AI detection of toxic comments

Exam:
Bachelor Project for B.Sc. Humanities and Technology, Summer 2022

Exam group no. S2225625766
Group members:

- Anne-Thea Rendtorff Uldal (atru@ruc.dk, study no. 68892)
- Oliver Weisel (oweis@ruc.dk, study no. 69226)
- Rasmus Rosendal Nielsen (raroni@ruc.dk, study no. 68843)
- Simone Emilie Jensen (siemje@ruc.dk, study no. 68939)

Supervisor:

- Jens Classen (classen@ruc.dk)

Abstract

This paper is a research of the challenges associated with detection of online toxicity. It is primarily based on a machine learning challenge posted to Kaggle.com by Jigsaw and Conversation AI. Thus our research on machine learning and AI identification of toxic comments is based on an attempt to fulfill this challenge.

Through an applied use of feministic epistemologies and feministic technoscience, this paper will elaborate on biases of language in correlation to the toxic comments and look into the occurrence of biases in gendered and politicized language.

As no universal formal definition of toxicity exists, this paper explores common traits between different contextual definitions of toxicity. We have found that considerations of intent vs. reaction are deeply essential to understanding the challenge of defining toxicity.

It can also be concluded that a simple and contextually relevant definition of toxicity is necessary to create data that is applicable for machine learning. Furthermore, this is confirmed by our applied dataset, which had the inherent definition of toxicity as detrimental to fair discussion and discouraging participation.

Using this definition of toxicity combined with binary classification gives the AI the ability to specifically choose between ‘toxic’ and ‘non-toxic’ when served a comment. And thus identification of toxicity can occur.

However, we find that this AI reflects the flaws of input and that the use of our model is generally hard to qualify. Thus highlighting the primary challenges of creating toxicity detection AI: 1. creating a valid interpretation of toxicity within the scope of machine learning, 2. Understanding the way our data inflicts bias on our AI and 3. Justifying the use of AI despite potential consequences.

Table of content

1. Introduction

The meaning of the word ‘toxic’ might be different depending on the use of the words, the intent of the user, and the social situation in which it was situated. For this reason, communication and understanding within this field are somewhat difficult in a real-life setting. Especially online, where the definition varies from platform to platform (Sheth et al., 2021).

This paper examines toxicity in an internet context and will attempt to uncover what challenges are associated with toxicity detection, identification, and definition. Thus we intend to study ML with the goal of identifying toxicity in comment sections. The paper will have the scientific approach of feministic epistemologies and feministic technoscience, which means the paper will look at gendered and politicized language to help identify potential biases in the AI and supporting datasets.

1.1 Problem Area

The presence of toxic content online has become a major issue for online platforms due to increased international contact and exchanges between countries. Social toxicity is commonly described in correlation to categories such as cyberbullying, racism, sexual predation, and other behaviors dubbed negative to society (Brassard-Gourdeau, E & Khoury, R. 2019).

Toxicity is difficult to define and take action against because it seems to have no formal definition. However, two main strategies were carried out by online communities to prevent toxicity from spreading; automated filtering and human surveillance. But given the sheer amount of online messages, it was realized that it was unrealistic for moderators to keep up with the message stream. Their solution was to pre-moderate comments, but this significantly slowed down conversations. Though simple toxic filtering later emerged, toxic users still managed to circumvent it (Brassard-Gourdeau, E & Khoury, R. 2019).

Given the progressive increase in online participation after the Covid-19 pandemic, a large number of communications moved to online platforms. The consequence of this is another rise in toxicity (Adinolf & Turkay, 2018). In light of these events, the problem of detecting toxicity remains of interest to both users and platform providers as it begs the question of whether it is even possible to properly moderate toxicity using AI technology.

Using artificial intelligence to handle toxicity moderation could potentially help online communities and their moderators. On the other hand, this technology also has the potential of being abused or unacceptably fallible. These challenges among other challenges of creating and implementing toxicity detection AI will be the subject of this research paper.

1.2 Problem Formulation

What are the current challenges in identifying and filtering “toxic” content on the internet using Machine Learning and AI?

1.3 Research Questions

- What defines content as “toxic”?
- How is toxicity interpreted in our dataset?
- How can we create an AI that identifies toxicity?

2. Theory & Methods

2.1 Theory

Fry: The Seven Stages of Visualizing Data

In this project, we use data visualization to present our dataset. We approach the task of visualizing data based on Ben Fry's data visualization theory (2007). This theory was selected as it provides structure and knowledge on optimal data visualization. The theory proposes that the optimal way to design data visualization is based on seven stages: Acquire, Parse, Filter, Mine, Represent, Refine and Interact.

The dataset we are working with has already been preprocessed to some extent. Furthermore, our primary focus is not to create perfect data visualization, but rather to communicate our results. For this reason, we judged the stages; Parse, Mine, and Represent to be most relevant to our work. In the following section, we will explain Fry's general theory and the specific traits of these three stages.

Ben Fry developed his theory based on a few key issues within the field. One such issue is 'information overload', which is the issue of too much information being presented at once. This is also referred to as the 'All-You-Can-Eat Buffet' of data visualization. He points out that sometimes less data is more telling when it comes to visualizing a pattern. In other words, it is important to distinguish necessary information from everything else (Fry, B. 2007).

Another issue is the problem of the presentation not fitting the unique properties of the data. While a bar-chart or a simple scatterplot can sometimes be the best way of presenting data, it is not always the case. Fry makes the point that careful consideration should go into choosing the visual presentation so that it best suits the specific properties of the data and the intended message (Fry, B. 2007).

To address these issues, he proposes that any data visualization project should follow the seven stages. One of these stages is the Parse-stage. The general goal of this stage is to provide a general structure for the acquired data. This is done by sorting it into relevant categories (Fry, B. 2007).

One of the following stages is the Mine-stage. This stage is where basic statistics and other strategies are applied to reveal patterns within the data (Fry, B. 2007).

Finally, we include the Represent-stage, which presents the data through simple models. In this stage it is not relevant whether the model is interesting or interactive, rather it is most important that the intended information is communicated clearly (Fry, B. 2007).

Fry's theory generally serves to inform us of common problems with data visualization and how to avoid them. Furthermore, our three selected stages provide some structure and method to our visualizations.

Langdon Winner

As we examine how to identify and filter toxic content on the internet, it is necessary to debate our results. We argue that this best be done using several perspectives/theories to ensure our results do not solely reflect

our own bias. Langdon Winner's (1989) theory on artifacts represents one of these perspectives. This section will therefore introduce and argue Winner's (1989) utility value for this project.

Langdon Winner (1989) is known as the philosopher, who first presented the controversial notion that technical things can have political qualities. One of his core principles was that technology has the ability to direct societal change and we should:

> "*…pay attention not only to the making of physical instruments and processes, although that certainly remains important, but also to the production of psychological, social, and political conditions as apart of any significant technical change*" (Winner, L. 1989. P. 17, line 16-18).

He argues that artifacts have effects on their surroundings. Whether big or small these effects have certain causality with greater events. Due to the gap between intended utility and actual utility, Winner's arguments may be considered important for both creators and consumers of new technology, because it represents the notion that there may be risks involved with the creation of new technology.

However, Winner also points out that "*to recognize the political dimensions in the shapes of technology does not require that we look for conscious conspiracies or malicious intentions*" (Winner, L. 1989. p 125, line 28-30). Meaning that while artifacts have the potential to be used negatively, this may not always be the case. The point is that we should not view new technology from a perspective where we only see the consequences.

Winner's theory also implicitly presents two choices. The choice of adopting new technology or not and the choice of considering implicit political features of a technology. Conclusively Winner's (1989) theory encourages discussion of technologies and their uses. In the case of toxicity, we believe Winner's theory may be adept at discussing the possible uses and misuses of being able to detect toxicity.

AI & Machine learning

Artificial intelligence (AI) is a branch of Computer Science that specializes in research on the improvement of a computer system's ability to mimic human cognitive functions using math and logic (Kotsiantis, S.B 2007).

In more accurate terms it can be described as technology used to create 'intelligent systems that can simulate human intelligence'. AI can be recognized by the fact that its system does not require to be pre-programmed explicitly. Instead, it uses algorithms to achieve new information as an artificial way of learning. By doing this the AI can evolve (Kotsiantis, S.B. 2007).

Machine Learning (ML) is a subbranch of AI primarily dedicated to extracting knowledge from data to enable machines to learn. Conversely to AI, which is a technology, ML can be considered a method. ML automates the creation of computational intelligence by enabling systems to learn without explicitly programming it (Kotsiantis, S.B. 2007). In conclusion, ML is therefore always AI, but AI doesn't necessarily have to be ML. ML can therefore be said to be the 'application of AI'.

In this project, we are using ML knowledge and principles to examine what challenges there are to creating an AI that can detect toxicity. The reason is that an AI system can be built using ML learning. Combining AI and ML would be a benefit for us because it would enable us to use more sources of input, increase operational efficiency and improve data integrity by reducing the chances of human error like miscalculation or oversight on details. The approach of making an AI using ML is furthermore attractive because it makes the predictive analytics of the task faster and the result more stable due to its benefits of computationally enhancing the process (Kotsiantis, S.B 2007).

However, while this approach is efficient, it is important to be aware that any results made by an AI created using ML contain bias based on its data. Whether this theory can be ideally applied to solve how to detect toxicity remains to be examined. But because our approach is limited by our specific knowledge of the subject, we are aware that there are likely better, but more advanced ML approaches. In conclusion, we are not able to argue that our approach is 'the best'. However, we can conclude on the benefits and challenges of our examined approach.

Supervised vs. unsupervised learning.

Supervised and unsupervised learning are both examples of different types of machine learning approaches. The approaches differ in the way that the ML models are trained and what conditions are required of the training set. An ML model consists of an algorithm trained by data called a 'training set', which either consists of labeled or unlabeled data (Surbhi, A. 2020).

In the case of supervised learning, the model learns using a labeled dataset to generate predictions for a problem. Unsupervised Learning is the exact opposite. The unsupervised learning model is self-organizing, meaning the model learns from discovering new patterns in an unlabeled dataset to predict an outcome instead of using comparison with an 'answer sheet' like supervised learning (Surbhi, A. 2020).

One model is therefore trained on what to recognize (supervised), while another does data mining (unsupervised). Here data mining refers to the process of finding patterns and correlations within large data sets to predict outcomes without having a predetermined answer (Surbhi, A. 2020).

Both approaches have in common that the input must be viable to read for the ML algorithm to function correctly. Therefore, data preprocessing is needed for the dataset to be used for the ML algorithms (Surbhi, A. 2020).

No approach is necessarily better than the other. They simply have different application uses. In our case, the application of supervised learning is more suited to examine our research area due to the nature of our problem being the detection/identification of known categories rather than searching for patterns in unstructured data. We will therefore examine toxicity detection using that approach.

However, as labeled data is a time-demanding resource to create, we have chosen to use a pre-prepared training set to accommodate this aspect of the theory in our study. This conclusion will be further elaborated on in the problem analysis.

Classification

In ML, classification is a supervised learning approach where an ML model learns from the data input given to it and then uses the input to categorically classify new observations (Brownlee, J. April 8, 2020).

There are different types of classification. For example, 'binary classification' and multi-classification. The type of classification is determined by what kind of distribution the classification is done in (Brownlee, J. April 8, 2020).

Binary classification is when a classification task has a binary outcome. For example, If the classification task is to label any given element as one of two classes, then it is 'binary classification' because it has a Bernoulli Distribution (a binary outcome) (Brownlee, J. April 8, 2020).

Multiclassification is on the other hand called a 'Multinoulli probability distribution' (categorical distribution) and is an outcome considering more than two states. Meaning the prediction task will have one of K's possible outcomes instead of one of two outcomes (Brownlee, J. April 8, 2020).

In this study, we can conclude binary classification can be used to examine toxicity based on the fact we are expecting a binary outcome for our classification, namely toxic/non-toxic. In this scope, an ML algorithm could therefore either be Logistic Regression, K-Nearest Neighbors, or Decision Trees as these are all ways to do binary classification. Using these methods is the same as choosing a 'classifier'

Logistic regression

Logistic regression is a type of regression analysis, but despite its name, it is a classification algorithm rather than a regression model (Thanda, A. 2022).

Regression analysis is a type of predictive modeling technique that can be used to find the relationship between a dependent variable (Y) and the independent variable (X). It can be used to forecast the effects, trends, future values, or the impact of an event (Thanda, A. 2022).

The primary difference between a classic linear regression model and a logistic regression model is however that a logistic regression model's range is bounded between 0 and 1 (Thanda, A. 2022).

In addition, logistic regression assumes the dependent variable to be either binary or dichotomous. Meaning it should have only two outcomes like 'toxic' and 'non-toxic' or be divided into two distinct groups that are mutually exclusive or contradictory. Basically, 'dichotomous' means that there should not be a high correlation between the independent variables (Thanda, A. 2022).

This type of regression is used to calculate the probability of a binary event occurring (Thanda, A. 2022), and we can therefore use it to predict a binary outcome like 'toxic' and 'non-toxic', especially because these groups are mutually exclusive. For example, if something is indeed toxic it cannot be non-toxic.

From a mathematical perspective, Logistic regression is strictly convex. A strictly convex function is a function where a straight line between any pair of points on the curve of the logistic function is above the curve. Except for the intersection points between the straight line and the curve itself (Thanda, A. 2022).

This translates to Logistic regression being faster from an optimization perspective because this allows there to be a ‘global minimum’ as searching for one point can allow searching for another. However, this does not necessarily translate to how long you wait for the model to give an output.

K-nearest Neighbors (KNN)

KNN is one of the simplest and most widely used classification algorithms. The concept is that any new data point will be classified based on similarity in the specific group of neighboring data points. This is done by finding the distance between the new data point and any K number of other datapoints and then evaluating the frequency of those points occurring near the initial point (Saji, B. 2021).

The result is labeled data placed in a space. However, this has the downside of the algorithm's time complexity potentially being quite slow based on the quantity of data the model is working with to do a classification task.

KNN is considered a ‘lazy learning algorithm’, which means it doesn’t perform any training when it is supplied training data. Instead, it stores the data doing the training time but only does the calculations when a query is performed on the dataset (Joby, A. 2021). Overall, this means that while KNN is predominantly used as a classifier it can possibly be considered more suitable for data mining.

This classifier is a non-parametric method, which means it actively can be used to avoid overfitting of the model because it doesn’t make any assumptions about the underlying data distribution (Joby, A. 2021). The KNN classifier is therefore suitable when using non-linear data like a binary outcome.

Decision Trees

A Decision tree is a flowchart-like tree structure. Here each node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

The three’s construction is ‘learned’ by splitting the training into subsets based on an attribute value test, which is derived recursively (GeeksForGeeks. 2022).

For example, the data is split into sets based on features considered toxic by finding the frequency of these attributes occurring when something is labeled toxic. This is called ‘recursive partitioning’.

Overall, Decision trees can handle high dimensional data, because of their structure. A decision three is furthermore considered an inductive approach (GeeksForGeeks. 2022). Meaning that its body of observations (the subsets) is derived from a general principle.

True vs. False vs. Positive vs. Negative

In an ML model there can be four possible outcomes to a prediction: True Positive (TP), False Positive (FP), False Negative (FN) & True Negative (TN).

- A true positive is an outcome where the model correctly predicts the positive class.
- A true negative is an outcome where the model correctly predicts the negative class.
- A false positive is an outcome where the model incorrectly predicts the positive class.

- And a false negative is however when the model incorrectly predicts the negative class

(Swalin, A. May 2, 2018).

The issue with these four outcomes is that while a true positive/negative outcome is desirable because it signifies the model works correctly, a false positive/negative outcome is the opposite because it signifies the model works incorrectly.

To use an example, a false positive would in our case result in a non-toxic comment being found toxic, and a false negative would be an outcome where a toxic comment is found non-toxic. Wrongfully deeming a non-toxic comment toxic or overlooking a toxic comment, because of such an error in an ML model's output, is not ideal.

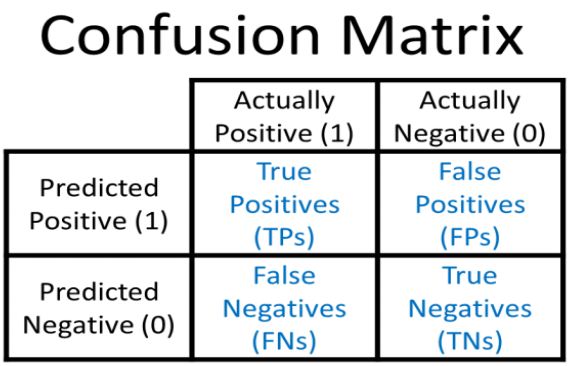
Another problem with false positives and false negatives are that they are hard to identify. This is because examining the ML model for these problems becomes a case of 'innocent until found guilty' (Manoa. n.d.). Meaning that if there is a lack of evidence to prove the model's results wrong, then the probability of simply incorrectly accepting the result as right becomes higher. This is especially the case when testing unlabeled input since labeled data works like an answer sheet to check and verify possible mistakes.

Performance metrics

Performance metrics can be described as figures or data which represent a model's actions, abilities, or overall quality. It can therefore be used to evaluate a model. For classification, the metrics precision, recall, and accuracy can be used to determine if the model is good or bad (Draelos, R. 2019).

Precision is how often a positive classification is correct or rather how often the model has TP outcomes, while Recall is how many of the TP outcomes are classified as positive/correctly. Finally, Accuracy is the fraction of the time when the classifier gives the correct classification. In other words, the time elapsed from start to end of the model's computation (Draelos, R. 2019).

A confusion matrix is on the other hand a table to visualize the performance of the model. The table consists of the model outcomes as seen below:



(Draelos, R. 2019)

2.2 Methods

GitHub

To more efficiently manage and track our progress for our toxicity detection experiment, we have used GitHub to help us facilitate a seamless collaboration on the code to avoid compromising our time-frame on other parts of the project.

Git is a freely accessible version control system. In other words, a free tool that can keep track of files and back-up all versions of them (Gaba, I. 2022).

This type of tool is widely used in software development due to its advantage of tracking and managing source code (Gaba, I. 2022). Furthermore, besides preventing possible loss of data, it also allows for mistakes to easily be rectified as former versions of a code can be restored with the click of a button. This makes it both useful and convenient for software developers to use, ourselves included.

To make and use Git-repository a host is needed (Gaba, I. 2022). In our case, we have chosen to use GitHub, because it is an affordable resource and features an open-source community (Gaba, I. 2022), which makes it convenient to facilitate multiple participants' contributions to a software. As a group, this is a valuable feature as it enables us to keep each other updated at all times, and search for inspiration from others.

Finally, an important reason we chose to use Git is that it is designed to be user-friendly (Gaba, I. 2022) and that we already had expertise with this tool. In conclusion, it allowed us to both synchronize our code to stay updated and track any changes without wasting time having to learn how to use the software, which is an advantage given our limited time-frame.

Kanban

Kanban is a project management method that uses a board to visualize the planning process and workflows (Sarandeska, I. 2019). We have chosen to use a digital Kanban tool found within GitHub. This allows us to stay updated, avoid work overlaps and manage our files on one platform. This lessens miscommunication, which can arise from complex communicative setups.

Kanban does not have specific rules to follow, but a set of guidelines and principles instead (Sarandeska, I. 2019). Not all these principles apply to our project. As such we have elected to mainly follow two principles whilst working with the Kanban method as we find these suitable for our project.

The first principle is to '*Respect the assigned roles and responsibilities for each task, and correlating caretakers*' (Sarandeska, I. 2019. p. 1), which means that one person is given authority over a task, and will remain in charge until feedback has been given. This avoids responsibility conflicts, thereby hopefully heightening the quality of work

The second principle we wish to follow is the principle of '*Pursuing incremental changes'* (Sarandeska, I. 2019. p. 1)**.** This ensures constant progress instead of stagnant periods, which can also heighten adaptability.

Finally, we have chosen to follow a number of practices in how we use our kanban board. This is to manage our workflow. The first practice is '*Visualizing the workflow'* (Sarandeska, I. 2019. p. 1) seen below to aid communication and understanding:

Idea Generator	Todo	In progress	Editing/acceptance	Done
- Brainstorming (All ideas generated by a member of the group can be written here)	- A list of elements deemed important to have in the project	- Task currently being worked on	- Solutions ready for feedback	- Collective list of completed tasks

We also '*Limit our work in progress'* (Sarandeska, I. 2019. p. 1)**.** Meaning we have a limit on how many tasks can be ongoing at a time. This upholds incremental changes instead of mountains of unfinished work. Another practice is to '*Manage our flow'*. This practice goes hand in hand with both the visualization and the last practice we follow, to '*Implement feedback loops'* (Sarandeska, I. 2019. p. 1). This implementation allows us to quality-check our work. In our project, we have specifically elected to do feedback in cohesion with all members present to make decisions in unison, and more diverse and thought out decisions.

We have in the group made two Kanban boards in parallel to avoid mixing up tasks or muddling up the Kanban board with fixing error tasks. An early snapshot of both kanban boards is shown below:

Figure 1 (Appendix 1):

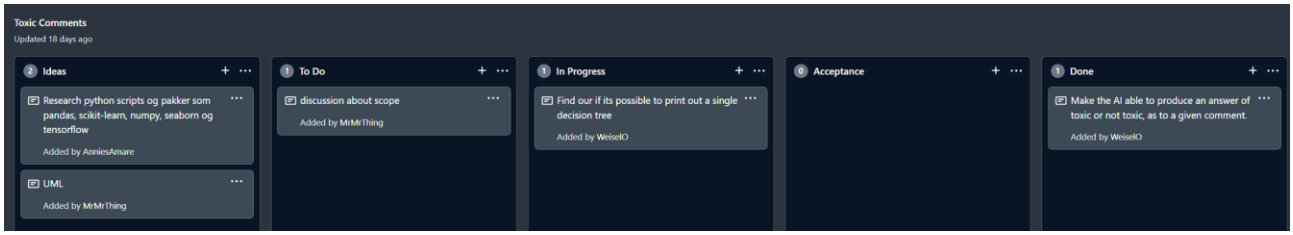
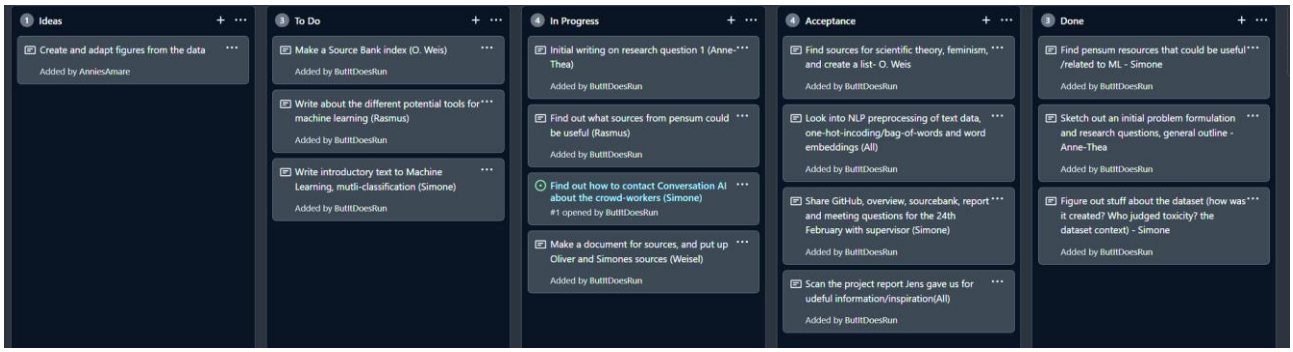


Figure 2 (Appendix 1):



TRIN-model

Here we will present and argue why we intend to use the TRIN-model (**T**echnology and **R**adical and **I**ncremental Design in **N**etworking) as a method to structure our examination for suggestions on how to create an AI that can detect toxicity.

The 'TRIN-model' is a model developed by Niels Jørgensen in collaboration with two other Roskilde University lecturers named Thomas Budde Christensen and Erling Jelsøe (Jørgensen, 2019).

The model is made up of six concepts: 1) Internal mechanisms and processes of technologies, 2) Artifacts of technologies, 3) Adverse effects of technologies, 4) Technological systems, 5) Models of technologies & 6) Technologies as innovation. However, we will only touch upon concepts 2 and 5 (Jørgensen, 2019) as these are the most relevant to our general goal of understanding our technological artifact and examining its use.

Each concept is presented as a 'step' in the model, and all steps contain an analysis of the concepts of the steps, and to a lesser extent analyses of theories. It does however not contain any kind of substantial method (Jørgensen, 2019). As such, we solely use the principles as a guideline on how to structure our research.

The step of 'Artifacts of technologies' is about the concept of an '*artifact*'. By Jørgensen's (2019) definition an '*artifact*' can be said to be an object, which has a technological function and is the outcome of a manmade process where the outcome has the purpose of fulfilling a human need. In Jørgensen's article, this step is used for examining the artifact of a technology (Jørgensen, 2019). We intend to use it for examining the specifics of our ML-based AI.

The step of 'Models of technologies' concern the concept of a '*model*'. In the TRIN-model Jørgensen (2019) classifies models as tools made to create or develop concrete artifacts. By his definition, models represent what enables an artifact to fulfill its purpose. Using this principle, we intend to examine different ML models, which can be used for toxicity detection. This is to evaluate the best or most appropriate toxicity detection approach for our problem using ML technology.

The purpose of this is to be able to have a concrete scope of technology to critically reflect and relate to the development and evaluation of an actual AI for toxicity detection. It also clarifies our intent and purpose behind the creation of our prototype.

In summary, we know from Jørgensen's (2019) article that the models' steps are a combination of technology definitions inspired by philosophers Carl Mitcham & Jens Müller. In other words, what aspects are important to cover in technical analysis. It can therefore be argued that by using the model's concepts to structure our technical analysis we not only aid in validating our choice of content in the analysis but also ensure that we have a shared knowledge base for understanding our technology.

Preprocessing

The first step in handling a new set of data is checking for missing or invalid data. These can come in a number of different types, but one of the most common ones is NaN which stands for 'Not a Number' and tells us there is no value in the given numerical field. The correct response to null values depends on the situation (Lawton, G 2022).

Secondly, we would look for useless or harmful data that is not needed in the data set for the particular use case.

When we are dealing with machine learning, we need to be able to work with any number of different input types, one could be letters and sentences. Standard machine learning is not capable of using these types without doing some preprocessing beforehand. Here we can use some methods that would transform the data into an input type that the classifier can use (Brownlee, J. 2017).

This is a transformation method that is used in NLP (Natural Language Processing) which takes a sentence and counts/vectorizes the words with a given template. This template is normally created beforehand because all the words used need a unique index. This approach is called 'bag of words'. One disadvantage of using 'bag of words', is that it does not contain any information about the structure of the original sentences (Devopedia. 2021).

N-gram is nearly identical to Bag of Words. The key difference is that N-gram can contain the structure of the sentences, by having a multidimensional array. Bag of words has a list over which words are in a sentence and how many, but N-gram can hold N amount of these indexes in a row and thereby keep the structure of that given sentence (Devopedia. 2021).

Three-Step Classification Model

Based on the principles of Data Science by Andreas C. Muller (2016), building a reliable machine learning model is a three-step process. This was further demonstrated by Hua Lu, a Professor of Computer Science from the research group PLIS (Programming, Logic and Intelligent Systems) situated in the Department of People and Technology, Roskilde University (RUC) during a lecture on classification, which our group attended the 10-03-2022 (Hua Lu. 2022).

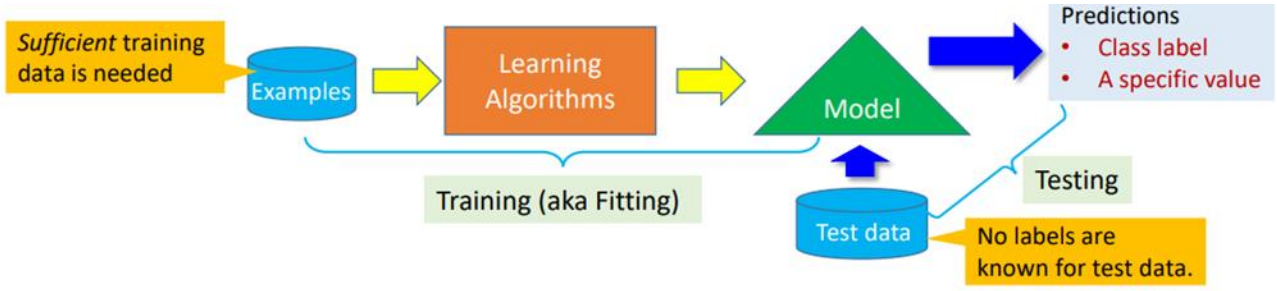
Hua Lu (2022) specifically applied the base data science principles from Muller (2016) to a classification example in the case of creating an ML model. Given Hua Lu's education and expertise, we will therefore argue that we can re-use his example as a valid method to accomplish classification.

In this report, we will refer to this method as the 'Three-Step Classification Model', and use it as a method to extrapolate one way to do a classification task.

The principles of creating a reliable ML are:

1. Model construction
2. Model validation
3. Model application/test

Hua Lu (2022) illustrates the processes of the three principles using this figure featuring classification as an example:



(Hua Lu. 2022)

The figure illustrates that classification consists of the three steps; model construction, validation, and application/test. For a classification task, the model is constructed by fitting an ML model using 'training data', which then is used to test and evaluate the model using performance metrics.

We intend to use this model to focus our technical analysis on ML technology because ML is a wide branch of AI. This means that if the analysis does not have a clear structure and a clear scope, we could end up including too much theory due to there being a multitude of ML approaches applicable for classification tasks. All are equally valid in their own way. The way we intend to do this is to apply our case scenario to each process of the model and argue for specific tools or algorithms to complete the processes. This will be done in two different ways to compare and argue either design.

Model construction

Model construction is the act of describing a set of predetermined classes. This is done using tuples, which as a sample make up a 'training set'. The tuples refer to labeled samples of the trained predictions (Muller, A.C & Guido, S. 2016).

The act of model construction is therefore called 'fitting' or 'training' because the training set is used to teach the model how to recognize the class that one wants to identify (Muller, A.C & Guido, S. 2016).

To do model construction classification algorithms are needed. The algorithms are chosen based on what type of classification problem is being dealt with (Muller, A.C & Guido, S. 2016).

For example, If the classification task is to label any given element as one of two classes, then it is 'binary classification' because it has a Bernoulli Distribution (a binary outcome) (Brownlee, J. April 8, 2020). In this case, algorithms appropriate for binary classification would be used.

For us, since we are dealing with the problem of whether or not something is toxic our classification problem is binary. Appropriate algorithms for us to do model construction could therefore include using either Logistic Regression, k-Nearest Neighbors, or Decision Trees.

Model validation

Model validation is the process of confirming whether the model achieves its intended purpose. This is done using a 'validation data set', which is a data set consisting of samples of the prediction task (Muller, A.C & Guido, S. 2016).

To validate the performance of a model 'performance metrics' is needed as a measurement, when running the validation/test data set through the ML algorithm. It is necessary to do this because assessing the efficiency of an algorithm is part of evaluating which algorithm is better suited for a specific application.

For this project, we have limited ourselves to using the performance metrics; confusion matrix and accuracy. The reason being these are popularly used performance metrics for classification, and we have experience with them. Something we determine will better aid our ability to evaluate our results.

Model Application/Test

Finally, we have the model application/test, which is testing the model in use. Here unlabeled data is given as input to test if the model can label elements correctly. The reason for using an unlabeled sample here is to avoid any false positives that could occur by only using the data set the model has been trained on to test it.

3. Scientific approach

3.1 Dimensions

This section includes a short explanation of how our bachelor dimensions are incorporated into this project.

Design and Construction

We include this dimension in our project by designing a toxicity detection AI during this project to test and supplement our research findings. We have organized a system development process for this technological artifact and used ML and AI theory and methods to analyze and evaluate our research results based on the artifacts' apparent risks, advantages, disadvantages, and functionalities. The intention is to use the practical knowledge gained from this to reflect on our conclusions.

Subjectivity, Technology, and Society

The focus of this project is the challenges presented by toxicity detection. These challenges include the relation between AI innovation and people as well as its effect on society. As such we have analyzed this technology's dynamic in cultural, political, and religious contexts incorporating Winner's perspective and the feminist epistemologies of science.

Technological systems and artifacts

This dimension is expressed through our use of data visualization and the technical analysis of our AI prototype. Our analysis involves the use of Fry's theory and the TRIN-model to better understand the purpose and creation of our AI prototype.

3.2 Feminist epistemologies of science

Introduction

This section will explain our scientific approach. We will elaborate further on why- and how this approach aids us in accomplishing our goals of identifying bias within the dataset and our AI prototype. Furthermore, this approach serves as a general basis for our understanding of bias and its inherently unavoidable nature.

Our scientific approach is guided by the field of feminist epistemologies of science. This branch of scientific theory is especially focused on understanding inherent biases in the creation of science. We chose this approach since Machine Learning and toxicity detection are fields prone to certain types of bias. Our goal is to educate ourselves on biases, including our own, to identify potential biases within our dataset and see if our program is susceptible to gender and ethnicity bias. We will also elaborate on feministic technoscience as it relates closely to our topic of research; toxicity detection AI.

Feministic technoscience studies arose in the field of biology and started as a critique of the status quo of scientific practices. It was meant to challenge the perceptions of women's role in science with a focus directed at biologist studies of sex and gender. A large quantity of their work has gone into researching the

scientific construction of gender and the cultural assumptions embedded in the language used. This is done to shine a light on a '*political unconscious*' (Sismondo, S. 2009. p 73).

To become more aware of the political unconsciousness, and possible bias in our research, we have elected to educate ourselves by looking into feministic epistemologies, and further into feministic technoscience, and diversity, as well as our group itself, being gender diverse, all in an attempt to better identify bias within our datasets and the AI we're training doing this project.

We also wish to reflect on the language we use to conduct science and create our AI. In the case of the feministic epistemologies the initial focus was gendered language (Martin, 1991. p 489). Similar to language, the technologies in use are also gendered, and are politicized. Due to the affordance of certain technologies furthering a specific use, and in a reserve correlation, making alternative uses more difficult. As an example, male-dominated work often relies on heavy tools, and the workers, therefore, rely on strength to do their jobs (Sismondo, S. 2009). In the same sense, an AI produced on datasets labeled solely by men would be prone to gender bias. Furthermore, an AI created to sort job applications might enforce a gender bias in especially male-dominated fields. In feminist epistemologies, these issues are addressed through reflectivity and awareness.

In our project, we aspire to have a feministic epistemological point of view. Our project is thematically centered around toxicity and people who take offense to comments on the internet. Because of this, we need an awareness that people might have different opinions as to what they react negatively towards. We have found that language for different ethnicities is a difficult problem to solve. The language we use is politicized. Not only in gender, but also in ethnicity and religion. A word used between members of the same group can have another meaning if the context or communicators change.

An example could be religious language such as our example with the word 'jihad'. A person within the religion of Islam might use 'jihad' in a positive context spreading positivity inside a community they might be a part of. But used in the context of Islamist extremism this might be used in a toxic manner, and might even be used as a threat.

> "*Ambiguity is a significant challenge. For instance, the meaning of the keyword ''jihad" in religion is referred to as a self-spiritual struggle, while it indicates intent to harm other individuals in the Islamist extremist ideology*"(Sheth et al., 2021. p 9. Line 17-20).

This is an example that could potentially be difficult for an AI to handle as it can hold more than one meaning, and this meaning is determined by a context our AI does not necessarily take into account.

A pragmatic approach to bias

In this project we are looking into AI and its ability to detect toxic comments arising in the Wikipedia comment section. Furthermore, we are looking into possible biases. To do this we are looking into AI, which can find some of these comments, and into feministic epistemologies and feministic technoscience, which can help us identify possible biases in an AI or in conjunction with the dataset the AI is trained on.

We intend to use machine learning to accomplish our goals of identifying these toxic comments but have run into difficulties as we try to teach this AI to classify words as toxic or non-toxic. One thing especially hard for us to detect is ambiguous words, like 'jihad'. This leads us into an ethical question of Kantianism and

further, intention. Our AI has little chance of knowing what intentions a commenter had while writing a comment as it can only look at the written words.

To see how potentially biased our dataset is, we've decided to conduct a small practical test of the dataset. In this test, all four of the group's members will individually look at a small sample of the dataset and determine for ourselves if the comment is or isn't toxic. By equipping us with the scientific approach of feministic epistemology, and having a diverse cast of decision-makers, in this case, two men and two women, we hope to have a higher accuracy at deciding what is and isn't toxic. Teaching the AI better and making a more just program. After the test is conducted we hope to find out if there is a bias present among the group members, and also see how we report the comments compared to how it is labeled in the dataset prior to the tests.

The initial talks within our group highlighted a problem in need of addressing: A risk of training an AI to be unjust or unknowingly giving such a machine a bias towards certain groups of people, or standpoints.

This is a concern mirrored in Sergio Sismondo's (2009) work; 'An Introduction to Science and Technology Studies' where he mentions different people not being included in science, and this paragraph specifically in the technology of science;

> *"discrimination is clearly unjust, and is inefficient because it reduces the pool of potential contributors to science and technology. [...] We might also ask in what ways science and technology would be different if Western minorities, and non-Westerners, were better represented"* (Sismondo, 2009. p 72. Line 12-18).

Sismondo's (2009) reflection on how a more diverse group of people could show better results in technologies produced is a factor that inspired us to engage in researching diversity. It directly inspired us to test the datasets we're working with within our group. Furthermore, we intend to use the knowledge gained from researching feministic epistemologies to create an awareness of potential biases within ourselves in order to reflect on them and produce the best possible work.

4. What defines content as “toxic”?

In this section, we will examine the meaning of the word “toxic” when used in reference to social behavior on the internet. For the purpose of this analysis, we will consider the creation of all types of toxic content as a type of toxic behavior. We will explain our reasons behind using the word “toxic” in this report and clarify the relevance of defining toxicity in this context.

The word “toxic” is a term commonly applied to describe poisonous pollution and waste products in the natural sciences. Formal use of the word regarding behavior is for the most part only prominent in corporate research on leadership and collaboration. However, when used in a casual manner the word can often be applied broadly to behavior that is perceived as negative or damaging. The Cambridge Dictionary (n.d.) has defined this informal use of “toxic” as:

1. very unpleasant or unacceptable
2. causing you a lot of harm and unhappiness over a long period of time

The first definition is applied broadly. But the second definition specifically refers to relationships. Cambridges’ SMART Vocabulary relates this to the topic of: “*Making people sad, shocked and upset*”. This definition is a good example of the word “toxic” being used informally to indicate bad or harmful behavior towards others (Cambridge Dictionary, n.d.).

Studies such as “*defining and detecting toxicity on social media*” (Sheth et al., 2021) and studies of “*toxic behaviors in esports games*” (Adinolf & Turkay, 2018) use different vocabulary to indicate harmful or aggressive behaviors. The reason is that formal definitions of toxic behavior are somewhat rare. Using specific definitions can prove useful as it better describes what specific behavior is being researched. However, it is also often a far more narrow definition than that of generally toxic behavior.

In this project, we will attempt to create a definition of toxic behavior as it is defined in an internet context. There are two main reasons for this decision.

The first reason is our data. Throughout this project, we will be working with a dataset that applies the adjectives ‘toxic’ and ‘severe toxic’. To analyze this data, we need a shared understanding of what constitutes toxic behavior and how it is applied within this specific context. Exploring a more general understanding of toxicity is necessary to qualify whether or not the definition applied within the data is an accurate definition of toxicity.

The second reason is the broad context of our project. We are working towards the goal of understanding the challenges in creating AI detection of toxicity in comments and online forums. Confirming the success of the AI requires an understanding of toxic behavior as it is perceived in an internet context. Simply choosing a synonymous expression to use would only be a reflection of our personal perceptions of toxicity. This perception would also likely be biased, as it is in our interest to choose an expression of toxicity that is easily identifiable by the AI. To avoid this bias we need to more broadly examine how “toxic” behavior is defined.

4.1 Toxicity in an internet context

In this section we compare studies of different aspects of toxic behavior on the internet. This is done to examine similarities and differences between studies handling the topic of toxic behavior. Our goal is to gain

a more general understanding of toxicity by examining the key aspects of how toxicity is defined within different sources.

How social networking sites such as Facebook, Snapchat, Twitter, etc. affect our social lives has been the subject of much research. Among this research is the study of cyberbullying and online harassment on social media. The research article: "*Defining and detecting toxicity on social media*" (Sheth et al., 2021) reflects on the problems involved with defining and understanding toxic behaviors on social media platforms.

In this research article toxic content is used very broadly to describe; "*[...] harmful content including disinformation, conspiracies, extremism, harassment, violence, and other forms of socially toxic material*." (Sheth et al., 2021, p. 1). But while this broad understanding of toxic content is applied, it is made clear that the definition of toxic content is largely based on intent and context.

What is perceived as toxic behavior depends on the community, the social exchange, and the prevalent culture. Friends may exchange otherwise toxic remarks in a joking or teasing manner. Members of the same racial background might also exchange comments that would otherwise be viewed as racist or offensive. In every case, toxicity is dependent on the context (Sheth et al., 2021).

The intent is also important when interpreting toxicity. A term such as "*jihad*" a "*self-spiritual struggle*" might be considered toxic as it is also used prominently by Islamic extremists with the intent to harm. Likewise, bullying is primarily defined by an intent to hurt the victim. The intent is hard to prove, even when given a fair amount of context. We are also faced with the problem that "*experienced harm is distinct from intent to harm*" (Sheth et al., 2021, p. 2). This means that an intent to harm must be deemed plausible from the surrounding social and cultural context, rather than from the victim.

The article argues that this understanding of toxicity poses a challenge when it comes to detecting toxicity using an AI. This is because it is impossible to use a simple keyword- or sentiment analysis to identify toxic content without disregarding the intent and context. Furthermore, toxicity is multimodal. Images, videos, and even sounds can be used in toxic content, all of which are not reasonably detectable by most modern AI technologies (Sheth et al., 2021).

It is even argued that there is often bias in the gathering of written texts to use as training materials for an AI. Identifying toxic content is dependent on perceived context and intent. This means that any training set will be biased by the cultural and ethnic identity of those gathering and labeling the data (Sheth et al., 2021). The article summarizes:

> "*To identify toxicity, it is necessary to understand the broader context beyond the situation and domain-specific content analysis, with reference to applicable human values, social norms, and culture, at the individual, group, and community levels.*" (Sheth et al., 2021, p. 3, line 44)

In conclusion, it is impossible to precisely define toxicity in a universally meaningful way. Instead, toxicity must be defined within specific domains, applying an understanding of the social and cultural norms prevalent within this domain. This means that the definition of toxic content can vary from community to community.

Another study by Kordyaka et al., (2020) attempts to gather a general understanding of what constitutes toxic behavior in gaming communities. They did this through a questionnaire study of 320 respondents, most of which were young adults from either America or India. The survey was answered by 214 male participants and 105 female participants, all of which were players of either DOTA 2 or LoL (League of Legends).

In this study they; "[...] *understand toxic behavior as an umbrella term used to describe different negative behaviors (e.g. harassment, flaming, trolling, and cheating) corroding team effort and harming the game ambiance while playing multiplayer video games*." (Kordyaka et al., 2020. p. 3, lin. 1). This definition indicates that, in the case of gaming, toxicity is usually defined as being detrimental to teamwork and general enjoyment. Interestingly this definition of toxicity disregards intent and rather focuses on the effects of certain behaviors.

The study also highlights toxic behavior as having negative consequences, meaning that; "*players who experience aggressive activities may choose to leave the game or initiate more toxicity in return, which may lead to a downward spiral*." (Kordyaka et al., 2020. p 14, lin. 52). This means that part of the indication of something being toxic is the way people react to it. Retaliating or leaving the game are both reactions to perceived toxicity within the game. Conclusively this study generally uses the reaction to the content to indicate whether or not it is toxic.

This also relates to their conclusion that people with prior experience with toxicity have a higher tolerance for toxic behavior (Kordyaka et al., 2020). Content that is seen as extremely toxic by a newcomer is perceived differently by people who have been a part of the community for a longer period of time, as they have experienced more toxic behavior. This indicates that reaction is again identified as a relevant indication of toxicity.

Overall both studies identify toxicity as behaviors that are harmful to the surrounding community. Both studies also argue that context is relevant as to whether or not something is experienced as being toxic. The research article by Sheth et al. (2021) presents intent as a crucial part of toxicity - While the study conducted by Kordyaka et al. (2020), presents the general reaction to content as an indication of toxicity. This means there is a clear distinction between toxicity based on intent and toxicity based on reaction.

While defining toxicity based on reaction makes it easier to detect, it encourages the ethical discussion of whether intent or reaction is more important. We will elaborate further on this topic in our final discussion.

4.2 A simplified definition of toxicity

The study by Kordyaka et al. (2020) is not the only study that applies a reaction-based understanding of toxicity. As highlighted by Sheth et al. (2021) in their research article there is a general tendency toward simplifying the otherwise nuanced concept of toxicity. This trend is especially prevalent within ML approaches (Sheth et al., 2021).

The reason for this is that the challenge of data gathering and the issue of defining toxicity are closely entangled. Gathering labeled data for ML is in itself a huge task, as exemplified by the project by Wulczyn et al. (2017). This project has the goal of building an ML model in order to label massive amounts of data as either toxic or non-toxic. The purpose of the project is to address key issues with data gathering. Crowdsourcing is both time-consuming and expensive. Furthermore, annotators have to be instructed and possible conflicts between their assessments have to be addressed (Wulczyn et al., 2017).

Creating the instructions for labeling toxic content also requires a clear definition of what constitutes toxic content. However, defining toxicity also requires a lot of data. This is the general cause of the paradoxical nature of the problem: Gathering data on toxic content requires a definition of toxicity and defining toxicity requires a lot of data.

One approach to this problem is the one applied by Wulczyn et al. (2017) in their project. They choose to apply a simplified definition of toxicity for pragmatic reasons. Specifically, they focused on personal attacks and harassment, posing the question: "*Does the comment contain a personal attack or harassment?*" (Wulczyn et al., 2017, p. 2) and having the annotations mark it as "*not an attack*" or indicate the target(s) of the attack (Wulczyn et al., 2017).

Similarly the dataset for the ML-classification challenge: "*Jigsaw Unintended Bias in Toxicity Classification*" (Conversation AI. 2019) asked annotators to; "*Rate the toxicity of this comment*" giving the option of labeling it as: "*Very Toxic*", "*Toxic*", "*Hard to Say*" or "*Not Toxic*". Further specifying that toxicity and the severity of it are based on the likelihood of it making you (the annotator) "*leave a discussion or give up on sharing your perspective*" (Conversation AI. 2019).

Furthermore, the comments in both datasets were evaluated by up to 10 annotators and finally labeled based on majority rule. While the data from the personal attack-study does indicate that information on the crowdworkers was gathered, the information is not applied directly within the study (Wulczyn et al., 2017). In other words, annotator disagreements are not closely examined and there is no discussion of potential cultural or personal bias applied by the crowdworkers. This means that a lot of information with potential relevance to the validity of their applied definitions of toxicity, including harassment and personal attacks, is simply disregarded.

Conclusively a simplified approach to defining toxicity can be necessary for pragmatic reasons, such as making it possible to gather data and inform annotaters. But the validity of the simplified toxicity-definition must also be part of the general evaluation of the dataset. Using these datasets means that we are not detecting toxicity in a broad sense, but rather detecting a specific interpretation of toxicity on a large scale.

Summary

While we can conclude that there is no universally formal definition of toxicity, there are certain defining traits applied to toxicity across different sources. Toxicity is often defined by a broad scope of harmful or socially disruptive behaviors, including but not limited to: harassment, cheating, mocking, threats, as well as obscenity, and general hostility towards others. The challenge of this broad definition can be addressed by defining toxicity within its specific context and applying knowledge of the surrounding context, social norms, and community.

Furthermore, it is an ongoing discussion whether toxicity should be defined by intent or reaction. While defining toxicity based on reaction has a practical advantage, especially within the context of the ML project, it can be argued that disregarding intent poses an ethical issue.

Finally defining toxicity within ML-projects has the inherent problem of getting enough data. While data is necessary to define toxicity, it is also necessary to define toxicity to some degree in order to collect relevant data. This paradox indicates that it is to some degree required for ML-projects to apply a simplified definition of toxicity to collect data.

5. How is toxicity defined within our dataset?

In this section, we explore our selected dataset. We explore how toxicity is defined as compared to a more general understanding. Furthermore, we explore the dataset using data visualization to present statistics and analysis of certain data patterns. We explore the implications of these patterns when applied to an ML approach. Finally, we provide a small sample as an example of a method for exploring potential data bias.

5.1 Understanding our dataset

Our dataset is from a Kaggle challenge named "*Toxic Comment Classification Challenge*", which is about identifying and classifying toxic comments online. The challenge was issued in 2018 by Lucas Dixon, a computer scientist in Google Research and the former Chief Scientist at Jigsaw. Jigsaw LLC, formerly known as Google Ideas, is an organization that once existed only as Google's think tank. Today the company uses technology to address a range of geopolitical issues on behalf of Google (Chang, L. 2016).

Lucas Dixon is known for working with ML, data visualization, and conversation AI (Google Research, n.d.). Conversation AI is an initiative to "*protect voices in conversation*" in order to protect the opinions of minorities and vulnerable groups when engaging in online discussions. The core values of the project include: Community, transparency, inclusivity, privacy, and topic-neutrality. All relate to the goal of making the internet a better platform for fair, open, and honest discussions rather than hostility and toxicity (Conversation AI. n.d.).

The initiative is about using ML to develop tools for combating online toxicity and harassment, and it is a joint research effort led by Jigsaw and the Google Counter-Abuse Technology Team (Conversation AI. n.d.). The "Toxic Comment Classification Challenge" is a part of the initiative and has had the general goal of exploring ways of using ML to prevent toxic behavior from hindering a meaningful discussion.

In the Kaggle challenge introduction, Dixon states:

> "*If conversations are so bad that people leave the discussion, then we have clearly failed to have a online discussion, let alone a good one! This was the basis for working with Wikimedia to create a dataset of comments from Wikipedia Talk pages [...]*" (Dixon, L. 2018. Page 1, lines 5-9).

He goes on to explain that the Wikipedia comments were then crowd evaluated for toxicity. In the crowd-evaluation, the term toxicity is interpreted as any content that hinders fair discussion, either by being rude, disrespectful, or likely to make people leave. This is similar to the sentiment behind the "Jigsaw Unintended Bias in Toxicity Classification"-challenge (Conversation AI. 2019), where toxicity is also evaluated based on the likelihood of the recipient leaving the conversation. We assume that our data was labeled in a similar manner, using a similar question. Furthermore, the comments in our dataset were evaluated for the specific type of toxicity present in the comment (obscenity, threats, insults, and identity hate).

The crowd-workers were presumably instructed to judge toxicity based on these criteria and furthermore asked to categorize content within the specific categories. Rather than interpreting toxicity individually, the subjects are told to apply a predefined understanding of toxicity. Furthermore, a category-limit is imposed on how toxic content is subdivided. This means that our data is not the result of a general interpretation of whether or not something is toxic. Instead, the results reflect what is interpreted as detrimental to the intention of creating a respectful discussion.

This is an indication that we are dealing with a reaction-based understanding of toxicity, and furthermore applying a simplified understanding of toxic content. This data cannot be applied to detect toxicity in general. But it is applicable for detecting the types of toxicity, that are likely to make people abandon interaction. Furthermore, the data is focused on the effects and reactions to toxic behavior, rather than intent. This is justified by the goal of creating an open discussion, meaning that intent is considered irrelevant if your perspective or opinion serves to discourage others from participating. While it is advantageous for pragmatic reasons to focus wholly on reaction, the results may cause unintentionally toxic content to become flagged, which could potentially prevent people from expressing themselves. This issue will be further discussed in our final research question.

Overall there is a lack of information about this data and its creation. This has led to a fair amount of uncertainty concerning its validity. It is not clear how many participants there were in creating this dataset. This leads to a problem of generalization. While we are building on the assumption that the results can be generalized, since crowdsourcing usually has a great number of participants, this cannot be confirmed.

Furthermore, there is no information about the participants, whether they are of varying nationality, cultures, ages, or affected by anything that might create a biased perspective. Because the data is created through crowdsourcing, we can assume that any number of internet users from anywhere around the world could have participated. However, it is not something we can say with absolute certainty.

Lastly, it is unclear what the exact instructions used in the examination were, only a general description made by Lucas Dixon. As mentioned above this description includes the notion that toxicity is any type of offensive content that might discourage people from participating in a discussion. But it is unclear how this was presented within the actual instructions.

This lack of information generally meant that we could not validate the dataset based on its description or its participants. Our options were either to find a new dataset or to find another way to validate the results of our current dataset. We chose the latter option based on the following reasons: The first reason is that the dataset is fairly large, providing a rich scope for potential ML and testing. The second reason is that the dataset is created with the intent of being used to research toxicity detection. This means the data is targeted toward the exact subject that we are examining. This minimizes the amount of work necessary to isolate relevant data.

It is also worth noting that creating or obtaining these kinds of data is both difficult, time consuming, and expensive. Because of this, it is beyond the scope of this bachelor project to obtain data that is specifically tailored to the focus of our research. Moreover, we would argue that being aware of flaws and biases contained within our dataset serves to validate our research, even if the data we apply is to some extent distorted.

In the end, we concluded that using the dataset, despite its inherent issues, was most likely the best possible option. However, this decision left us with the problem of having to validate the data, using only the data itself.

5.2 Validating our data

In this section we make a mathematical and statistical analysis of the dataset. This is done in order to understand and validate the interpretation of toxicity that is present within the data. Because there is so little information about how this dataset was created, we need to examine the data itself to better understand what it tells us. In order to do this, we have applied Fry’s method of data visualization. Specifically, we have

applied the steps “Parse”, “Mine” and “Represent” to our data, in order to get a clearer understanding of the data.

The parsing process is made simple by the fact that our data is already subdivided into categories. The degree of toxicity can be divided into: non-toxic, toxic and severe toxic. While the content sub-categories remain as: obscene, threat, insult, and identity hate. This results in two areas of assessment: Toxicity and aggression-type.

As is it also possible for a comment to not be labeled with any sort of offensive content we will include the ‘clean’ sub-category. This is not an indication of toxicity, but rather an indication that the comment is not labeled as obscene, threatening, insulting, or containing identity hate. This is due to the data containing comments that are labeled as toxic, while not being sub-categorized. This means that for a comment to be deemed “Pure” it has to be both non-toxic and clean.

Figure 3 provides a simple overview of these categories (Appendix 1)

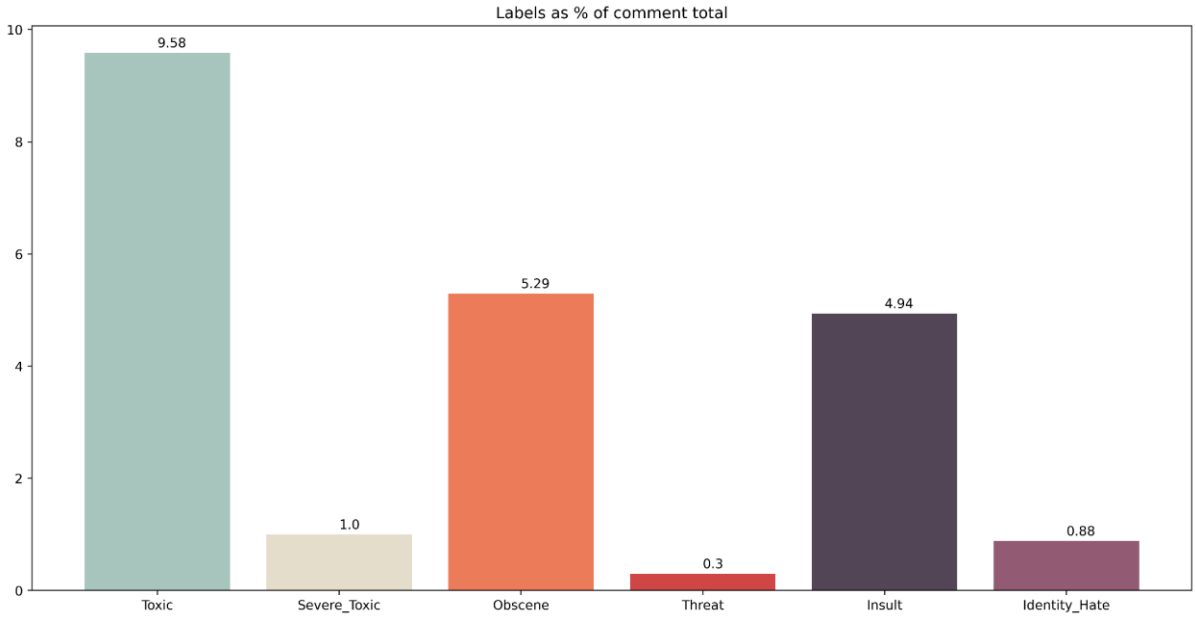
Aggression-type / Toxicity	Obscene	Threat	Insult	Identity hate	Clean
Toxic					
Severe toxic					
Non-toxic					Pure

In this figure the assessment-categories are shown. Anything that is labeled within the red area is considered toxic. The yellow area is for comments that are considered non-toxic, meaning that they wouldn’t be a cause for leaving a discussion, despite containing offensive content. Anything within the green area is labeled as a pure comment, meaning it is not toxic and not offensive in any way. These are the type of comments that Dixon would deem ideal as they encourage thoughtful and polite discussion.

Now that we have a way to structure our data using these categories we move on to the “Mine” and “Represent” phases. We approach these by first discerning general patterns using statistics and then presenting these patterns using simple visual models. Once this is done we apply supervised machine learning using classification to examine how the data is interconnected and represent this using more advanced visual models.

There is a total number of 159.571 labeled comments in the dataset. Out of all these comments, 143.346 are categorized as pure and 15.294 are labeled as toxic. The following figures show the distribution of labels. The result is shown as a percentage of the total number of comments. Note that a comment can be labeled with more than one category at a time.

Figure 4 (Appendix 1)



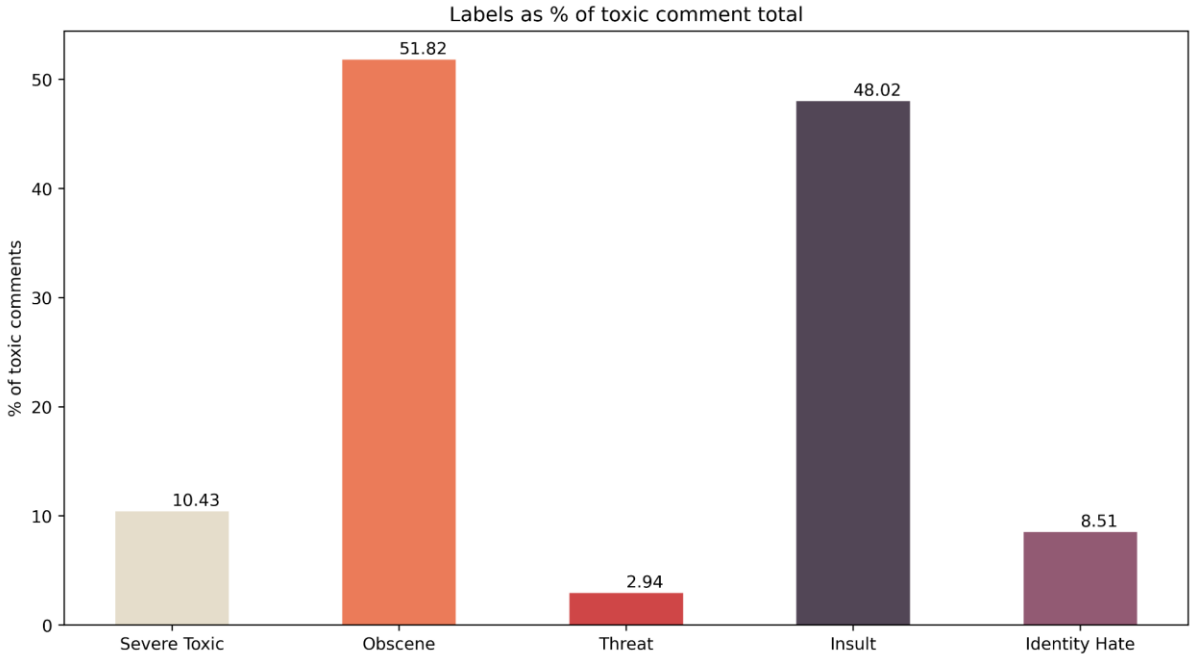
Right away we can tell that out of all the comments around 10% are labeled as toxic and only 1% are considered severe toxic. We can also tell that the most prominent aggression-types are insults and obscene content, with around 5% each. Threats are the least prominent aggression-type, with only 0.3% of the comment total being labeled as threats.

Research on the distribution of personal attacks within Wikipedia Talk Pages indicates that the number of personally attacking comments is only around 1 % (Wulczyn et al., 2017). This could indicate that applying a broad definition of toxicity has resulted in more content being labeled as toxic.

Overall the dataset contains the most data for pure content. This means that there is more information about what isn't toxic, rather than what is actually toxic. But this is not necessarily a bad thing. Any AI trained on this dataset will inevitably have more training on what isn't toxic, rather than what is. While this might increase the risk of false negatives (toxic comments labeled as non-toxic), it also decreases the risk of false positives (non-toxic comments labeled as toxic). Knowing this we have to be especially aware of the risk of false negatives. We will also discuss the potential consequences of false negatives vs. false positives in our discussion later on.

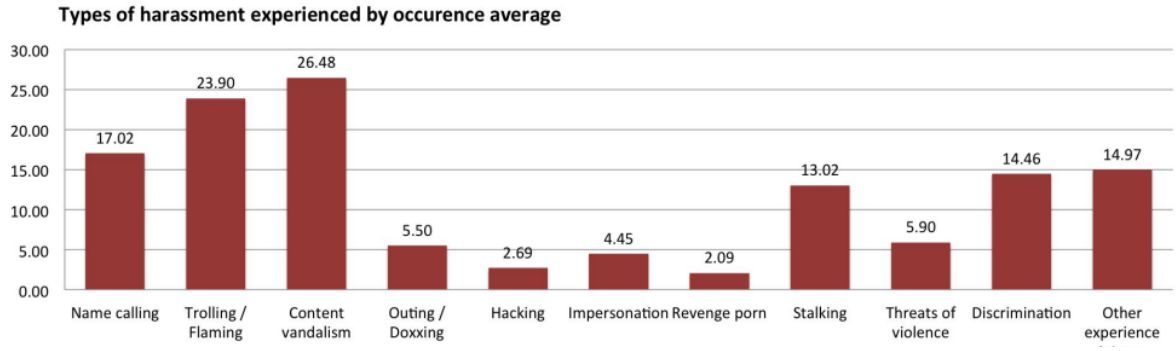
Aside from deriving knowledge of what is and isn't toxic, the data presents a certain correlation between toxicity and the aggression-types. This data is represented using a bar-diagram. The bar represents the percentage of toxic comments that are labeled with a specific aggression-type.

Figure 5 (Appendix 1)



We can see that the most prominent types of toxicity are insults at ca. 48% and obscene content at ca. 52%. The least prominent type of toxicity is threats at ca. 3%.

We can compare these results to the results presented in the 2015 Wikimedia Harassment Survey-report (Tsouroupidou, 2016).



(Tsouroupidou, 2016, p. 18, figure 16)

The Wikimedia Survey (Tsouroupidou, 2016) uses different classifications of harassment from the dataset we are using. But there are some similarities. For example, we can tell that our dataset shows a smaller amount of threatening content than the survey, where threats amount to around 6 % of the harassment experienced by their users. This could have potential consequences for our ability to detect the true amount of threatening comments, assuming that the Wikimedia Survey is more accurate than our data.

Toxicity in our dataset is primarily defined by obscenity and insults. This correlates somewhat with the Wikimedia Survey, where “Name calling” at 17% and “Trolling / Flaming” at 23% are prominent types of harassment, both of which can include insults and profanity (Tsouroupidou, 2016). But it is also an indication that our data contains a very broad interpretation of toxicity. Anything coarse or insulting can

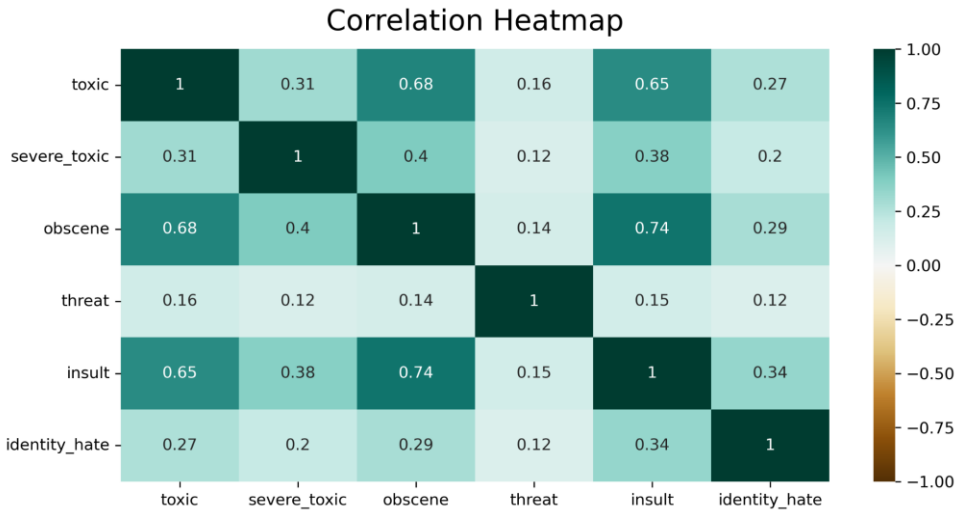
potentially be toxic. This is a problem of applying very general labels rather than categorizing the data based on actions, such as “Name calling”. Any model we train on this data will adopt this very basic understanding of toxicity and will therefore be more likely to label content as profane or even slightly insulting as toxic.

It is also worth noting that 15% of harassment in the survey is some “Other experience”. This could indicate that the survey-options are insufficient for the participants understanding of harassment.

The dataset we are working with has no option to select an aggression-type that is not already indicated. This means that the dataset somewhat blindly assumes that toxicity always falls within the category of “Severe toxic”, “Obscene” etc. But judging by the Wikimedia Survey (2015), this might not always be the case. “Content Vandalism”, “Discrimination” and “Stalking” are all prominent categories that don't necessarily fit any of the agression-types provided within our dataset. These types of toxic behavior would either show up as content labeled both “Clean” and “Toxic”, meaning there is no identifiable aggression-type - or they would be labeled with an ill-fitting aggression-type. Both of these options make the data harder to validate, as it doesn’t provide annotators the visible option of indicating that the inherent understanding of toxicity and its sub-categorization is insufficient.

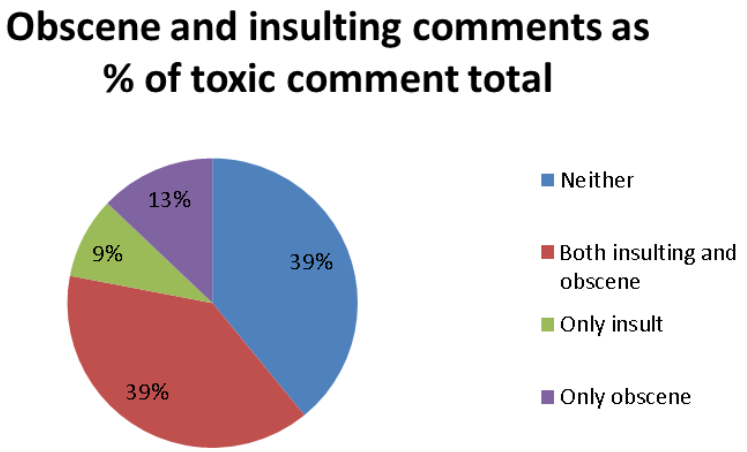
Finally, we decided to make a correlation heatmap for the entire dataset. This was primarily to further examine the correlation between toxicity and the different aggression-types. But it is also to examine if the aggression-types are in any way interconnected. The heatmap scales from 1 to -1 and measures linear correlation. With 1 indicating a direct proportionality and -1 indicating that they are inversely proportional. A value closer to 0 indicates that there is no proportional relationship between the two values.

Figure 6 (Appendix 1)



The heatmap indicated that there is a strong correlation between toxicity and obscene content and also a strong correlation between toxicity and insults, confirming our previous results. Interestingly there is an even clearer correlation between insults and obscene content. This could indicate that these two categories are related. Analyzing the numbers in the database indicates that 5954 toxic comments are also labeled as both obscene and insulting. This means that ca. 39% of all toxic comments are labeled as both obscene and toxic. We chose to illustrate the related data in a pie chart.

Diagram 1 (Appendix 1)



The chart shows that any toxic comment is more likely to be labeled both insulting and obscene than only labeled as either an insult or obscenity. This indicates that any comments within our dataset that contain insults are also very likely to also contain obscene content (and vice versa). This means that any model trained on this dataset will lack data to distinguish insults from obscenity. Furthermore, the correlation between these two aggression types and toxicity again indicates that our model will be very likely to interpret any insulting or obscene content as toxic.

This is not necessarily a problem, as our goal is to detect toxicity in general, not to detect its specific subcategorization. However, it does indicate that the data we are applying is somewhat biased towards specific types of toxicity. Making it better at detecting for example insults/obscenity and less proficient at detecting other types of toxicity such as threats.

5.3 Sampling our data

The following sample is intended as an example of how data-bias could potentially be addressed in future studies. We have chosen to include this sample despite validation issues, as the sample illustrates how data can potentially be assessed in accordance with our scientific approach - which dictates that certain norms and biases can affect our scientific study and practice. This section will elaborate on the creation of the sample, explore some of the related results and elaborate on how this could potentially serve as a basis for further research.

Initially, we wanted to gather a sample through a survey, which would allow a diverse group of people to judge a sample of our dataset for toxicity and thereby allow us to discover potential differences between groups. However, due to the general scope of our project, we decided that we did not have the time nor resources to create and process a large survey. Instead, we decided to do the survey among ourselves to examine our own biases and potential differences between gender.

While this is by no means ideal practice, it still serves as a proof of concept. It conveys the *potential effects* of looking into the subject of cultural and gender-specific interpretations of toxicity.

The sample itself is created with SQL through a few queries of the original dataset. Initially, we experimented with completely random samples of 100 comments. But this posed the problem of getting

enough toxic comments within the sample. In the end, we decided to compose a sample from 100 random comments, 50 that were originally deemed toxic and 50 that were originally deemed non-toxic. Thus we achieve a total of 100 comments with a 50-50 distribution of toxicity, despite the comments themselves being selected at random.

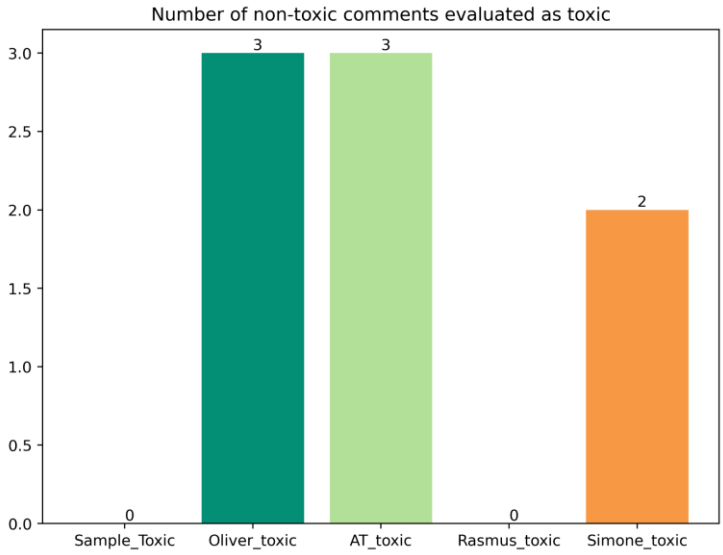
All comments were evaluated without any knowledge of their context. Since the comments were selected at random and the dataset itself provided no context references, it was not possible for our sample to include an evaluation of the context. This could be the reason for the noted discrepancies between our evaluation and the original one, as it is unclear from our sources whether or not the comments were originally evaluated within context.

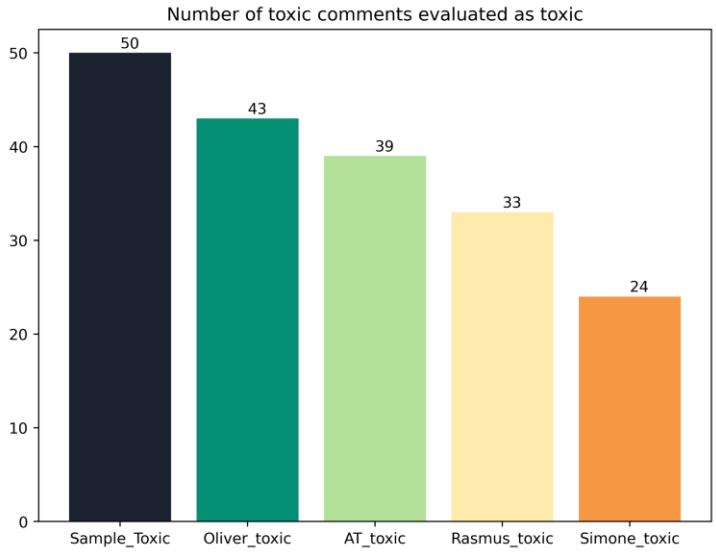
However, we did note some interesting tendencies from our sample. We present these data as proof of concept. Note that this data is *not* included for scientific study, but rather to argue that the general subject of data-bias deserves even further examination.

Figure 7 shows the distribution of comments evaluated as toxic between the non-toxic data and the toxic-data. On the left (figure 8) is the number of non-toxic comments that were evaluated as toxic. On the right is the number of toxic comments that were also evaluated as toxic. The labels indicate who provided the sample. In future studies, we would recommend making participants anonymous instead of providing a binary indication of gender. This is to avoid participants being judged personally based on the outcome.

However as we are attempting to reveal our own potential biases, we did not feel the need to make the samples anonymous.

Figure 7 & Figure 8 (Appendix 1)





We see the trend that our evaluation of the sample is generally less toxic than the original evaluation. This could be due to cultural or personal differences as opposed to the people who made the original evaluation. But without data on the original survey participants, we have no way of asserting that assumption.

However, the data does show that only a small number of non-toxic comments were re-evaluated as toxic. Most notably there was majority agreement, 3 out of 4, on one comment:

> "*All i can remember, in fact, is schnarquing your mom last night.*" (id: 7d20be69f46253ac, Appendix 2, test_labels file)

This comment was re-evaluated as being toxic by most parties, due to obscenity and insulting content. Furthermore, some originally toxic comments were re-evaluated as being non-toxic. One such comment is the following:

> "*Your edits to 2100. Don't be stupid. 2100 can also refer to a number, hence we shouldn't pretend that it doesn't, by redirecting in to 21st century*." (id: 865b4efa8a37c739, Appendix 2, test_labels file).

While calling someone stupid is considered an insult, none of us deemed this comment as being toxic, despite its original classification as such. Generally, these two comments could indicate that the general evaluation of the dataset is somewhat different from what we might personally consider as toxic.

Furthermore, the sample also shows the unanimous evaluation that the following toxic comment is actually non-toxic:

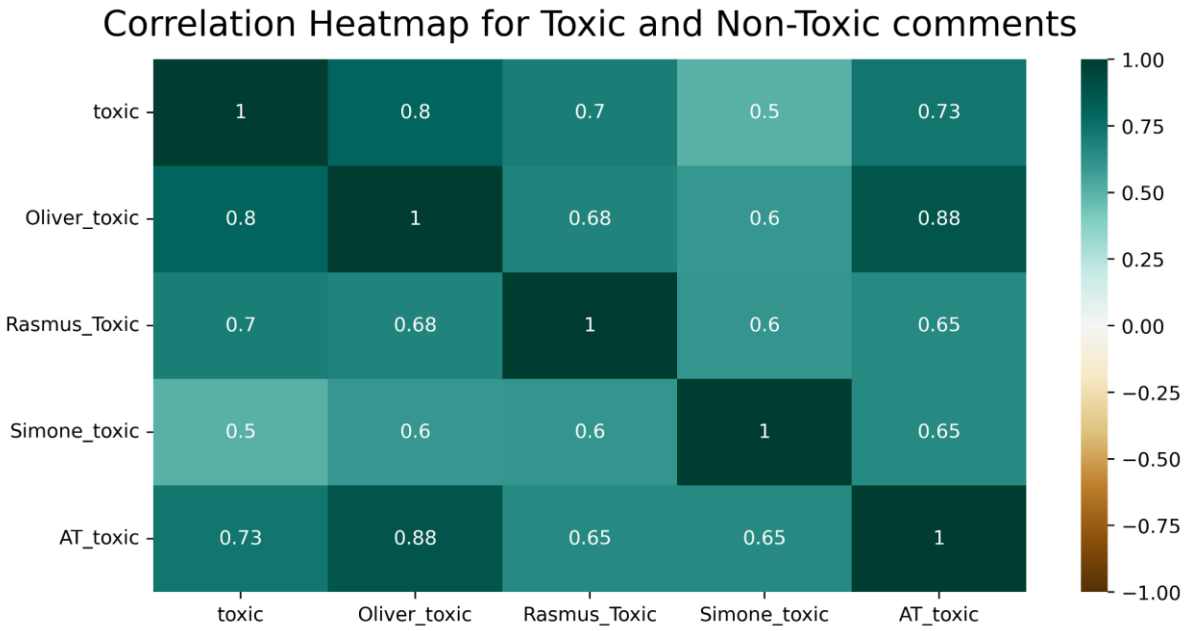
> *"Homosexuality and the punishment being equal to adultery. Are we sure that al-Qaradawi said that people who commit homosexual acts should be stoned, as the Wikipedia article currently asserts? It all depends on the Arabic words he used. If he said the punishment should be the same as those who commit zinnah, then it's 100 lashes for those who are unmarried and stoning to death for those who are married. That's not the same as saying ALL homosexuals should be stoned. [...]*" (id: f744c62f0e46f093. Appendix 2, test_labels file).

While the subject of homosexuality and the punishment thereof can be considered taboo or at least somewhat offensive, we evaluated the content of this comment as being non-toxic. This could potentially be an indication that the dataset itself contains a bias towards mentions of homosexuality. However, it would require more samples and a thorough data examination to truly assert this statement.

The discrepancies between our sample evaluation and the original evaluation may serve as proof of the problems of defining toxicity and creating a dataset based on this definition. As no dataset is without bias, it is also impossible to create a model without a certain bias. However, we would argue that awareness of definition-problems and potential data-bias is essential to creating an acceptable model.

It is also worth noting that while there are some discrepancies there is a strong overall correlation between the original sample evaluation and our evaluation, as illustrated in figure 9.

Figure 9 (Appendix 1)



As we can tell by the correlation heatmap there is no visible correlation between the toxicity evaluation and gender. The correlation between male participants and female participants matches the overall average for correlation between participants. However, this could also be due to the rather small scale of this sample.

Summary

There is generally a lack of knowledge on the specifics of our dataset. But based on information on similar datasets and the data itself it becomes possible to analyze the datasets' inherent definition of toxicity. Toxicity within our dataset is broadly defined as content that is detrimental to fair discussions, such as comments that are suppressive or offensive enough to discourage participation. Furthermore, the data indicated that the most prevalent types of toxicity are obscenities and insults.

Our analysis also reveals that the dataset contains more non-toxic content than toxic content, indicating that it is unbalanced. Research also indicated that the dataset is not representative of the actual distribution of toxic content on Wikipedia Talk Pages (Wulczyn et al., 2017).

Finally, we did an experimental sampling of the dataset, which indicates that there are some minor discrepancies between our interpretation of toxicity and the one applied within the dataset. However, the potential bias of the dataset requires further study in order to prove anything conclusive.

6. How can we create an AI that identifies toxicity?

In this section, we will elaborate on the possibilities and challenges of AI in the field of toxicity-detection and identification. By using the TRIN-model as a framework we first present our general understanding of our artifact: The AI created through our ML example. Then we will present the different possibilities for ML within the field. Finally, we will introduce our own work with AI as an example and explain our design choices and limitations.

6.1 The AI as a technological artifact

In this section we will clarify how an ML-based AI fulfills the definition of a technological artifact. We do this by elaborating on three main points: 1. Technological function, 2. Creation through a manmade process and 3. Outcome and general purpose. Our goal is to first and foremost understand the specific traits of the AI we are attempting to create.

Technological function

Throughout this project the AI has been referenced in correlation with the more generally intended function: Detecting or identifying toxicity. However, within the scope of this project, the precise goal of our AI is far more narrow.

As mentioned previously, formally defining toxic behaviors within an internet context requires a distinction between intent and reaction. Furthermore, most definitions of toxicity are related to context, social ties, and community - meaning that there is generally a broad understanding of the term. This multifaceted definition of toxicity disallows making any one definition of what qualifies as ‘toxic’. But making an ML model that can detect and label toxicity requires the predictive task to be quantifiable and strictly defined, otherwise, it is impossible to evaluate the model and its ability.

This has led to most ML-projects interpreting toxicity pragmatically. As our AI is created through ML on an existing dataset, we are extrapolating our definition of toxicity from the dataset itself. As our previous analysis has shown the selected dataset defines toxicity as content that hinders fair discussion, especially insults or obscene content. We can therefore conclude that this is the exact type of toxicity that our AI should be able to detect or identify.

However, this also requires a clear understanding of what is implied by ‘detection’ or ‘identification’. Detection and identification are seen as synonymous within this report and the words will therefore be used interchangeably. Our general goal is only to identify toxicity in a binary fashion, by categorizing content as ‘toxic’ or ‘non-toxic’. This is despite the fact that the dataset has different toxicity types and could therefore possibly be used to subdivide toxicity into specific categories. However, we would argue that these categories; obscenity, threats, insults, and identity hate, are implicitly included in the broader definition of toxicity within the dataset.

This means that detection of toxicity is defined as a binary classification of content that fulfills the definition of toxicity interpreted from the dataset. In other words, the AI will have successfully detected toxicity if: 1. The input content is categorized as ‘toxic’ or ‘non-toxic’ and 2. The content categorized as toxic fits the toxicity-definition from the dataset, while the non-toxic content does not.

It is worth noting that this is a very broad definition of success and furthermore one that requires a qualitative evaluation, which can be subject to bias. This is because the definition of toxicity as: ‘content that hinders fair discussion, especially insults or obscene content’, is broad and somewhat vague. This subjects the definition to a high degree of personal interpretation. Because of this a broad qualitative assessment is necessary to validate results as either successful or unsuccessful. A problem that we will address in the evaluation of our AI example.

Conclusively the technological function of our AI is defined as: Detecting toxicity that hinders fair discussion through the binary classification of content as either ‘toxic’ or ‘non-toxic’.

Creation through a manmade process

In this case ‘a manmade process’ refers to the process through which our artifact is created. While our development-section will go into specific detail on the creation of AI, the intent of this section is to examine how the creation process is affected by humans. Our goal is to present the ways in which people can affect the AI through the process of creation by using our AI as an example.

The argument for examining the process of creation and the potential biases that occur is based on Langdon Winner’s (1989) notion of technologies' effects on societal change.

In accordance with Winner (1989) we acknowledge that there is a need to reflect on social repercussions when designing, planning, or making any artifact - Thereby owing to how artifacts can shape our society and in retrospect, the world. This is especially relevant in the context of toxicity, which is a topic for much debate due to its direct correlation with everyday online behavior. Because of this, a reflection on biases and how we affect the process of AI creation is an inseparable part of our project.

First, we will begin by acknowledging that our general lack of experience and understanding of ML-processes has naturally affected the complexity of our AI-example. Within this project ML is primarily used as a blackbox technology, meaning that the specifics of applied parameters and configuration is - for the most part - unknown to us. Because of this, we rely primarily on the default configuration. This is based on the assumption that the developers selected a default configuration that is applicable to most ML-projects. While this is the best solution for our project due to scope and timeframe, more experience with ML would likely result in parameters more attuned to our specific purpose.

Secondly, we must clarify that the quality of our example, like any other ML-based AI, is a reflection of the data-input.

Our dataset is a result of Conversation AI’s research on toxicity. This means that the general purpose of the dataset remains the same; studying toxicity. Furthermore, our dataset was created in partnership with Wikimedia and the comments for the dataset were sourced from Wikipedia Talk Pages (Conversation AI. 2019). This means that the data is also a reliable reflection of actual internet content while limited to the context of Wikipedia Talk Pages. There is also the advantage of the dataset being used in a Kaggle competition with the goal of creating toxicity detection. This means that the dataset is already labeled and contains a test-sample with no annotations, so that any AI may be tested on content from the same context as the training-content.

However, in using this dataset we are inscribing their definition of toxicity into our project. Furthermore, we are limited by the amount of information that has been released on this specific dataset. This hinders

validation and bias-assessment, as we have no specific knowledge of the annotation-work that went into the dataset or knowledge of who the annotators were. While examining the data can to some extent provide knowledge on present biases, we cannot deny that the lack of knowledge affects our ability to judge the quality of our AI. Ideally, we would have knowledge of the questions posed to annotators as well as information on the identity of the annotators, such as sexuality, gender, and culture. A dataset containing such information would provide a better argument for the validity of the dataset and therefore the validity of the final model.

Finally, the goal of this project has also affected the creation process for the AI. While our ML-example provides an opportunity for discussion and reflection, creating an applicable and accurate AI is not the main goal of this project. This means that the creation of the AI has an experimental nature and is used more as a tool for examining potential challenges and gaining knowledge, rather than a goal in itself.

A project focused entirely on creating an optimal AI for the detection of toxicity would likely result in an AI that is different from our example. While the goal of the AI is to detect toxicity, the goal of this project is to understand the process and the challenges of detecting toxicity using ML. Because of this creating the AI has been a lesser effort, than understanding and examining the general process of creation.

In conclusion, our artifact, which is the outcome of our work and the work of Conversation AI, is affected by human choice in many ways. The key ways are: Our lack of experience, the advantages and constraints of the dataset from Wikimedia, and finally the general scope of this project.

Outcome and general-purpose

The definition of a technological artifact requires that it has the purpose of fulfilling a human need. In other words, it is the outcome and the general purpose of the artifact. In the case of our AI, we will explain the expected outcome and the general purpose that the AI would ideally fulfill. This is done to present what purpose the AI fulfills within this project and compare it to the purpose of a fully developed and improved AI with the same intended function.

The expected outcome of our ML-example is not necessarily an applicable AI. While our AI can certainly be argued to have some usability in the field of toxicity detection, it is not intended to be a viable solution for the problem of toxicity online. The AI is instead expected to illustrate some of the challenges when creating ML-models for toxicity detection. However, this requires the AI to fulfill some basic, practical expectations.

One expectation is that the final model will be functional and able to make a somewhat accurate ‘prediction’ i.e. define new content as either ‘toxic’ or ‘non-toxic’ on a binary scale. In this case, ‘new content’ refers to content that is not part of the labeled training- or test-data. This is necessary to evaluate the model, but it is also a crucial part of our research. By evaluating the model it becomes possible for us to see how the general trends of our dataset affect the outcome. Furthermore, it makes it possible to reflect on our specific use of methods and how our lack of experience potentially affected the model. Finally, if the model is not at least ‘somewhat accurate’ it becomes clear that the entire process is subject to fundamental error. This would mean that our general process cannot be viewed as an acceptable example of ML-development. This would mean that our gained knowledge and research are not generally applicable to the process of creating an ML-based AI. In making a functional model we confirm that our process is at least fundamentally correct, which serves to accredit the knowledge gained in the process.

However, ideally, the purpose of the AI would be more general. Since the intended function of the AI is to identify toxicity it could functionally be applied to any number of use cases within online moderation and content-filtering.

The utility value of toxicity detection is that it enables an AI to recognize toxicity on an online site. This means that the manual task of regulating online content is, to some extent, automated through AI. While definitions of toxicity vary they all share the feature of 'being harmful' to some extent. Therefore 'toxic' content is understandably not desirable. However, when moderators have to manually detect toxicity, it becomes an overwhelming task on any online site featuring a large user base - and as such it requires a lot of moderators (Brassard-Gourdeau, E & Khoury, R. 2019). An AI could potentially decrease the amount of manual labor and therefore also the number of moderators necessary.

Furthermore, an AI applies its inherent rules universally. This means that the way AI judges content will always be representative of the data it is trained on (Pratt, M.K. n.d.). As a result, the amount of individual assessment in terms of toxicity would decrease. Put simply the AI applies the same bias on everything, in contrast to moderators who are each subject to their own personal bias. While an AI cannot exactly be deemed fair in the sense that it is unbiased, it can be deemed fair in the sense that all content is processed the same way.

Based on the toxicity definition from our specific dataset an ideal AI would be able to regulate online discussions on Wikipedia Talk Pages. It would serve as a way to flag comments that are suppressive to certain views and expressions, resulting in a more nuanced and fair discussion. However as no perfect AI can exist, it would simultaneously in itself be suppressive to the types of comments that are wrongfully flagged as toxic. One way this could be addressed is by having moderator surveillance of the AI. The moderator would confirm that the flagged comments are toxic and as such minimize the number of false positives. This means that ideally, the AI could serve as a helpful tool for moderators to create fair discussion. But it would not be possible for the AI to entirely replace the site moderators.

In conclusion, our AI is only intended to function as an example of the ML-process when creating a toxicity detection model. But in an ideal use case applying a more well-developed version of our AI could aid moderators in regulating the Wikipedia Talk Pages and generally decrease the workload associated with moderation.

Summary

In the scope of our project, an ML-based AI fulfills the definition of a technological artifact, because it has the technological function of detecting toxicity that hinders fair discussion, and is created by us as part of a manmade process proven by the biases present in the finished prototype, and is designed with the intention to fulfill moderators' need to be aided in regulating the Wikipedia Talk Pages for toxic content.

One of the core traits of our AI is that it is designed to detect only a quantifiable and strictly defined type of toxicity, namely toxicity that hinders fair discussion. This is done through binary classification where a comment is classified as either 'toxic' or 'non-toxic'.

Our example of a toxicity detection AI is furthermore susceptible to bias due to the data-input from our chosen dataset and because we focus on examining the challenges of making such an AI rather than the creation of one. As a result, the AI has a lack of complexity both due to our knowledge and because we

outsourced the creation of specific parameters and configuration, so the AI itself ended up built on black box technology.

We do, however, still argue that the intention or purpose of the AI has not changed, and therefore if an ideal and well-developed version of our AI was created, it should aid moderators in regulating the Wikipedia Talk Pages and decrease the workload associated with moderation.

6.2 Possibilities of ML

In this section, we will analyze the results of our ML prototype tests against ML theory, and evaluate what possibilities we see in our conclusions, and what alternatives we believe to consider.

Models of Technologies

While we believe our intended AI to indeed have the potential to aid moderators moderating websites like Wikipedia Talk Pages, a badly executed AI may do the opposite. A big problem seen in AI is bias. An AI can end up discriminating against others, and this can lead to outrage(Hao, K. 2019).

Morally, even when an AI works, its utility is decided by who and how they intend to use it. Considering that AI potentially can be used to violate people's right to free speech. Creating an AI essentially functioning like a snitch is a risk in itself even when the intention is good. This is especially the case for binary classification since it is very clear-cut in options. There are simply no gray areas, so either the AI is wrong or it is right.

In terms of binary classifiers, the more popular options include logistic regression, k-nearest neighbor, and decision trees. However, all these have in common that they are somewhat rigid in execution. As such, they all have the potential for errors.

Comparing them, logistic regression seems less suitable for a toxicity prediction task, because it theoretically shouldn’t handle non-linear solutions well given its roots in regression logic. KNN is considered slower than logistic regression because its time complexity increases correspondingly to its dimensions. The decision tree, however, does not have this problem, because of its tree structure (Varghese, D. 2018). Based on efficiency one could assume a decision tree to be faster than both the logistic regression and KNN option.

Conversely, logistic regression can derive confidence levels about its predictions, whereas KNN and decision trees only output the labels (Varghese, D. 2018). Which would mean Logistic regression would possibly have better accuracy.

In conclusion, a decision tree theoretically looks like the better classifier when testing how well the binary approach can detect toxicity because it scales well. Since it can handle more input more efficiently, it means that computationally it could prove more efficient regarding optimization in the realm of more data for better toxicity detection.

From our tests, the decision tree classifier further seemed to perform best given our confusion matrix results, and given our accuracy results seem to only count for the labeled data. However, this may not be the case if proper optimization is factored in when using the other algorithms.

Overall, our test showed that toxicity is unlikely to ever be detected perfectly, but according to tests, it should be possible to do it correctly to some extent. The model being only as good as the training data provided, we can observe that much of the error margin we experienced was due to an inadequate data set. The fact that toxic and non-toxic examples were unevenly distributed meant that the model became more likely to have false positives/negatives outcomes, and this is not desirable.

Conversely, giving the model more data may also result in the opposite outcome. Namely risking more unintended bias, because the huge amount of data would mean more margin for error. According to Jigsaw (Hanu, L. 2021), a way to address this type of bias would be to specify the prediction task as Jigsaw did with their subcategories in the data set we used. Another option is to switch to another NLP method than n-gram, so more context is factored in when the model gives a prediction. Since this could reduce the data, which the AI does not understand.

A third option could be deep learning. Khieu & Narwal (2021) argues their binary classification model LSTM was significantly optimized by supporting their binary classification NLP operations by using deep learning. Deep learning is essentially another way to optimize data mining for large quantities of data. What can be concluded from this, is that the amount of data a model is trained in is indeed a key step to improving an AI's performance.

This would also address the issues seen in our tests, such as the model making errors due to lacking understanding of languages, special characters, and lack of data. Or so one would think, however, multilingual toxic text detection is a challenge all on its own (Guizhe, S. et. al. 2021).

The language gap that exists between languages is difficult to bridge due to translation errors, which can happen when extracting grammatical and semantic features from text. Grammatically similar languages may be assumed to be easier to translate, but this is not a guarantee. Therefore, the best option is to train the model on more language material. The problem is just that not much of such language material is necessarily available, and this is an issue since such material optimally ought to be heavily supplemented for an AI for it to work well in this aspect.

Summary

Improvement or possibility for optimization in toxicity detection AI technology can be summarized to be three components; 1) a good algorithm time complexity because this optimizes input processing. 2) Equal representation for class distribution in a data set, so the AI does not make mistakes based on lack of data. 3) NLP method, because this is what allows the AI to read and analyze the ML model's input.
We concluded that within traditional binary classification algorithms, a decision tree may be the better option because of its good time complexity. On the other hand, the NLP method we used does not factor in the overall contexts for a sentence, and it may not be the best NLP option. Finally, we can conclude the amount of data an ML model is trained on defines its performance.

6.3 Development

Libraries

Here we will present libraries used for program development and explain their individual use and functions.

NumPy

NumPy is an open-source Python library. This library consists of multidimensional array objects and routines for processing these arrays. Its utility lies in mathematical and logical operations like linear algebra and matrices (Great Learning Team. 2022). In our case, we chose this library due to our intentions of using a confusion matrix as one of the performance metrics to validate our ML model. Later we also employed it to perform vectorization due to a necessity of making our text input readable for the computer.

Matplotlib

Matplotlib is an open-source library known for creating static, animated, and interactive visualizations in Python. Built on NumPy arrays, it is useful for low-level data visualization and easy to use. Furthermore, it provides flexibility when making various plots (W3schools. n.d.). We used this library for data visualization when examining our data set.

Pandas

Pandas is an open-source Python library used for data cleaning and analysis. Meaning it is a tool to analyze, manipulate, and manage data structures. Pandas primarily use the standardized data science data-structure 'data frame', which is a structure that organizes data into a two-dimensional table of columns and rows (pandas. n.d.). It functions like a spreadsheet and as a result, it is simpler to understand, use and visualize. Because Pandas can take data input from various data files like CSV, JSON, Microsoft Excel, and SQL (pandas. n.d.), we initially chose it in consideration for its input versatility, but we have since used it for both import, export, and preprocessing of our data set.

Scikit-learn

Scikit-learn (sklearn) is a Python targeted open-source data analysis library. It features extensive Machine learning tools such as algorithmic methods like classification, regression, and clustering. Sklearn is designed to interoperate with popular Python libraries like NumPy, SciPy, and Matplotlib (ActiveState. 2021). For this reason, the library is versatile enough to support the integration of other libraries. We chose it due to this interoperability, its popularity, and because of the syntax similarity to other Python libraries. In our development, we use sklearn's algorithmic methods as the functionality to build our ML model.

Development process

In this section we will elaborate on our experience and process during the development of our toxicity detection AI prototype. This includes exploratory data analysis, preprocessing methods, model construction, validation, test, and final program improvements. All of which we have executed based on the Hua Lu (2022) inspired method: Three-Step Classification Model.

Exploratory data analysis

Prior to coding we had to verify that our dataset was applicable for our use case and could be used for testing our theories on a toxicity detection AI.

We did this through exploratory data analysis on our data. For this, we used the open-source web application Jupyter Notebook, and the libraries Pandas, NumPy, Sklearn, and Matplotlib. Primarily Matplotlib, as we intended to do data visualization on our findings to better illustrate our results. The code for this can be

found in appendix 3. The results are presented and analyzed in research question two, where we validate why and how we used our chosen dataset for this project.

Preprocessing

Based on research question three, we concluded that we were detecting toxicity through binary classification. Firstly, we use feature selection to focus our classification exclusively on toxicity. This means that we isolate data on the toxicity label, disregarding other labels such as: servere_toxic, obscene, threat, insult, and identity_hate.

Because we are applying supervised learning, we knew our input needed to be viable for our ML model. As a result, we had to find a way to transform our text input into a numeric feature instead of a string. We did this employing the natural language processing (NLP) technique 'bag of words' (BOW), also called 'count vectorization' in sklearn. As the name implies it is a process of encoding words as vectors, then counting their frequency to apply meaning. Specifically, we create a vector space where every word is indexed. Sentences are then represented by a vector that contains information on how often every word appears.

The sklearn's function 'CountVectorizer()' does this automatically, and has the added feature of removing the punctuation marks and converting all words to lowercase. All of our count vectorization is done using this function.

An alternative to this approach would have been to use the NLP technique N-gram. However, in terms of usage, we found 'BOW' more practical as the default options required less RAM. N-gram considers sentence structure to get an idea about the context of the particular word. Conversely, BOW doesn't extract context from a word. Rather it just builds vocabulary for the data set (Malik, F. 2019). As such, N-gram is more suited for our agenda. But due to hardware limitations, we could not implement N-gram and were therefore forced to use BOW.

Model construction

We created and trained an AI using three individual ML models. This was done using the pre-made training set from the Kaggle dataset. The code was built to run a single classifier at a time and was modified for each model. We ran each classifier separately before comparing and evaluating them.

While this may not have been ideal for testing, it was useful for comparing models and choosing the most adequate classifier for the prediction task. Model construction was mostly based on pre-made algorithms, and sklearns functions for each classifier. We only had to re-name the classifier variable based on the algorithm we wished to apply.

The downside to the general approach of applying existing algorithms and default values was that our entire model construction ended up being black box programming. It was therefore hard to examine non-visible parameters and potential flaws with our approach.

Model Validation & Test

To test which of our models was the most accurate, and thereby, which classifier is the most suited for the type of toxicity we wished to detect, we decided to compare their accuracy to validate the models. As such, we did a classic ML split with 80% for training the classifier and the last 20%, used for validations. The accuracy scores were as follows:

Classifier	Accuracy
Decision Tree Classifier	0.93839887200376 (0.94)
K-Nearest Neighbors	0.8690271032429892 (0.87)
Logistic regression	0.9574808083973053 (0.96)

Summed up, we can evaluate that logistic regression had the best accuracy score, decision trees coming in second. This is a deviation from our initial theoretical assumption about decision trees being the most suited for this type of classification task. However, this could change given we had yet to test it on unlabeled data. We therefore tested it on unlabeled data and displayed those results using a confusion matrices:

Logistic Regression

	Positive	Negative
True	0,1081056377109653 (10,81%)	0,31342669669963763 (31,34%)
False	0,5139359514249339(51,39%)	0,06451865635099402(6,45%)

Decision Tree

	Positive	Negative
True	0,1432442137564065 (14.23%)	0.2921881630920902 (29.21%)
False	0.4787973753794927 (47.88%)	0,08575718995854144 (8,85%)

K-nearest Neighbor

	Positive	Negative
True	0,12144419416968628 (12,14%)	0,30522638984102113 (30,52%)
False	0,5005973949662129 (50,05%)	0,07271896320961055 (7,27%)

The models all scored highly on false positive outcomes. In conclusion, the models are good at wrongfully deeming something toxic. Secondly, the models seem to predict non-toxic content correctly more often than toxic content. Of course, both tendencies seem to confirm that our dataset is skewed in favor of classifying non-toxic content rather than toxic content. Finally, the matrices seem to indicate the same as our initial expectations, namely that a decision tree classifier predicts the truest positive outcomes

We additionally did small qualitative tests to better understand the models. We created an input field to write a sentence, which the model could then predict. After testing it, we concluded that the model specifically had trouble understanding special characters, other languages, and swear words. The result was that the model labeled these types of content as toxic even when it was not. Especially swear words were an issue, because the nuances in the overall context were lost, which led to these being labeled toxic regardless of how the model was trained.

Program improvements & changes

During our model construction we experimented with implementing N-gram to address the issues of lost context. But we found that the N-gram method uses exponentially more RAM to store an Arraylist than BOW. For example, the default setting for N-gram used in Sklearn required 227 GB of RAM. There are solutions for this, but considering the scope of this paper, we disregarded implementing these techniques.

All the models are prone to wrong outcomes. This is possibly due to the content of our training set. Ideally, we should therefore get or create a more equally balanced dataset. Ideally, we would gather this data ourselves. But it has not been possible within the scope of this project. Furthermore gathering enough toxic comments to create a balanced dataset could pose a challenge, as our research has previously revealed that the general percentage of toxic comments on Wikipedia Talk Pages is only around 1 %.

Model evaluation

The prototype AI was primarily made to demonstrate the challenges of creating a toxicity detection AI.

When evaluating our model we can conclude that while the accuracy test did seem promising, it cannot dismiss the fact that it doesn't apply to unlabeled data. Furthermore, small qualitative tests revealed that our AI is prone to falsely identifying foreign languages, special characters, and swear words as toxic. In conclusion, the model works well in tests, but not in practice.

Possible optimizations such as more data and implementation of N-gram could improve the models. But as of now, it remains to be the subject of future testing.

6.4 Program description

Program structure

Toxic.py

Toxic.py was built to run and test our preprocessing methods and classifiers. The program will create an AI by training it to look through sentences and find the toxic ones. In Toxic.py there is a function that is called predictDataSet() which when given a Dataframe with the data set, it can make a prediction on every data point and then export the original data set with an extra column with the predictions. For documentation see appendix 1, diagram 2.

Matrix.py

Matrix.py was created for validation of our AI, the program is to create a confusion matrix based on a training set that trains the AI, a test set that needs to be predicted by the AI, and results set which have the actual results of the comments from the test set.

The program will run the preprocessing and make the AI just like Toxic.py, but it then goes into a loop that makes the predictions on the whole testing set and simultaneously compares the predicted values to the predicted one. After it has run through the test and result set it will give four numbers which are True Positive, True Negative, False Positive, and False Negative. For documentation see appendix 1, diagram 3.

Datasets

There are three data sets the two python programs will require.
Train.csv is the dataset used for training the AI. This dataset is split and used to train and validate the AI by making an accuracy score.

Test.csv is an unlabeled test set. It only contains comments and ids. Comments are needed for testing and ids serve as identification for the comment.This test set is used to test the AI with unlabeled data.

Test_labels.csv is the dataset that contains the labeled test data. This is because matrix.py needs the correct result for validation and the creation of a confusion matrix.

User Guide

If you want to recreate our findings using our code, here is a guide explaining the necessities for running the program.

The program requires python to be installed along with the libraries: Pandas, NumPy, and Scikit-learn (sklearn). We recommend using pip to install the libraries.

Import the datasets. These can be downloaded - or found in appendix 2. In our code, these are imported as Microsoft Excel and CSV files. But these can be used interchangeably.

```
data = pd.read_excel('C:/Users/rasmu/Desktop/Train.xlsx', names = ['id', 'sentence', 'toxic', 'severe_toxic', 'obscene' , 'threat' , 'insult', 'identity_hate'])
test = pd.read_csv('C:/Users/rasmu/Desktop/Test.csv', names = ['id', 'comments'])
result = pd.read_csv('C:/Users/rasmu/Desktop/test_labels.csv', names = ['id','toxic', 'severe_toxic', 'obscene' , 'threat' , 'insult', 'identity_hate'])
```

Once this is done, the program should be able to run. The default classifier used is Decision Tree Classifier, but if it is desired this can be changed on line 53 using the proper sklearn function.

```
53    classifier = LogisticRegression()
```

The program comes preloaded with K-Nearest Neighbor and Logistic Regression in the libraries, but more can be added.

7. Results and discussion

7.1 Toxicity based on intent or reaction

Our research has indicated that toxicity can be defined in various ways, most of which are dependent on the surrounding context, personal views, and community. Furthermore, we have provided examples of how ML-studies pragmatically approach the task of defining toxicity. In this section, we discuss the ethical implications of the pragmatic approach which values reaction above intent.

For the sake of this discussion, we will apply the understanding of moral judgment processes explained by Cushman et. al. (2013). Moral judgment is defined as a competitive interaction between two independent processes: The judgment of outcome and the judgment of the mental state of the perpetrator. The final ruling on the morality of a situation is only achieved when both processes have been taken into account. This means that any moral ruling is dependent on a judgment of both outcome and mental state (Cushman et. al., 2013).

In the case of our toxicity detection AI, the lack of context information poses a substantial challenge for the assessment of intent. The available information on the mental state of the perpetrator is limited to a single comment. While some comments can contain wording and references that indicate a causal relationship to other comments - e.g. retaliation or reaction - there is generally not much to judge intent on. This means that judgments are based almost entirely on the perceiver's reaction to the content i.e. the outcome.

This has the consequence that our model is not able to discern any mental state factors that might alleviate the perceived severity of toxic behaviors.

A relevant mental state factor could for example be if the perpetrator was being provoked or victimized. This consideration could be especially relevant because it has been shown that retaliation is a somewhat common reaction to toxicity (Kordyaka et al., 2020). Furthermore, other mental stage factors such as the surrounding community and their common behaviors are also disregarded, despite possible conditioning that would cause a higher tolerance for toxic behaviors (Kordyaka et al., 2020).

While it can be argued that disregarding these factors results in unfair judgment, Cushman et. al. (2013) reveals that the general focus on outcome is not uncommon for moral judgment. In the case of "*Blame blocking*" (p.8), the absence or presence of a harmful outcome is the dominating factor of moral judgment. In the case of toxic content, it means that intending to be toxic and failing, is judged less harshly than being an unintentional perpetrator of toxic behavior.

On this basis, it could be argued that making judgments entirely based on the outcome is not entirely unethical, as the resulting moral judgment would be more affected by the outcome either way. But doing so also means disregarding the factors of context and community, which are proven to be a major part of how we define behaviors as toxic.

In conclusion, it is possible to make a moral argument against the necessity of intent assessment. However, in the specific case of toxicity, this results in several relevant factors being disregarded from any kind of toxicity evaluation. As such limiting the scope of our AI model to the detection of toxic outcomes has the added effect of limiting our general definition and understanding of toxic behaviors. In other words, toxicity becomes a factor decided only by the outcome.

7.2 Suppression of content through AI fallibility

As our results have indicated our AI model is not perfect and we would argue that no such thing as a perfect AI is achievable within the field of toxicity detection. Therefore we will discuss the potential consequences of false positives vs. false negatives - and how these problems could potentially be addressed. Furthermore, we will discuss arguments for and against online moderation in the context of suppressing harmful content.

We have previously discussed the problem of our AI being potentially suppressive to comments that are wrongfully flagged as toxic, the false positives. We surmised that having moderators monitor the AI's classifications could potentially address the problem of false positives.

This estimated solution could in itself cause problems. One being the potential amount of data. Assuming that the AI would monitor pages with large userbases and countless forums, even a minor classification error could end up being replicated across massive amounts of content. For example, if the AI simply classifies any usage of the word "fuck" as toxic, the resulting flagged data would have a significant amount of false positives. This would require the moderators to reassess massive amounts of classified data, diminishing the value of implementing an AI.

In contrast, false negatives are not as detrimental to the general use of the AI. While falsely classifying toxic content as non-toxic is not ideal, this problem could potentially be solved by having users report content that is visibly toxic. However as the general trend of just ignoring toxic content has been documented (Kordyaka et al., 2020), it would not be a perfect solution. By supplying the AI with more data on toxicity, perhaps through a balanced dataset, it would be possible to decrease the number of false negatives. In the case of our AI, the dataset is especially skewed with far more data on non-toxic content than toxic content. As such providing it with more data on toxicity could potentially improve its ability to classify toxic content.

In general false positives could be considered the biggest problem for this type of AI, because it inherently causes unfair censorship of content.

A study by Sherrik & Hoewe (2018) on the moderation of online newspapers revealed that evidence of censorship directed towards specific viewpoints could be the cause of reader backlash towards presented opinions. We can speculate that this type of causality would similarly present itself on the Wikipedia Talk Pages in the occurrence of wrongful censorship. This could cause a general backlash towards the site itself. Furthermore, the added awareness that certain viewpoints are suppressed, could skew potential discussion in favor of suppressed views, as is also demonstrated by Sherrik & Hoewe (2018). This would defeat the ideal function of the AI, which is to facilitate fair discussion.

A possible solution to this issue involves clear communication and careful consideration of the moderation practices. According to Sherrik & Hoewe (2018):

> "*News and other sites with an interest in the quality of conversation on their sites should be cautious with and conscientious about their comment moderation policies as well as the language used to express those policies*" (Sherrick, B & Hoewe, J. 2018, 2018, p. 471)

This could indicate that an applying an AI as a tool for moderation is risky. The AI can be designed specially to minimize the number of false positives and thereby the amount of falsely flagged comments. However, this would provide no guarantee against backlash, as the site's communication about their moderation policy is also relevant. Documenting the use of the AI as a moderation tool would require though information on the

applied AI and how it was tested and evaluated. This might prove a more difficult task than that of communicating a general moderation policy that is enforced by site moderators.

To summarize the fallibility of AI makes it a somewhat risky solution for online moderation, with one of the bigger problems being false positives and thereby an unwarranted suppression of content.

Overall it is indicated that the AI, no matter how ideal, cannot be a solution to the moderation problem in itself. Any application of the AI would simultaneously require some sort of human qualitative evaluation. So while it can be argued that using AI would in some cases decrease moderator workload, it will not entirely eliminate the need for moderation work.

7.3 Our AI model

Using sklearns count vectorization as our NLP technique means that when we vectorize a sentence it outputs an array that contains indexes of the words of that sentence. This is practical because we get a sequenced list containing n-words, which will then be used to encode the likelihood of the phrase appearing again.

While it is a quick, and easy way to do NLP, splitting the sentences into individual words and then evaluating them separately ruins the sentence structure. Had it been clustered knowledge like the weather or stock markets this would not be an issue. However, language is built on accumulated knowledge from centuries of grammar, dictionaries, and added meaning. As a result, the sentence structure in language determines the meaning, and the structure can therefore not be ignored. (BiText. 2017)

The sklearn count vectorization method ignores this. As an example, our black box test repeatedly deemed a sentence that used the word ‘fuck’ to further enforce an otherwise completely positive compliment, as toxic. Through this, we realized that the AI was responding to commonly used swear words. In other words, it recognized the context behind the word, but not the context of the overall sentence.

This is a flaw in a considered design for a toxicity detection AI as toxic language often appears as more than swearwords. The implication is that an AI ought to have some language understanding for it to succeed in toxicity detection. In conclusion, the applied method works in theory and can be applied to the AI, but the result can be argued to be an AI that only works on the training data.

From this, we can also conclude that the performance of an AI, algorithm functionality aside, solely relies on its input. This is also possibly why AI bias is so hard to fix. Regarding bias, many types can exist. The ‘unknown unknowns’ is the kind of bias you don’t realize you possess and is input to the machine. Then there is the ‘lack of social contexts’ for the AI, which inevitably means an AI gets taught to frame problems in a certain way to fill this gap. Finally, there is the ‘definition of fairness’, where politics or religion plays a role (Hao, K. 2019)

Morality always has intent behind it, whether belief, rules or necessity (Hao, K. 2019). Due to this fact, the AI technology can be argued to always fulfill Winner’s concerns regarding an artifact driving/affecting societal change. While this bias may not be as intentional or unintentional as it could be, it is still an aspect to consider when trying to view AI technology without looking for conscious conspiracies or malicious intentions as Winner advises against.

Summarized, AI technology requires the ability to analyze full sentence structures to not be flawed by design, and only have high accuracy for validations tests. Bias is furthermore an implication of the technology rather than a side effect. Meaning it cannot be removed, only acknowledged and controlled. Finally, this remains an ongoing issue, because an AI's performance can be tied to the amount of input it is given. If input is considered equal to new bias, then an AI may be described as more biased the better its performance.

8. Conclusion

Our research has so far revealed that there are many aspects of creating a toxicity detection AI that needs to be taken into consideration. We can conclude that the general approach of gathering data and applying machine learning to create an AI is not enough in itself to provide a meaningful solution to the problem of AI toxicity detection.

Identifying and filtering content into categories of toxic and non-toxic is a task faced with a number of fundamental challenges.

The first challenge is defining toxicity in a simple and practically applicable way. This requires that we apply a specific understanding of toxicity that is tailored to the community and the overall cultural context. For this definition to be considered valid it has to be validated within the context of its appliance. In a broader context defining toxicity requires a discussion on the ethics of making definitions based on outcome and reaction, rather than intent.

The second challenge is understanding the way our data inflict bias on our AI. In order for any AI initiative based on machine learning to be considered valid, there has to be a high degree of data clarity. Specifically, any use of data requires thorough documentation and validation, as data can reflect many different types of bias. Addressing this challenge requires quality testing of applied data and evaluations of bias on results.

The final challenge is that of reflecting on the outcome of machine learning processes and understanding the specifics of a created AI. We would argue that applying AI on the basis of perceived convenience is a generally flawed approach to problem-solving with AI. Applying any type of AI requires an understanding of the applied ML approach and its implications. Furthermore, it is relevant to consider what the AI is actually capable of, based on its data input.

Finally, we have found that discussing and understanding the potential consequences of the technologies before they are implemented, can sometimes prove that these technologies are perhaps risky or less than optimal solutions for the intended problem.

9. Reflection

The original intent of this project was to simply create an AI to detect toxicity and test its use. However, as the process of doing this was explored we encountered many interesting challenges. Because of this the primary goal instead became to explore these challenges and their implications for the potential uses of ML-based AI in the field of toxicity detection.

While the scope of our applied use of ML to create an AI is rather small, the challenges we encountered are relevant to many similar projects within the field. Even if we cannot justify the usability of our prototype AI, we can at least assume our research of the associated challenges to be of relevance to other projects.

This research can also be considered the basis for further discussion on the subject of bias in data and its consequences for ML approaches, as well as how these are researched.

10 References

10.1 Bibliography

Pensum

Muller, A.C & Guido, S. (2016). *Introduction to Machine Learning with Python* (1st ed pp. 10-25). O'Reilly media, Inc

Fry, B. (2007). *Visualizing Data: Exploring and Explaining Data with the Processing Environment* (1st ed. pp. 1-18). Ch. 1. O'Reilly Media, Inc.

Hu Lu. (2022). *Data Science and Visualization* (DSV, F22). 4. Classification (I). Roskilde University. Appendix 1, Source 1: Hua Lu lecture slides

Jørgensen, N. (2019). *Digital signatur; En eksemplarisk analyse af en teknologis indre mekanismer og processer* (pp. 1-58). Ch. 1- 4.

Sismondo, S. (2009). *Introduction to Science and Technology Studies* (2 ed. pp. 72-80). Wiley-Blackwell. Ch. 7

Winner, L. (1989). *Technology as forms of life. In The whale and the reactor. A search for limits in an age of high technology* (pp. 3-18). Chicago: University of Chicago Press.

Websites

ActiveState.(21 Sep. 2021). *What Is Scikit-Learn In Python?*. Retrieved from: https://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in-python/. (Last visited: 30-05-2022)

Adinolf, S., & Turkay, S. (2018). Toxic Behaviors in Esports Games: Player Perceptions and Coping Strategies (p. 365-372). Retrieved from: https://doi.org/10.1145/3270316.3271545. (Last visited: 31-05-2022)

BiText. (2017). *How Phrase Structure can help Machine Learning for Text Analysis*. Retrieved from: https://blog.bitext.com/limitations-of-traditional-approaches-to-text-analysis. (Last visited: 31-05-2022)

Brassard-Gourdeau, E & Khoury, R. (2019). *Subversive Toxicity Detection using Sentiment Information*. Retrieved from: https://aclanthology.org/W19-3501.pdf . (last visited: d. 27-05-2022)

Brownlee, J. (April 8, 2020). *4 Types of Classification Tasks in Machine Learning*. Retrieved from: https://machinelearningmastery.com/types-of-classification-in-machine-learning/. (Last visited: 24-03-2022)

Brownlee, J. (2017). *A Gentle Introduction to the Bag-of-Words model*. Retrieved from: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. (Last visited: 29-05-2022)

Cambridge Dictionary. (n.d.). *Toxic definition.* Retrieved from: https://dictionary.cambridge.org/dictionary/english/toxic. (Last visited: 31-05-2022)

Chang, L. (2016). *Google has a new plan to fight internet trolls, and it starts and ends with AI*. Retrieved from: https://www.digitaltrends.com/cool-tech/conversation-ai-trolling/. (last visited: 16-02-2022)

Conversation AI. (July 28, 2019, b). Jigsaw Unintended Bias in Toxicity Classification. Retrieved from: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification. (Last visited: 24-03-2022)

Conversation AI. (n.d.). *Conversation AI*. Retrieved from: https://conversationai.github.io/. (Last visited: 16-02-2022)

Cushman, F., Sheketoff, R., Wharton, S. & Carey, S. (2013). *The development of intent-based moral judgment*. Retrieved from: https://www.researchgate.net/publication/234133369_The_Development_of_Intent-Based_Moral_Judgment. (Last visited: 31-05-2022)

Devopedia. (2021). *N-Gram Model*. Retrieved from: https://devopedia.org/n-gram-model. (Last visited: 25-05-2022).

Draelos, R. (2019). Measuring Performance: The Confusion Matrix. Retrieved from: https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/. (Last visited: 31-05-2022)

Gaba, I. (2022). *What is GitHub And How To Use It*?. Retrieved form: https://www.simplilearn.com/tutorials/git-tutorial/what-is-github. (Last visited: 31-05-2022)

GeeksForGeeks. (2022). *Decision Tree*. Retrieved from: https://www.geeksforgeeks.org/decision-tree/. (Last visited: 31-05-2022)

Google Research. (n.d.). *Lucas Dixon*. Retrieved from: https://research.google/people/LucasDixon/. (Last visited: 31-05-2022)

Guizhe, S., Huang, D. & Zhifeng, X. (2021). *A Study of Multilingual Toxic Text Detection Approaches under Imbalanced Sample Distribution*. Retrieved from: https://www.proquest.com/docview/2532337044. (Last visited: 31-05-2022)

Great Learning Team.(11 Jan. 2022). *What is Numpy in Python | Python Numpy Tutorial*. Retrieved from: https://www.mygreatlearning.com/blog/python-numpy-tutorial/. (Last visited: 30-05-2022)

Hao, K. (2019). *This is how AI bias really happens—and why it's so hard to fix*. Retrieved from: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. (Last Visited: 31-05-2022)

Hanu, L. (2021). *How AI Is Learning to Identify Toxic Online Content. Retrieved from*: https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/. (Last visited: 31-05-2022)

Joby, A. (2021). *What Is K-Nearest Neighbor? An ML Algorithm to Classify Data*. Retrieved from: https://learn.g2.com/k-nearest-neighbor. (Last visited: 31-05-2022)

Khieu, K & Narwal, N. (2021). *CS224N: Detecting and Classifying Toxic Comments*. Retrieved from: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6837517.pdf. (Last visited: 31-05-2022)

Kordyaka, B., Jahn, K., & Niehaves, B. (2020). *Towards a unified theory of toxic behavior in video games* (30[4], 1081–1102). Retrieved from: https://doi.org/10.1108/intr-08-2019-0343. (Last visited: 31-05-2022)

Kotsiantis, S.B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. Retrieved from: https://books.google.dk/books?hl=da&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=machine+learning+supervised+learning+classification&ots=CZoxvxZBmq&sig=seR74EqEwct3ZBBkp4LtK-1MZBo&redir_esc=y#v=onepage&q=machine%20learning%20supervised%20learning%20classification&f=false . (Last visited: 24-03-2022)

Lawton, G (2022). *Data Preprocessing.* Retrived from: https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing . (Last Visited: 29-05-2022)

Dixon, L. (2018). *Hi! and Welcome to our first toxicity classification challenge*.

Retrieved from: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/46064. (last visited: 16-02-2022)

Martin, E. (1991). *The Egg and the Sperm: How Science Has Constructed a Romance Based on Stereotypical Male-Female Roles*. Signs: Journal of Women in Culture and Society, 16(3), 485–501. Retrieved from: https://doi.org/10.1086/494680. (Last visited: 31-05-2022)

Manoa. (n.d.). *Practices of Science: False Positives and False Negatives.* Retrieved from: https://manoa.hawaii.edu/exploringourfluidearth/chemical/matter/properties-matter/practices-science-false-positives-and-false-negatives. (Last visited: 31-05-2022)

pandas. (n.d.). *Package overview*. Retrieved from: https://pandas.pydata.org/docs/getting_started/overview.html. (Last visited: 30-05-2022)

Pratt, M.K. (n.d.). *machine learning bias (AI bias)*. Retrieved from: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-bias-algorithm-bias-or-AI-bias. (last visited: d. 27-05-2022)

Saji, B. (2021). A Quick Introduction to K – Nearest Neighbor (KNN) Classification Using Python. Retrieved from: https://www.analyticsvidhya.com/blog/2021/01/a-quick-introduction-to-k-nearest-neighbor-knn-classification-using-python/. (Last visited: 31-05-2022)

Sarandeska, I. (2019). *What is Kanban – A Quick Guide*. Retrieved from: https://kanbanzone.com/2019/what-is-kanban-quick-guide/. (Last visited: 31-05-2022)

Sherrick, B & Hoewe, J. (2018). *The effect of explicit online comment moderation on three spiral of silence outcomes*. Retrieved from: https://www.researchgate.net/publication/305887739_The_effect_of_explicit_online_comment_moderation_on_three_spiral_of_silence_outcomes. (Last visited: 31-05-2022)

Sheth, A., Shalin, V. L., & Kursuncu, U. (2021). *Defining and detecting toxicity on social media: context and knowledge are key*. Neurocomputing (p.312–318). Retrieved from: https://doi.org/10.1016/j.neucom.2021.11.095 . (Last visited: 31-05-2022)

Surbhi, A. (January 29, 2020). *Supervised vs Unsupervised vs Reinforcement*. Retrieved from: https://www.aitude.com/supervised-vs-unsupervised-vs-reinforcement/ . (Last visited: 24-03-2022)

Swalin, A. (May 2, 2018). *Choosing the Right Metric for Evaluating Machine Learning Models — Part 2*. Retrieved from: https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428 . (Last visited: 24-03-2022)

Thanda, A. (2022). *What is Logistic Regression? A Beginner's Guide*. Retrieved from: https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/. (Last visited: 31-05-2022)

Tsouroupidou, K. (2016). *Harassment Survey 2015 - Results Report*. Retrieved from: https://commons.wikimedia.org/wiki/File:Harassment_Survey_2015_-_Results_Report.pdf. (Last visited: 31-05-2022)

W3schools. (n.d.). *Matplotlib Pyplot*. Retrieved from: https://www.w3schools.com/python/matplotlib_pyplot.asp. (Last visited: 30-05-2022)

Wulczyn, E., Thain, N., & Dixon, L. (2017). *Ex Machina*. *Proceedings of the 26th International Conference on World Wide Web*. Retrieved from: https://doi.org/10.1145/3038912.3052591. (last visted: 31-05-2022)

Varghese, D. (2018). *Comparative Study on Classic Machine learning Algorithms*. Retrieved from: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 . (Last visited: 31-05-2022)

10.2 Appendices

Appendix 1: diagrams, figures & sources

Diagram 1: Pie_chart

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *Pie_chart*. Roskilde University

Content: A PNG file

Diagram 2: Predict dataset function

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (31 May, 2022). *Predict dataset function*. Roskilde University

Content: A PNG file

Diagram 3: Confusion matrix

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (32 May, 2022). *Confusion matrix*. Roskilde University

Content: A PNG file

Figure 1: Kanban board 1

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (31 May, 2022). *Kanban sceenshot 1*. Roskilde University

Content: A PNG file

Figure 2: Kanban board 2

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (31 May, 2022). *Kanban sceenshot 2*. Roskilde University

Content: A PNG file

Figure 3: Overview of label categories

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (15 May, 2022). *Class overview*. Roskilde University

Content: A PNG file

Figure 4: labels_as_%_of_total

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *labels_as_%_of_total*. Roskilde University

Content: A PNG file

Figure 5: toxicity_by_label_%

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *toxicity_by_label_%*. Roskilde University

Content: A PNG file

Figure 6: Toxic_correlation_map

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *Toxic_correlation_map*. Roskilde University

Content: A PNG file

Source 1: Hua Lu lecture slides

Source: Hu Lu. (2022). Data Science and Visualization (DSV, F22). 4. Classification (I). Roskilde University.

Content: A pdf file

Figure 7: Non_toxic_sample_comparison

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *Non_toxic_sample_comparison*. Roskilde University

Content: A PNG file

Figure 8: Toxic_sample_comparison

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *Toxic_sample_comparison*. Roskilde University

Content: A PNG file

Figure 9: toxic_total_sample_comparison_heatmap

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (24 May, 2022). *toxic_total_sample_comparison_heatmap*. Roskilde University

Content: A PNG file

Appendix 2: Toxicity detection AI test prototype

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (31 May, 2022). *Toxicity detection AI test prototype*. Roskilde University

Content: a zip with our code for the test AI introduced in development

Appendix 3: Data visualization

Source: Uldal, AT. R., Nielsen, R.N. & Weisel, O & Jensen, S. E. (25 May, 2022). *Data visualization*. Roskilde University

Content: a zip with our code for our data visualization