipped with an Intel Core i5 processor. After the experiments, the data was exported to Google Sheets where the resulting graphs and tables were generated.

CaWT Procedure

In total, we ran 6 parent experiments with the following CWs: 70%, 75%, 80%, 85%, 90%, 95%. Within each parent experiment, we conducted 10 sub-experiments with the intent of creating more accurate data, through multiple tests, to represent each parent experiment's CW. These sub-experiments maintained the same CW and class distribution as all the other sub-experiments within the parent experiment. For example, the 10 sub-experiments within the 70% CW parent experiment, all maintained 70% as the CW but consisted of unique phrases to fill each category, due to the random parsing of the program (Table 1). Each sub-experiment involved running the output training dataset with the corresponding CW against the SVM algorithm in order to train it. Once trained, the SVM algorithm was then run against a testing dataset and the scores were printed.

Constructing Datasets

Furthermore, to create each training dataset efficiently, we wrote a scraping program to collect data phrases from the published dataset to create our proposed method. The program can parse through HateXplain with a given percentage, and based on that percentage, outputs a new dataset with cultural speech representing that percentage of the new CWTD. Additionally, the program ensures a 1:1 ratio of HS to clean speech within each CWTD and category: HCS, HNCS, CCS, CNCS (Table 1). While parsing, the program selects random phrases within the correct category. Particularly, the 10 sub-experiment CWTDs for a certain CW will all contain the same amount of phrases in each category with a 1:1 HS to clean speech ratio but all the phrases may be different. Furthermore, the overall dataset sizes remained the same for experimental purposes while the CW percentage varied (Table 1). The overall dataset size was determined to be 3,640 since it is the largest size that accommodates the CWs – 70%, 75%, 80%, 85%, 90%, and 95% – while maintaining the range limitations of each category of HateXplain (Table 1). Specifically, since the amount of HS equals the amount of clean speech, each CW multiplied with the dataset size must be an even number. Additionally, since HateXplain only contained 640 HNCS phrases, this limited the datasets' sizes. The largest size that satisfies these constraints is 3,640. For this reason, if given the input 0.7, the program outputs a training dataset derived from HateXplain which contains 70% cultural speech and 1,274 HCS phrases (Table 1).

The program also selected the first 910 phrases – 502 HCS, 96 CCS, 282 CNCS, 30 HNCS – of the HateXplain testing dataset to create a new testing dataset for the experiments (Table 1). HateXplain's testing dataset is completely different from the training dataset, ensuring that all the testing phrases are unique from the training phrases. The fact that all the HateXplain datasets are in a random order justifies only selecting a specific portion of the dataset. Importantly, the CW does not affect the testing dataset whatsoever. Instead, this dataset remains static with the same 910 phrases throughout the full procedure.

ACKNOWLEDGMENTS

We would like to express our gratitude to those who helped us in our research process. Professor Sandra Kübler from Indiana University, Bloomington exposed Shlok to the world of machine learning and computational linguistics. She also connected him with Zuoyu Tian, a PhD candidate for Linguistics at Indiana University, Bloomington. Tian provided Shlok with the foundational knowledge in machine learning to conduct the experimentation. To that, we express our appreciation to both Tian and Professor Kübler for taking time out of their busy schedule to aid in our research.

Finally, Shlok would like to thank his family for supporting him throughout his research. Despite his initial unfamiliarity

Number of Phrases in Each Category								
Sum of Phrases in Category		*Cultural Weight*						*Testing Dataset*
Grouped Category	*Category*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*	
Cultural Speech	Hate	1274	1365	1456	1547	1638	1729	502
	Clean	1274	1365	1456	1547	1638	1729	96
Cultural Speech Total		**2548**	**2730**	**2912**	**3094**	**3276**	**3458**	**598**
Non-Cultural Speech	Clean	546	455	364	273	182	91	282
	Hate	546	455	364	273	182	91	30
Not Cultural Speech Total		**1092**	**910**	**728**	**546**	**364**	**182**	**312**
Grand Total		**3640**	**3640**	**3640**	**3640**	**3640**	**3640**	**910**

Table 1: Number of phrases in each category. The table displays the number of phrases in each category for all six parent experiments and the testing dataset.

DOI: https://doi.org/10.59720/23-156

with machine learning, they encouraged him to take on the challenge, nonetheless.

Thanks to all those involved for helping Shlok discover his new passion for machine learning and computational linguistics.

Received: June 11, 2023
Accepted: September 14, 2023
Published: March 20, 2024

REFERENCES

1. Kovács, György, et al. "Challenges of Hate Speech Detection in Social Media." *SN Computer Science*, vol. 2, no. 2, 2021, pp. 1-15. https://doi:10.1007/s42979-021-00457-3
2. Wiegand, Michael, et al. "Detection of Abusive Language: The Problem of Biased Datasets." *Association for Computational Linguistics*, vol. 1, 2019. https://doi:10.18653/v1/N19-1060
3. García-Díaz, José A., et al. "Evaluating Feature Combination Strategies for Hate-speech Detection in Spanish Using Linguistic Features and Transformers." *Complex & Intelligent Systems*, 2022, pp. 2893–2914. https://doi:10.1007/s40747-022-00693-x
4. Mahesh, Batta. "Machine Learning Algorithms - A Review." *International Journal of Science and Research*, vol. 9, no. 1, 2020, pp. 381-86. https://doi:10.21275/ART20203995
5. Park, Ji H., et al. "Reducing Gender Bias in Abusive Language Detection." *Association for Computational Linguistics*, 2018, pp. 2799-804. https://doi:10.18653/v1/D18-1302
6. MacAvaney, Sean, et al. "Hate Speech Detection: Challenges and Solutions." *PLOS One*, vol. 14, no. 8, 2019, pp. 1-16. https://doi10.1371/journal.pone.0221152
7. Sap, Maarten, et al. "The Risk of Racial Bias in Hate Speech Detection." *Association for Computational Linguistics*, 2019, pp. 1668-78. https://doi:10.18653/v1/P19-1163
8. Schmidt, Anna and Michael Wiegand. "A Survey on Hate Speech Detection Using Natural Language Processing." *Association for Computational Linguistics*, 2017, pp. 1-10. https://doi:10.18653/v1/W17-1101
9. Razo, D and Sandra Kübler. "Investigating Sampling Bias in Abusive Language Detection." *Association for Computational Linguistics*, 2020. https://doi:10.18653/v1/P17
10. Buckland, Michael and Fredric Gey. "The Relationship between Recall and Precision." *Journal of the American Society for Information Science*, vol. 45, 1994. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L
11. Mathew, Binny, et al. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14867-75. https://doi:10.1609/aaai.v35i17.17745