A Machine Learning Approach to Identify Toxic Language in the Online Space

Lisa Kaati
Stockholm University
Stockholm, Sweden
Email: lisa.kaati@dsv.su.se

Amendra Shrestha
Mind Intelligence Lab
Uppsala, Sweden
Email: amendra@mindintelligencelab.com

Nazar Akrami
Uppsala University
Uppsala, Sweden
Email: nazar.akrami@psyk.uu.se

***Abstract*—In this study, we trained three machine learning models to detect toxic language on social media. These models were trained using data from diverse sources to ensure that the models have a broad understanding of toxic language. Next, we evaluate the performance of our models on a dataset with samples of data from a large number of diverse online forums. The test dataset was annotated by three independent annotators. We also compared the performance of our models with Perspective API - a toxic language detection model created by Jigsaw and Google's Counter Abuse Technology team. The results showed that our classification models performed well on data from the domains they were trained on (F1 = 0.91, 0.91, & 0.84, for the RoBERTa, BERT, & SVM respectively), but the performance decreased when they were tested on annotated data from new domains (F1 = 0.80, 0.61, 0.49, & 0.77, for the RoBERTa, BERT, SVM, & Google perspective, respectively). Finally, we used the best-performing model on the test data (RoBERTa, ROC = 0.86) to examine the frequency (/proportion) of toxic language in 21 diverse forums. The results of these analyses showed that forums for general discussions with moderation (e.g., Alternate history) had much lower proportions of toxic language compared to those with minimal moderation (e.g., 8Kun). Although highlighting the complexity of detecting toxic language, our results show that model performance can be improved by using a diverse dataset when building new models. We conclude by discussing the implication of our findings and some directions for future research.**

I. Introduction

Toxic language is widespread on social media platforms and includes content that, for example, incites violence, expresses direct hate towards individuals and groups, and extremist propaganda. For the target of a toxic message, the content may cause emotional distress, and a toxified space may lead to further toxic language and hate speech and may inspire individuals to commit acts of violence. The problem of hate speech is not specific or limited to a certain territory or culture – it is spread all over the internet. Since it is impossible to monitor everything that is said online manually, the research community is dealing with this issue by exploring technical solutions to detect toxic language. Technological solutions to detect toxic language are used by a number of different actors. For example, social media companies may use automatic methods to detect messages that violate their policies or that violate the law, while law-enforcement use similar technologies to detect threats to the security of individuals and groups in society.

While most people have an intuitive sense of what toxic language is, there is still no consensus on how to define it. Generally, people have different opinions and levels of tolerance for what is considered toxic. Toxic language is also context-dependent: the same person can judge a message differently depending on the context in which it occurs. Most big social media companies have their own policies on what kind of content is allowed on their platforms and what is not. These policies tend to change over time to capture new forms of toxic language.

There are several challenges when developing technologies for detecting toxic language. These challenges are concerned with different areas such as definitions of what should be considered toxic or not, lack of high-quality training data, language and domain-dependent algorithms, the performance of algorithms on new types of data, and biases in the results of the algorithms. The aim of this paper is to look into some of these challenges. As a first step, we created three different machine learning models that can recognize toxic language. Second, we have created a dataset that can be used for testing and evaluation of toxic language. The dataset consists of comments from 21 different online environments, and the data is different from what the models are trained on. We use the dataset to evaluate our models and to compare the performance of our models with Google Perspective - a free API that uses machine learning to identify toxic comments. Finally, we also assessed the level of toxicity on 21 different platforms to explore how well our models perform and to compare the level of toxicity in different online environments.

Outline

This paper is outlined as follows. In Section II we describe some of the previous work on automatic detection of toxic language. In Section III we describe how we have trained three different models to detect toxic language. Section IV contains

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

a description of how we created a dataset for evaluation and comparison of our models and Google Perspective API. Section V contains an analysis of the level of toxic language on 21 different forums/platforms. Section VI contains a discussion of the results and finally, some conclusions and directions for future research are presented in Section VII.

II. Related work

Toxic language is a broad term that captures various forms of offensive or harmful language. In [18] toxic language is described as an "umbrella term" that comprises several different types of language, including offensive language, abusive language, and hateful language. Google's Counter Abuse Technology team who developed the Perspective API described toxic language in online conversations as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion".

The majority of automated approaches for identifying toxic language are aimed to classify a piece of text as toxic or not toxic. In some cases, the text is classified into a specific type of toxic language, such as misogyny, antisemitism, xenophobia, abusive language, or threats. Most automated toxic language detection technologies rely on natural language processing or text mining technologies.

There have been many attempts to detect toxic language in previous research. The simplest of these approaches are dictionary-based methods, which involve developing a list of toxic words. If a word from the toxic list is present in a text – the text is considered toxic. A dictionary-based approach to detecting toxic language in Swedish is described in [11]. There are also dictionary-based approaches that have been extended and combined in various ways. One approach that includes the use of a combination of natural language processing and automated reasoning to detect directions of toxic language is described in [19], and another approach uses dictionaries containing implicit and explicit offensive and swearing expressions annotated with contextual information is presented in [21]. In [8] two different types of hate speech are analyzed: generalized hate and targeted hate (directed at individuals or entities). Directed hate is defined as hate language towards a specific individual or entity, while generalized hate is defined as hate language towards a general group of individuals who share a common protected characteristic, e.g., ethnicity or sexual orientation. Other approaches have considered threatening language and in [4] a threat dictionary that can be used to index threat levels from texts across media platforms is described. The threat dictionary shows convergent validity with objective threats in American history, including violent conflicts, natural disasters, and pathogen outbreaks.

In [16] machine learning is used to separate hate speech, profanity, and other texts. The results showed that distinguishing profanity from hate speech is challenging. The use of bag-of-words approaches tends to have high recall but leads to high rates of false positives since the presence of offensive words can lead to the miss-classification of tweets as hate speech, something that was noticed in [13]. Kwok and Wang found that, often, the reason that a tweet was categorized as racist was because it contained offensive words.

With the recent advancement in text analysis, the use of deep learning-based technologies has become a leading approach for automatic detection of toxic language and hate speech. One example of such an approach can be found in [3] where levels of hate are measured on some online platforms. The approach was later used in [12] to analyze the levels of toxic language used in right-wing extremist communities online.

Despite the many differences in how to detect toxic language, nearly all systems rely on a training dataset, which is used to teach the system what is and is not considered toxic or not. One of the issues with training data is that the resulting models relies heavily on the quality of the data. However, creating training datasets that are large enough, varied, theoretically-sound and that minimize biases is difficult, takes a lot of resources and requires deep expertise [22].

III. Detecting Toxic Language

We have trained and tested three different classification models to detect toxic language. The models are based on Robustly Optimized BERT Pretraining Approach (RoBERTa) [15], Bidirectional Encoder Representations from Transformers (BERT) [7], and Support Vector Machine (SVM).

A. Creating a dataset

To train the models, we used data from 8 different sources. Using data from diverse sources for training ensures that the resulting model has a broad understanding of toxic language. The datasets are from different social media platforms and include different expressions of toxic language. In some of the datasets, the data was labeled with several different categories, such offensive, profanity, abusive, bullying, etc. Since all categories are not related to what we consider toxic language, we discarded some categories and used only two different labels: toxic and non-toxic.

The datasets we have used are the following:

1) **HateXplain** - A dataset that is a combination of Twitter and Gab posts annotated with the labels hate, offensive or normal provided by Mathew et al. [17].
2) **Twitter 1** - A multilingual Twitter dataset for hate speech against immigrants and women provided by Basile et al. [2]. We used the English tweets labeled as hate speech or normal.
3) **Twitter 2** - A dataset with annotated Twitter data provided by Perifanos et al. [20]. We use the English part of the data.
4) **Twitter 3** - A dataset provided by Golbeck et al. [10] with tweets labeled as harassing and non-harassing.
5) **CAALDYC** - A dataset provided by Noman et al. [1] with YouTube comments annotated for abusive language detection.
6) **Forum data** - Annotated dataset with data from a white supremacist discussion forum provided by Gibert et al. [6].

7) **Twitter 4** - A dataset with from Twitter with abuse-related labels provided by Founta et al. [9].
8) **Twitter 5** - A dataset consisting of tweets annotated with the labels hate speech, offensive language, and neither provided by Davidson et al. [5].

TABLE I
THE DATASETS INCLUDED IN THE CONSTRUCTION OF THE FINAL DATASET.

Data source	Toxic	Non-toxic
1. **HateXplain**	Yes	Yes
2. **Twitter 1**	Yes	Yes
3. **Twitter 2**	Yes	No
4. **Twitter 3**	Yes	No
5. **CAALDYC**	No	Yes
6. **Forum data**	No	Yes
7. **Twitter 4**	No	No
8. **Twitter 5**	No	No

As a first step, we trained a RoBERTa model by combining all datasets. The model had a precision of 0.79, recall 0.62 and F1-score 0.70. One way to improve the model is to increase the quality of the data. In order to do so, we trained a RoBERTa model for each dataset. When a model could classify toxic and/or non-toxic posts with an accuracy higher than 80% we used the data for training our final model. If the accuracy was lower than 80% did not use the data.

Table I shows the datasets we included in our construction of a dataset. For some datasets, we only included one class (toxic or non-toxic). Two of the datasets were not used at all.

Our final dataset was constructed by combining 18,259 non-toxic text and 12,576 toxic texts. All texts were converted to lowercase. @user mentions, "#", and ULRs were removed from the text.

B. Classification models

We trained three different models to detect toxic language: a RoBERTa model, a BERT model, and an SVM model. RoBERTa and BERT are language models based on transformer architecture. Instead of training the language models from scratch, we utilized a pre-trained RoBERTa and BERT model made available from the transformers library Huggingface[1] and fine-tuned them with our created dataset.

We used roberta-base version of RoBERTa and bert-base-uncased of BERT. For both models, the maximum sequence of token was fixed to 256 tokens. The experiment was done with 10 epochs, and the batch size was kept at 8. During the training process, we chose the best performing model measured by accuracy on the validation set. In the case with RoBERTa, Adam optimizer was used with a small learning rate of 5e-6. Similarly, while training the BERT model, Rectified Adam optimizer [14] was used with learning rate 3e-5.

In the case of the SVM model, hyper-parameter tuning was done utilizing grid search to estimate the optimal parameters of the classifier. To select features, we used TF-IDF (Term Frequency–Inverse Document Frequency) numerical statistics that reflect how important a word is to a text in a collection of texts. English stop words, i.e., words that do not add much meaning to a sentence, e.g., a, the, is, are, were removed from the text before applying TF-IDF.

The dataset was split into three parts, 75% was used for training, 15% for model validation, and the remaining 10% was kept for testing purposes. We used the same training, validation, and testing dataset to prevent biases in the evaluation of the different models. The results from experiments on the different models are shown in Table II. The RoBERTa model and the BERT model both had a F1-score of 0.91, while the SVM model had a somewhat lower F1-score (0.84).

TABLE II
CLASSIFICATION RESULTS FOR THE THREE DIFFERENT MODELS.

Model	Precison	Recall	F1-score
RoBERTa	0.88	0.93	0.91
BERT	0.93	0.88	0.91
SVM	0.88	0.81	0.84

IV. EVALUATION OF THE TOXIC LANGUAGE DETECTION MODELS

One of the challenges with toxic language detection models is to understand what we can expect when it comes to performance. Most toxic language detection models are trained, evaluated, and tested on similar data (data from the same domain). Machine learning models need to be tested extensively to understand how they operate in the wild – on new unseen data. To examine the performance of the different models, we tested their ability to classify data from different social media platforms.

A. Creating test data

We created a set of randomly selected comments from 21 discussion forums. The idea of choosing different forums was to evaluate how well the different toxic language detection models perform on different kinds of data.

From each of the 21 forums, we randomly selected between 75 and 100 comments. The forums we used and the number of comments from each forum are listed in Table III. Initially, 100 posts from each forum were selected, but some of the posts consisted of images or emojis and could not be classified.

Three independent expert annotators (psychologists and social scientists) independently classified each post as non-toxic, toxic, or unsure. The result of the classification is shown in Table IV. When a majority decision (2 out of 3 raters) was needed to consider a post toxic, 32% of the posts were considered toxic, and 68% of the posts were non-toxic. When a full agreement was needed (3 out of 3 raters) 22% of the posts were considered toxic, 52% of the posts were non-toxic, and 26% could not be classified.

When a majority decision was used to determine if a post was toxic or not, the inter-rater agreement between the raters

[1]https://huggingface.co/docs/transformers/index

TABLE III
THE DIFFERENT FORUMS/ENVIRONMENTS USED TO TEST OUR MODELS.

Forum/environment	Description	Comments
8kun	An imageboard website composed of user-created message boards.	100
Airliners	A forum for discussions about airplane information and aviation.	100
Alternate history	A forum for discussions about alternative history.	96
Bitcoin	A forum for discussions about bitcoins.	100
Blackpill	An incel (Involuntary celibates) discussion forum. Incels are heterosexual men who blame women and society for their lack of romantic success.	97
Bodybuilding forum	A forum for discussions about training and bodybuilding	75
Gates of Vienna	An anti-muslim counter jihad blog portal with comments	100
Incels.co	An incel (Involuntary celibates) discussion forum.	100
Incels.net	An incel (Involuntary celibates) discussion forum.	100
Lookism	An incel (Involuntary celibates) discussion forum with a focus on appearance/looks.	99
Lookmaxxing forum	An incel (Involuntary celibates) discussion forum with a focus on appearance/looks.	94
Looks theory	An incel forum with a focus on theories that try to explain how looks affect attraction.	95
Looksmax	An incel (Involuntary celibates) discussion forum with a focus on appearance/looks.	100
Mumsnet	A forum for discussions about family life and children	99
Ni**ermania	A discussion board with condescending jokes and racist comments about Black people.	100
Non Cucks United	An incel (Involuntary celibates) discussion forum.	98
Stormfront	One of the most well-known white supremacy discussion forums.	100
The Sims	A forum for discussions related to the game Sims.	83
The Suicide Project	A discussion forum for sharing stories of desperation and depression.	99
Women Only Forum	A discussion forum for women only.	100
Youre not alone	An incel (Involuntary celibates) discussion forum.	99

TABLE IV
CLASSIFICATION RESULTS OF 2051 POSTS BY THREE INDEPENDENT RATERS.

Classifier	Non-Toxic	Toxic	Unsure
Raters			
Rater 1	1343 (65%)	680 (33%)	28 (1%)
Rater 2	1365 (66%)	609 (30%)	77 (4%)
Rater 3	1293 (63%)	727 (35%)	31 (2%)
Voting			
Majority of raters	1394 (68%)	657 (32%)	0 (0%)
Full agreement only	1066 (52%)	445 (22%)	540 (26%)

was measured using the Fleiss' kappa reliability measure. The Fleiss' kappa measure was .68 [Confidence Interval, 95%; .66, .71], which can be interpreted as a moderate level of agreement. When a full agreement was used to determine if a post was toxic or not, the Fleiss' kappa reliability measure was 1.0 [Confidence Interval, 95%; .97, 1.0]. The reliability is at the maximum (full agreement between the raters) level since all posts where the raters disagree are removed.

B. Comparing the models

To examine of the performance of the different models we used two different datasets: one with posts that were annotated as toxic or non-toxic with a **majority** decision (2051 posts) and one with **full agreement** which means posts annotated as toxic or non-toxic when all raters agreed on the decision (1511 posts).

We ran our three different models (RoBERTa, BERT, and SVM) and Google Perspective on the test data. The results are shown in Table V and VI. The results show that the quality of the data is important. When the models were run on the test data that was built on a majority decision, the RoBERTa model had a ROC score of 0.79 and Google Perspective .77. The BERT model and the SVM model had lower scores of 0.66 and 0.62, respectively.

When the models were examined on the test data that was built on full agreement from the raters (i.e., only posts that all three raters classified as toxic or non-toxic), the RoBERTa model showed a ROC score of .86 and Google Perspective .84. The BERT model and the SVM model still had lower scores .72 and .66 respectively. As can be seen in Table VI, the RoBERTa model correctly classified 83% of the toxic posts as toxic, and Google Perspective correctly classified 78% of

TABLE V
CLASSIFICATION OF THE POSTS THAT THE MAJORITY OF THE RATERS CLASSIFIED AS NON-TOXIC (N=1394) AND TOXIC (N=657).

	Non-toxic posts (majority)		Toxic posts (majority)	
Model	Classified as toxic	Classified as non-toxic	Classified as toxic	Classified as non-toxic
RoBERTa	212 (15%)	1182 (85%)	477 (73%)	180 (27%)
BERT	66 (5%)	1328 (95%)	245 (37%)	412 (63%)
Google Perspective	199 (14%)	1195 (86%)	448 (68%)	209 (32%)
SVM	100 (7%)	1294 (93%)	199 (30%)	458 (70%)

Model	**ROC score [Confidence Interval, 95%]**	**F1-score**
RoBERTa	.79 [.76, .81]	0.71
BERT	.66 [.64, .69]	0.51
Google perspective	.77 [.75, .79]	0.69
SVM	.62 [.89, .64]	0.42

TABLE VI
CLASSIFICATION OF THE MESSAGES THAT ALL RATERS RATERS CLASSIFIED AS NON-TOXIC (N=1066) AND TOXIC (N=445).

	Non-toxic posts (full agreement)		Toxic posts (full agreement)	
Model	Classified as toxic	Classified as non-toxic	Classified as toxic	Classified as non-toxic
RoBERTa	114 (11%)	952 (89%)	371 (83%)	74 (17%)
BERT	36 (3%)	1030 (97%)	209 (47%)	236 (53%)
Google Perspective	105 (10%)	961 (90%)	346 (78%)	99 (22%)
SVM	64 (6%)	1002 (94%)	166 (37%)	279 (63%)

Model	**ROC score [Confidence Interval, 95%]**	**F1-score**
RoBERTa	.86 [.84, .89]	0.80
BERT	.72 [.69, .75]	0.61
Google perspective	.84 [.82, .86]	0.77
SVM	.66 [.62, .69]	0.49

the toxic posts as toxic. The BERT and SVM models did not perform well and could only correctly classify 47% and 37% of the posts, respectively. When classifying non-toxic posts, the BERT and the SVM models performed best with 97% and 94% correct classification, respectively. Google Perspective and the RoBERTa model correctly classified 90% and 89% of the posts.

V. LEVELS OF TOXICITY ONLINE

There are many ways to use automatic classification to analyze toxic language online. One way is to use toxic language classification is to measure the level of toxic language in different online environments. If the level of toxic language is computed using the same methods, the level of toxic language can be compared in a meaningful way. It is also possible to analyze the level of toxic language and changes over time. To estimate the level of toxic language in different digital environments, we used the RoBERTa model. For a one-year period (2020), we selected a representative sample of posts from 21 different forums, the sample size was chosen to achieve a margin of error below 1%, and a confidence level above 95%. Table VII shows the sample size that was selected from each forum/environment and the number of posts that were posted during 2020. The level of toxic language is presented as the percentage of posts that contains toxic language in the sample set.

TABLE VII
NUMBER OF POSTS SELECTED FROM EACH FORUM/ENVIRONMENT.

Forum/environment	**Year 2020**	**Sample size**
8kun	141586	8994
Airliners	108674	8824
Alternate history	381569	9368
Bitcoin	121263	8899
Blackpill	36247	7593
Bodybuilding.com forum	28635	7192
Gates of Vienna	11340	5200
incels.co	2370743	9565
Incels.net	203439	4249
Lookism	1872070	9555
Lookmaxxing forum	30658	7313
Looks theory	125380	8921
Looksmax	3127130	9575
Mumsnet	47235	7981
Ni**ermania	34377	7509
Non cucks united	7620	4249
Stormfront	175799	9107
The Sims	166536	9080
The suicide project	1714	1455
Women only forum	11249	5181
Youre not alone	7301	4148

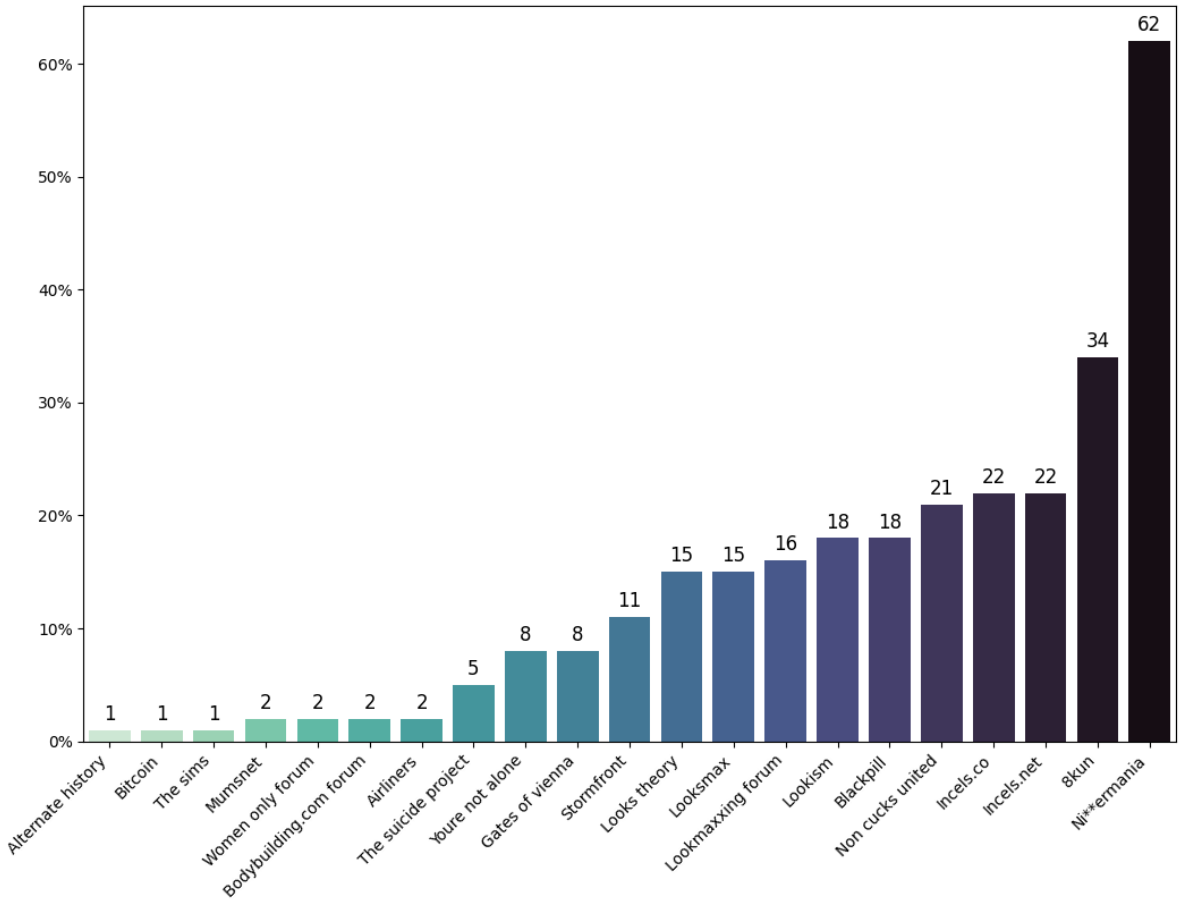

Fig. 1. Levels of toxic language in a number of digital environments.

Figure 1 shows the levels of toxic language in the different environments. Ni**ermainia had the highest proportion of toxic language with more than 60% of the posts containing toxic language. Around 34% of the posts on 8kun contained toxic language. The level of toxic language on the different incel forums varied between 8% and 22%. The white supremacy forum Stormfront had a toxicity level of 11%. The least amount of toxic language was found on Alternate history, The Sims, and Bitcoin with 1%.

The results reported above are in line with what can be expected considering the nature of the studied forums. The forums for general discussions that have moderation have a lower amount of toxic language than forums such as N**ggermania, which is dominated by aggressive racist jokes and images. The level of toxicity on the several different incel forums that are analyzed differs. As expected, the forums that are more focused on discussions about appearance and looks have a somewhat lower level of toxicity compared to the incel forums that have more general incel discussions.

VI. Discussion

We used several different datasets to create a set of models for recognizing toxic language. Using different datasets was aimed to obtain a model that could recognize toxicity on many platforms. When testing our models (on similar data as the training data), the F1-score was 0.91 for the RoBERTa and the BERT models and 0.84 for the SVM model.

To test how well the classification models work in the wild, we used a dataset consisting of annotated posts from 21 different online environments. The data was annotated by three independent expert annotators. When creating the test data, 26% of the post was marked as unsure, which means that at least one of the three annotators was not sure about the character of the post. The Fleiss' kappa reliability measure was 0.68, which indicates that classification of toxicity is a difficult problem. To examine the performance of our models, we created two different test datasets: one with posts that all annotators agreed were toxic or non-toxic and one with posts that were annotated as toxic or non-toxic with a majority decision.

The RoBERTa model and Google Perspective had the best performance on the test data, with the RoBERTa model slightly better. Both models had a ROC score over 0.84 on the dataset with full agreement.

The results from the tests of the models on the created test data reveal some interesting observations. The performance decreases significantly: for the RoBERTa model, the F1-score decreases to 0.8 on the full agreement dataset and 0.71 on the

majority dataset. For the BERT model, the F1-score decreases to 0.61 (full agreement) and 0.51 (majority). For the SVM the decrease is even higher: the F1-score goes from 0.84 to 0.49 (full agreement) and 0.42 (majority). These results show that the classification of toxic language is a challenge and that applying machine learning models to different domains is coupled with challenges too. Specifically, all models had a decent performance on the same type of data as they were trained on, but when we used data from other domains, the performance decreased to such an extent that some of the models were useless.

VII. Conclusions and Future Work

Our results show that classifying online posts as toxic or not is a complex problem - for both humans and computer algorithms. When using classification models as a tool to detect toxic language, it is essential to take into consideration that many models perform well on similar data as they are trained on, but the performance decreases significantly when they are run on new unseen data.

There are several interesting directions for future work. One interesting direction would be to improve the models by providing more training data. In particular, data that the models seem to have difficulties in classifying. An in-depth analysis of the miss-classified posts is needed to understand more about what kind of training data needs to be provided. A study of what kind of biases are built into the models would also be both natural and highly relevant direction for future work.

Acknowledgements

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala University, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

[1] N. Ashraf, A. Zubiaga, and A. Gelbukh. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ. Computer science*, 7, 2021.

[2] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[3] T. Berglind, B. Pelzer, and L. Kaati. Levels of hate in online environments. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 842–847, 2019.

[4] V. K. Choi, S. Shrestha, X. Pan, and M. J. Gelfand. When danger strikes: A linguistic tool for tracking america’s collective response to threats. *Proceedings of the National Academy of Sciences*, 119(4):e2113891119, 2022.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May 2017.

[6] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[8] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *ICWSM*, pages 42–51, 2018.

[9] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of International Conference on Web and Social Media(ICWSM)*, pages 491–500. AAAI Press, 2018.

[10] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, Q. Gergory, R. K. Gnanasekaran, R. R. Gunasekaran, K. M. Hoffman, J. Hottle, V. Jienjitlert, S. Khare, R. Lau, M. J. Martindale, S. Naik, H. L. Nixon, P. Ramachandran, K. M. Rogers, L. Rogers, M. S. Sarin, G. Shahane, J. Thanki, P. Vengataraman, Z. Wan, and D. M. Wu. A large labeled corpus for online harassment research. In P. Fox, D. L. McGuinness, L. Poirier, P. Boldi, and K. Kinder-Kurlanda, editors, *Proceedings of the 2017 ACM on Web Science Conference, WebSci*, pages 229–233. ACM, 2017.

[11] T. Isbister, M. Sahlgren, L. Kaati, M. Obaidi, and N. Akrami. Monitoring Targeted Hate in Online Environments. *Second workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*, 2018.

[12] L. Kaati, K. Cohen, and B. Pelzer. *Heroes and scapegoats : right-wing extremism in digital environments*. European Commission and Directorate-General for Justice and Consumers. Publications Office, 2021.

[13] I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In M. desJardins and M. L. Littman, editors, *AAAI*. AAAI Press, 2013.

[14] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *CoRR*, abs/1908.03265, 2019.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[16] S. Malmasi and M. Zampieri. Detecting hate speech in social media. *CoRR*, abs/1712.06427, 2017.

[17] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, 2021.

[18] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. Toxicity detection: Does context really matter? In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *ACL*, pages 4296–4305. Association for Computational Linguistics, 2020.

[19] B. Pelzer, L. Kaati, and N. Akrami. Directed digital hate. In *ISI*, pages 205–210. IEEE, 2018.

[20] K. Perifanos and D. Goutsos. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7), 2021.

[21] F. Vargas, F. Rodrigues de Góes, I. Carvalho, F. Benevenuto, and T. Pardo. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online, Sept. 2021. INCOMA Ltd.

[22] B. Vidgen and L. Derczynski. Directions in abusive language training data: Garbage in, garbage out. *PLoS ONE*, 15(12), 2020.