Machine Translation and Toxicity Detection in Finnish: A FinBERT Approach

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Data Analytics
August 2025
Anni Eskelinen

Supervisors:
Professor Filip Ginter

UNIVERSITY OF TURKU
Department of Computing

Anni Eskelinen: Machine Translation and Toxicity Detection in Finnish: A FinBERT Approach

Master of Science (Tech) Thesis, 60 p., 1 app. p.
Data Analytics
August 2025

In the age of social media, an overwhelming amount of content is generated by users, making automated content moderation essential for maintaining safe online spaces. While English dominates much of the internet, the need for content moderation extends to smaller languages, such as Finnish, where resources and tools for automatic toxicity detection are still limited. This thesis investigates the feasibility of building an effective Finnish toxicity detection model using unified datasets created through machine translation as a form of cross-lingual transfer.

The thesis builds on previous work that introduced a toxicity detection model for Finnish and two Finnish toxicity datasets: a machine translated Jigsaw dataset and a manually annotated test set built from Suomi24 comments. FinBERT, a Finnish pre-trained transformer-based model, is fine-tuned on machine-translated data and evaluated on a new manually annotated corpora made for the purposes of the thesis. The thesis explores how well data from other cultures works in the Finnish context, whether models generalize across datasets, and how safe and useful the models can be in practical use. The thesis uses both quantitative experiments and qualitative analyses, such as error examination and prediction explainability using integrated gradients.

Despite differences in cultural context, language, and label distributions, results show that unified translated datasets can support the development of robust models. The best-performing model achieved competitive results that were better than the existing model, although the model tended to prioritize recall over precision, occasionally flagging non-toxic content as toxic. While the resulting model is not a replacement for humans, it can serve as a valuable aid in moderation workflows and data preprocessing.

Alongside its theoretical contributions, the thesis offers practical resources: a new Finnish toxicity detection model, a new manually annotated test set and the machine-translated datasets, as well as code for unifying datasets, model training, and inference.

Keywords: natural language processing, language technology, artificial intelligence, machine learning, toxicity detection, machine translation

Contents

List of Figures

List of Tables

List of acronyms

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

DL Deep Learning

FFN Feed-forward Network

FN False negative

FP False positive

FPR False Positive Rate

GPT Generative Pre-trained Transformer

IAA Inter-Annotator Agreement

LLMs Large Language Models

ML Machine Learning

MLM Masked Language Model

NLP Natural Language Processing

RNN Recurrent Neural Network

ROC AUC Area Under the Receiver Operating Characteristic Curve

SOTA state-of-the-art

TN True negative

TPR True Positive Rate

TP True positive

1 Introduction

Social media platforms have huge user bases and the amount of content in terms of text posts and comments is massive. This has brought forth the need for automatic content moderation in online settings. Automatic toxicity moderation (e.g., detecting insults or hate speech) is one method that can help with that. English is a lingua franca meaning a lot of the content people consume and make is in English, but the web is still inherently multilingual and many smaller languages have a need for content moderation. Yet, there are no such models for every language and effort is needed to build these models.

Natural language processing (NLP) is a field of artificial intelligence (AI) which deals with text and speech data. It can incorporate machine learning (ML), or even deep learning (DL) methods. Deep learning methods are very commonly used nowadays and the most popular method has been the Transformer architecture used in large language models (LLMs) with ChatGPT most notably paving the way for generative models.

This work is a continuation of a previous work [8] where toxicity detection was explored through cross-lingual transfer learning. In the paper, a machine translated Jigsaw dataset [37], a manually annoted dataset based on Suomi24 comments [38], and a Finnish toxicity detection model were first introduced along with their results.

Machine translation as a form of cross-lingual transfer is a method that has been used to get data to many smaller languages to build models [8] or used as a way to translate data for English models to make predictions [15]. Cross-lingual transfer will be explained in detail in Chapter 2. The Jigsaw dataset as well as other datasets used in the thesis will be presented in Chapter 3 and related work regarding toxicity and similar tasks, will be introduced in Chapter 2.

The main research question I intend to answer with this work is as follows:

> To what extent is it possible to use machine translated unified datasets to build robust models for toxicity detection in Finnish?

To get an answer to this question, I will also try to answer the following sub-questions:

1. How well does a corpus from another culture fair in the Finnish context?
2. Does a model trained on one or multiple corpora work on a different corpus?
3. How safe and useful can the model be in actual use?

To answer these research questions I will do several experiments. I will use the existing model fine-tuned for the toxicity detection text classification task [8] to get results on the other translated datasets introduced in Chapter 3.2, explore the results of previous work, and get results on a new manually annotated dataset. Then, I will train a new model on a unified dataset consisting of more translated data and test it on many different datasets, and finally qualitatively analyze what the models can or have been used for, what kind of misclassifications there are with the best model, and what the model sees before making a prediction.

In summary, the aim of this thesis is to build a robust classification model for Finnish that can try to solve the issue of toxicity moderation by incorporating many different types of texts from different corpora in the training data to make the model as comprehensive as possible. The best model is available in HuggingFace[1] and the new annotated dataset as well as the translated datasets are available in Github[2].

[1] https://huggingface.co/annieske/bert-base-finnish-cased-toxicity

[2] https://github.com/anniesk/masters-thesis

2 Background

In this Chapter I will introduce the transformer architecture, BERT and FinBERT, and explain fine-tuning to a specific task. In addition, I will define toxicity, introduce related tasks and some previous results, explain what cross-lingual transfer is, as well as present commonly used metrics for evaluating machine learning models.

2.1 Transformers and FinBERT

Here I will introduce the architecture of transformers, BERT and FinBERT, and talk about fine-tuning models for different tasks.

2.1.1 Transformer

The transformer architecture was first introduced in the paper "Attention is all you need" [39] in 2017 and quickly gained popularity specifically in the field of NLP. The architecture overcame issues of capturing long-range dependencies in texts that Recurrent Neural Networks (RNNs) had, by discarding recurrence entirely and instead relying on a mechanism known as self-attention. Self-attention allows the model to consider all parts of a sequence (in my case, text) simultaneously and to dynamically weight their importance when computing contextual representations. This enables each token to attend to every other token in the sequence regardless of distance, allowing the model to capture complex syntactic and semantic relationships.

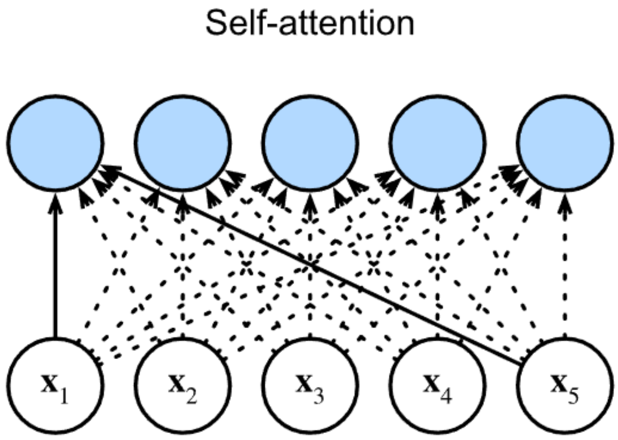

Figure 2.1: Self-attention visualized.[1]

In the self-attention mechanism, each token is projected into three vectors: a query Q, a key K, and a value V. These vectors are used to compute how much attention each token should pay to the others. The attention score between two tokens is calculated by taking the dot product of a query and a key. After applying a softmax to these scores, the result is used to weight the value vectors, producing a context-aware representation for each token. This lets the model integrate relevant information from across the sequence at every layer.

The transformer architecture consists of encoder and decoder structures. The encoder is tasked with reading and interpreting the input sequence and transforming it into a sequence of continuous representations. Each token in the input is first embedded into a fixed-dimensional vector space of 512. This was chosen as the fixed size because computational cost increases quickly with self-attention, as it scales quadratically with sequence length. Since the model does not inherently recognize the order of tokens, positional encodings are added to the embeddings to incorporate sequence information. These vectors are then processed through a

[1] https://d2l.ai/chapter_attention-mechanisms-and-transformers/self-attention-and-positional-encoding.html#comparing-cnns-rnns-and-self-attention

stack of identical layers, each consisting of two key components: a multi-head self-attention mechanism and a position-wise feed-forward network (FFN).

The multi-head self-attention mechanism allows each token in a sequence to attend to every other token including itself across multiple representation subspaces. Each head computes scaled dot-product attention independently using its own learned Query, Key, and Value projections. The outputs of all heads are concatenated and passed through a linear layer. The position-wise feed-forward network which consists of two linear layers with a non-linear ReLU activation in between, follows the attention mechanism and processes each token's representation independently and identically.

The decoder mirrors the structure of the encoder but introduces one crucial difference: in addition to processing its own inputs through self-attention and feed-forward layers, each decoder layer includes an encoder-decoder attention mechanism. This allows the decoder to attend not only to previous tokens in the output sequence but also to the encoder's representations of the input. During training, the decoder uses a masked self-attention mechanism to ensure that each token can only attend to earlier positions, preserving the autoregressive property needed for generation. Like the encoder, each component of the decoder is followed by residual connections and layer normalization.

Based on the transformer architecture, there have been encoder-only, decoder-only, and encoder-decoder models. BERT [6], which will be discussed below, is an encoder-only model which is suited for tasks like classification, question answering, and sentence embedding as the encoder processes the entire input and is able to generate conceptualized representations. GPT [28], on the other hand, is a decoder-only model and is able to generate text one word at a time. An encoder-decoder model can be used for sequence-to-sequence tasks such as machine translation, where a variable length input is mapped to a variable length output. This was the task the Transformer was tested on in the original paper [39].

2.1.2 BERT, FinBERT, and Fine-tuning

BERT (Bidirectional Encoder Representations from Transformers) [6] is a transformer-based architecture, specifically the encoder. It is a "multi-layer bidirectional Transformer encoder based on the original implementation" of Transformers which has been discussed above. There are two different sized versions of BERT: the base model has 12 layers, hidden size of 768, 12 self-attention heads and 110M parameters while the large model has 24 layers, hidden size of 1024, 16 self-attention heads and 340M parameters in total.

BERT uses a "masked language model" (MLM) pre-training objective, which "randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context." [6]. The MLM task allows the model to fuse the left(-to-right) and right(-to-left) context instead of only using one or the other, which was previously the generally used method to train word embedding vectors. This is where the term bidirectional in the name BERT comes from. As a second pre-training task, "next sentence prediction" is used which "jointly pre-trains text-pair representations" [6]. During pre-training, the model is trained on huge amounts of unlabeled textual data from BooksCorpus (800M words) [47] and English Wikipedia (2,500M words). The training data can be characterized as fairly neutral, but it can still have biased predictions for the MLM task, which will in turn affect fine-tuned models[2].

The BERT model uses WordPiece embeddings [44] with a token vocabulary of 30,000. A tokenizer is needed to encode text into a numerical representation for the model. An example of how the "bert-base-cased" model tokenizes input texts is displayed in Figure 2.2 as text decoded from the numerical embeddings. The

[2]https://huggingface.co/google-bert/bert-base-cased#limitations-and-bias

```
This is a text meant to visualize how texts are tokenized by the bert-base-cased
model.
[CLS] This is a text meant to visual ##ize how texts are token ##ized by the be
##rt - base - case ##d model . [SEP]
```

Figure 2.2: Example of tokenized text in English.

first token of every tokenized text is "[CLS]" which is a special classification token, and the final token is "[SEP]", which can act as a separator for sentence pairs [6]. The input representation for a token is "constructed by summing the corresponding token, segment, and position embeddings" [6]. The model is limited to 512 subword tokens on the input, meaning that longer texts need to be truncated.

FinBERT [40] is a model trained on Finnish which follows the training paradigm of the BERT model [6]. The model was trained on massive amounts of Finnish data from different sources such as YLE news articles, online discussions from Suomi24 and from internet crawls of the Finnish internet, and by doing language detection on Common Crawl. To ensure the quality of the training data, it was cleaned and filtered by first removing machine translated and generated texts by using a simple machine learning model trained on the FinCORE corpus [17], which includes several labels to categorize texts by their register (or genre) and one of them happens to be machine translation. Then the remaining documents were filtered using language detection and other heuristics.

A tokenizer and a dedicated BERT vocabulary for Finnish were also needed and built, which consisted of 50,000 wordpieces. For training the model, the same tasks were employed as for the original BERT model. Cased and uncased versions of the FinBERT model were made, as well as base and large versions of the model. To evaluate the FinBERT model, testing on several different fine-tuning tasks was performed, including two sequence labeling tasks where a label is predicted for each word, and two text classification tasks where a label is predicted for the whole text. The FinBERT model obtained better scores than those of the multilingual BERT.

Fine-tuning a model to a downstream task is done by taking the pre-trained model and tuning it to some specific task with annotated data. This is straightforward and done "by swapping out the appropriate inputs and outputs" [6] for each task "and fine-tuning all the parameters end-to-end" [6]. Fine-tuning is inexpensive compared to fully training a model from scratch and is a very common method to train models for different tasks. In the case of text classification, the fine-tuning is done by "simply attaching a dense output layer to the initial ([CLS]) token of the top layer of BERT" [40].

2.2 Transfer Learning and Cross-lingual Transfer

Transfer learning is a popular concept within natural language processing which "aims at improving the performance of target learners on target domains by transferring the knowledge contained in different but related source domains" [48]. As deep learning has become the norm for many different tasks, large amounts of data are needed but huge labeled datasets are almost impossible to get. Especially with the BERT and GPT models came a way to pre-train models on huge amounts of text data, and then later use that model by fine-tuning it on a task with a much smaller dataset [13]. This is what transfer learning is in its simplest form, as it can mean the same as fine-tuning a pre-trained model to some specific task because it leverages the knowledge gained from the pre-training.

Transfer learning can be categorized based on several criteria [48], one being the availability of data. In inductive transfer learning, the target domain contains some labeled data, allowing the model to be directly trained on the target task using transferred knowledge. In contrast, transductive transfer learning assumes that labeled data exists only in the source domain, with the goal of applying this knowledge to unlabeled instances in the target domain. When neither domain has labeled data and the task does not involve supervised prediction, the problem falls under unsupervised transfer learning. Another criterion on which transfer learning can be categorized, is the feature and label space of the source and target domains. If both the input features and output labels are the same across domains, the scenario is called homogeneous transfer learning. If there are differences in feature types, label sets, or both, the scenario is termed heterogeneous transfer learning.

In terms of methodology, transfer learning approaches are also classified based on what is being transferred. Instance-based methods focus on reweighting or selecting source domain instances that are most relevant to the target domain. Feature-based methods aim to find a shared or transformed feature space that reduces the distribution gap between domains which can involve either transforming only the source features to match the target (asymmetric) or finding a common latent space for both domains (symmetric). Parameter-based approaches transfer model parameters or use prior distributions learned from the source domain to guide learning in the target domain. Finally, relational-based methods are designed for relational or structured domains and aim to transfer logical relationships, dependencies, or structural patterns.

Cross-lingual transfer means leveraging pre-trained multilingual models and fine-tuning in one language and testing in another. It is generally used as the zero-shot version, where there is only test data for the new language, and all the training and validation for the task happens in a language which has more training data. An example of this is register (genre) labeling or register identification [17] where there have been good results using the XLM-RoBERTa (XLM-R) model [4] which is a transformer-based model that was pre-trained on text in 100 different languages and which outperformed multilingual BERT and obtained state-of-the-art (SOTA) performance on tasks such as classification, sequence labeling and question answering. The results for register identification indicate that multilingual zero-shot models trained on larger corpuses can beat models trained on only one language for which there is less data [30], [32]. Zero-shot cross-lingual transfer especially benefits smaller languages with limited annotated data.

For toxicity detection, in the paper by Eskelinen, Silvala, Ginter, *et al.* [8] XLM-R with zero-shot cross-lingual transfer on the Jigsaw dataset performed poorly. Machine learning as a mode of cross-lingual transfer was instead tested and performed better. Thus, instead of using a multilingual model like XLM-R for cross-lingual transfer, in this thesis the focus is translation as a type of transfer with the monolingual Finnish BERT model, as that was deemed to be the best performing approach, at least with the Jigsaw dataset.

2.3 Defining Toxicity

To define the task of toxicity detection, I have to first define toxicity. Toxicity itself is an umbrella term that can fit many different, but similar terms under it such as hate speech, abusive language, offensive language, and harmful language. These terms in turn can have their own definitions, and there can be variation about the definition even within the task [10].

In my thesis I have decided to adopt the same definition as in the paper by Eskelinen, Silvala, Ginter, *et al.* [8] as I use their fine-tuned model for some of my experiments. The same definition is used by the Jigsaw dataset [36] and Perspective API [24] which I will describe in more detail later in Section 3.1. The definition for toxicity is thus as follows:

> ”rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion”

Now that toxicity has been defined, toxicity detection (or identification or classification) is now easy to define as it is the detection of toxicity from texts, mostly online as most of the discourse happens on the internet in the many social media sites. This task can be either multiclass, where only one label is predicted for each text, or multilabel where the model can predict any number of labels or no label at all. In my thesis the focus is on multi-class, specifically binary with only two labels as to unify the many different datasets.

2.4 Previous Work on Toxicity and Related Tasks

There are many different tasks that can fit under the toxicity umbrella, and they include overlapping but yet different characteristics. Many datasets use their own definition for the specific subtask they are researching and build their dataset on that. This has led to the datasets being incompatible in the sense that annotations may differ greatly due to the definitions, and what is considered toxic may not be toxic in another dataset. These points have been brought up in many previous pieces of research [5], [9], [20], [33]. Related tasks that are similar to toxicity detection are at least hate speech detection, offensive language detection, abusive language detection and aggressive comment detection.

2.4.1 Labels and Annotation

As there is no one standard definition for toxicity or related tasks there are also no standard definitions for labels. Different use cases also exist and many researchers have opted to build their own datasets with their own definitions and guidelines for annotation. Some datasets use a binary classification where there are only two labels, others use a multiclass setting with varying amounts of labels and some use a multilabel distribution where many labels can co-exist at the same time. As such, labels within different datasets can vary greatly. Some of the variation regarding the datasets used in this thesis can be seen later in Chapter 3 in Table 3.3.

Many datasets use crowdsourcing for the annotation process [5], [9], [10] where workers on a data annotation platform are paid to annotate texts based on carefully built instructions. For some datasets the guidelines are not as detailed as for others, and as a matter of fact less detailed instructions should make the annotations correspond to how the general population categorizes offensive language or hate speech [10]. The number of annotators can vary from two to even 20 making data annotation a costly venture if large-scale datasets consisting of hundreds of thousands of comments are annotated.

The inter-annotator agreement which tells how well the annotators agreed on labels, has been low for many datasets which highlights the difficulty of the task for even humans, as the topic of toxicity can be quite subjective. Many datasets have reported an IAA of around 60%, e.g., Waseem and Hovy [41] reported a Kappa score of 0.57 and Eskelinen, Silvala, Ginter, *et al.* [8] reported a unanimous IAA of 63% but some have reported higher or lower results from as low as 15% to as high as 72%.

2.4.2 Languages

A great deal of toxicity (and toxicity related) datasets are in English as it is a lingua franca, although many related tasks have datasets also in other languages. The toxic comment collection paper [31] gathered many of the toxicity related datasets together and showed that there are datasets in e.g., German, French, Arabic, Italian and Portuguese. Finnish is unfortunately not one of them and we have to rely on machine translated data or cross-lingual transfer using multilingual models to get any reasonable results out of a classifier.

As mentioned, there are some smaller datasets in other languages (e.g., German, Arabic, Italian) [10] and datasets which have comments from many languages exist as well [9]. Code-switching is a phenomenon in people's speech where a multilingual person uses multiple languages in their speech with having either adapted word(s) or morphological or phonological properties from the other language [26], e.g., a person speaks in Finnish and says a word in English in the middle because it feels perhaps more natural. Code-switching in itself is a topic that has been studied in the NLP community [43] and recently the code-switching capabilities of LLMs have been evaluated [12], [46] with GPT-3.5 and GPT-4 showing promise in their ability for cross-lingual understanding although they are not quite there yet.

2.4.3 Dataset Sources and Bias in their Creation

Many of the toxicity and related datasets are from Twitter (now X) [31]. It is probable that toxicity in tweets differs from other social media outlets due to the previously 140, now 280 character without a subscription and other social media sites might not have that kind of small limit. Different social media sites also have different target groups and the people using them might have differences in how they write comments, which might affect how models that are built on only one type of dataset perform on other datasets when the domain changes.

For the collection of comments from Twitter or other online forums, many datasets have used keywords which can be taken from e.g., Hatebase.org which has been used by many datasets (e.g., [5], [41], [45]). This method will bias the distribution in the sense that more of the texts will be labeled toxic. This does not represent the real world where toxicity is still rather scarce and e.g., in the paper by Founta, Djouvas, Chatzakou, *et al.* [9] it was mentioned that "In the grand scheme of things, abusive tweets are quite rare (between 0.1% and 3%, depending on the label)". Although some might say that the internet can be a hateful place, most of the comments are neutral as opposed to some datasets e.g., Davidson, Warmsley, Macy, *et al.* [5] where hate speech or offensive language was found in 81% of the tweets whereas for e.g., the Jigsaw dataset [36] the number was only 11% because the data was gathered in a different way and from a different domain.

A problem with having Twitter datasets, is that due to Twitter's API's rules, saving the actual tweets in the datasets is not allowed[3]. Instead, datasets often include the ids of tweets and the tweets' texts are gotten through the API. This means that any tweet that is deleted is gone for good, be it deleted by the user, or if the user's account is suspended or deleted. Even if the tweets would be allowed to be saved as text, content compliance[4] means that any deleted tweets would still have to be deleted in 24 hours. This is especially a problem when it comes to toxicity datasets because racist or otherwise toxic tweets might be deleted later. This makes the distribution of classes in the dataset later different from the original intention.

Now as Twitter has been rebranded as X, academic access has been rescinded after almost two decades and data collection is not possible on the same scale anymore [1]. As explained by Blakey [1], this made a big impact for especially social researchers regarding e.g., research towards negative views on minorities as datasets were no longer available. The paper also notes that in 2023, X filed a lawsuit against a nonprofit researching hate speech.

2.4.4 Transfer

Transfer learning and cross-lingual transfer were introduced in Chapter 2.2 but in this section the focus is on transfer learning in toxicity detection and related tasks. This part is mostly implemented from the paper by Eskelinen, Silvala, Ginter, *et al.* [8].

Multilingual models have been widely applied to toxicity related tasks such as hate speech detection and offensive or abusive language classification. For example, Pelicon, Shekhar, Martinc, *et al.* [23] showed that a multilingual BERT-based classifier can achieve performance comparable to monolingual models in offensive

[3] https://developer.twitter.com/en/developer-terms/policy#4-d

[4] https://developer.twitter.com/en/developer-terms/policy#3-c

language detection. Similarly, Eronen, Ptaszynski, Masui, *et al.* [7] demonstrated that zero-shot cross-lingual transfer can yield competitive results in abusive language detection. However, challenges still remain. Nozza [21] highlighted that zero-shot transfer in hate speech detection may be hindered by language-specific, non-hateful taboo expressions that models incorrectly interpret as hate speech. Likewise, Leite, Silva, Bontcheva, *et al.* [18] reported that zero-shot transfer failed to produce accurate results for toxicity detection in Brazilian Portuguese.

As was also mentioned in Chapter 2.2, machine translation can be considered a mode of transfer learning. Machine translation has been used in toxic language detection to get more data by data augmentation [29] i.e. adding backtranslated texts (going through translation from e.g., English - Finnish - English) to generate parallel sentences and add noise to the data, or by translating data to English to be able to use ready-made English models [14]. Kobellarz and Silva [14] found that comments that were analyzed as toxic in Portuguese were not as toxic when translated to English, however, the same behavior may not apply to other language pairs.

2.5 Commonly Used Metrics

Commonly used metrics in the field of NLP and text classification specifically are precision, recall, F1, accuracy and ROC AUC. All the metrics except ROC AUC are based on a confusion matrix with true positive (TP), false positive (FP), true negative (TN) and false negative (FN) counts in the simples binary case. TP is how many positive samples the model predicted correctly, TN is how many negative samples the model predicted correctly and FP is how many negative samples the model predicted incorrectly which is also known as a Type-I error. Lastly, FN is how many positive samples the model predicted incorrectly. This is also known as a Type-II error. A simple example of a confusion matrix can be seen in Table 2.1.

	ground truth positive	ground truth negative
predicted positive	TP	FP
predicted negative	FN	TN

Table 2.1: Confusion matrix showing TP, FP, TN, FN

From the confusion matrix several evaluation metrics can be calculated. In the simplest binary case, they are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Accuracy is calculated with the number of correct predictions divided by the total number of predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

Precision is calculated with the number of true positives divided by the total positives. This focuses on Type-I errors which focus on false positives [3].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

Recall is calculated with the number of true positives divided by all ground truth positives. This focuses on Type-II errors which focus on false negatives [3].

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{2.4}$$

F1-score is the harmonic mean of precision and recall. It gives good results on imbalanced classification problems.

For F1-score there are many ways to calculate the metric, micro and macro averages possibly being the most popular with their difference being that micro "Calculate metrics globally by counting the total true positives, false negatives and false positives" whereas macro "Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account"[5]. In my thesis I

[5]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

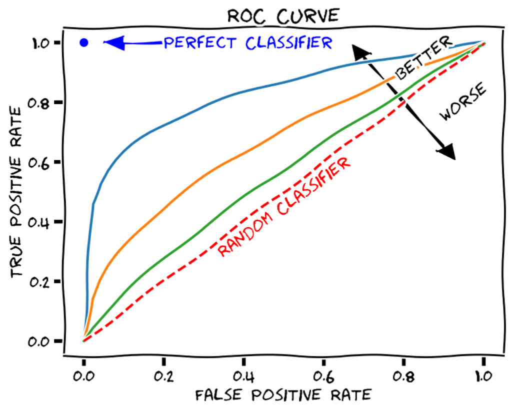

Figure 2.3: Example of a ROC curve [42].

use both versions when applicable, but in the binary case the micro average corresponds to accuracy and macro average is used instead for F1, precision, and recall. Extensions of the metrics exist for different types of tasks such as multiclass and multilabel and ready-made solutions for many metrics are available for example in the python library scikit-learn[6].

Area under the receiver operating characteristic (ROC) curve, also known as simply area under the ROC curve (AUC), is a common machine learning algorithm evaluation method which considers all possible threshold values instead of just one. It is meant for binary tasks, but extensions also exist for other types of tasks. The AUC is calculated by first calculating the ROC curve by using the true positive rate and the false positive rate at every possible threshold and then making a graph of TPR over FPR. The AUC (area under the curve) is then calculated with that graph. An example of a ROC curve is shown in Figure 2.3.

The mathematical formulas for TPR, FPR and both ROC and AUC are shown below. Any good model should do better than 50% (as a linear line in the middle from bottom left to upper right) as shown in Figure 2.3 and the perfect model which is correct all of the time would have a TPR rate of 1.0 and thus an AUC of 1.0.

[6] https://scikit-learn.org/stable/index.html

$$\mathrm{TPR}(t) = \frac{TP(t)}{TP(t) + FN(t)} \tag{2.5}$$

TPR is also known as recall or sensitivity and the mathematical formula is the same as the one introduced above, only this time all thresholds are used.

$$\mathrm{FPR}(t) = \frac{FP(t)}{FP(t) + TN(t)} \tag{2.6}$$

FPR is calculated with false positives divided by all ground truth negatives with all thresholds used in the case of the ROC curve.

$$\text{Receiver Operating Characteristic (ROC)} = \{(\mathrm{FPR}(t), \mathrm{TPR}(t)) \mid t \in [0, 1]\} \tag{2.7}$$

ROC curve defined as a parametric curve.

$$\text{Area Under the Curve (AUC)} = \int_0^1 \mathrm{TPR}\left(\mathrm{FPR}^{-1}(x)\right) \, dx \tag{2.8}$$

This is the integral version of the AUC but there are several ways to display the formula.

3 Data

In this section I will introduce the many datasets used in the experiments of the thesis and explain how the labels of the different datasets were unified for further experiments.

3.1 Jigsaw Toxicity Dataset

Jigsaw toxic comment classification [36] was a classification challenge made by Jigsaw, a Google team which builds models for fake news detection, toxicity detection etc. with the goal to make the internet safer. The dataset is in a multilabel format, and includes six subtypes of toxicity to train and test on. These labels are toxicity, severe_toxicity, obscene, threat, insult and identity_attack. The data was crowd-sourced and human-rated but no specific information about the annotation process or the labels can be found in the Kaggle competition site, only on the Perspective site [24], [25]. The definitions for the different types of toxicity are described in Table 3.1 below. Note that the label "toxicity" uses essentially the same definition as what was defined as the umbrella term in Section 2.3.

The data contains 223,549 comments, 159,571 and 63,978 in train and test, respectively. The distribution of the labels is shown in Table 3.2. Note that the total amounts to more than the number of examples, as the data is multilabel and thus only the "No label" matches directly to the number of examples. Most of the examples in both the train and test have no label (approximately 90%), meaning that

Label	**Definition**
Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Severe Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Threat	Describes an intention to inflict pain, injury, or violence against an individual or group.
Obscene (or profanity)	Swear words, curse words, or other obscene or profane language.
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
Identity Attack	Negative or hateful comments targeting someone because of their identity.

Table 3.1: The definitions of the labels as described in the Perspective API [24]

Label	**Train**	**Test**
Toxicity	15,924	6,090
Severe Toxicity	1,595	367
Threat	478	211
Obscene	8,449	3,691
Insult	7,877	3,427
Identity Attack	1,405	712
No label	143,346	57,735

Table 3.2: Label distribution in the Jigsaw Toxicity Dataset [8].

most of the examples in the dataset are neutral or non-toxic. The label distribution is thus highly imbalanced, but models the nature of the real world in the sense that most of the texts and discussions online are neutral and not toxic at all as mentioned already in Chapter 2.4.3.

Some labels co-occur more than others as can be seen in Figure 3.1, where small values close to zero indicate no correlation between the labels, while higher values closer to one suggest correlation and that the labels tend to appear together. In particular, labels "obscene" and "insult" tend to co-occur together and both of them also tend to co-occur with "toxicity". From here it can be noted that despite

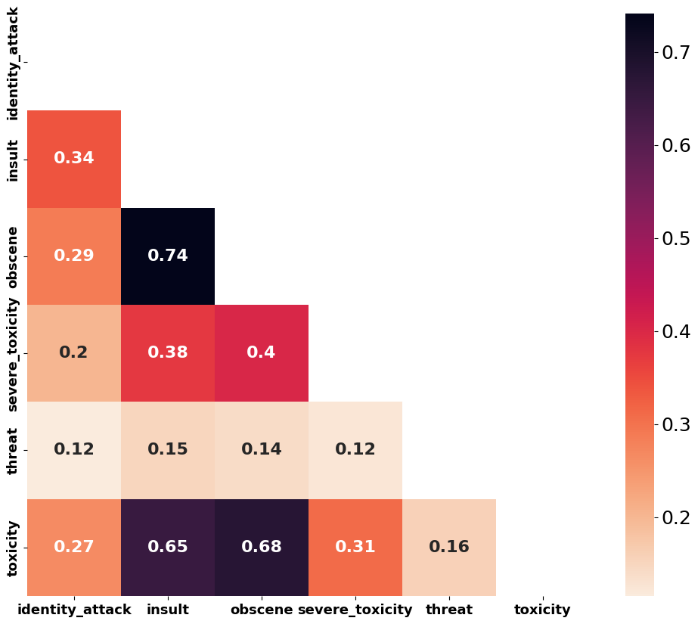

Figure 3.1: Correlation matrix of the labels of the original train dataset calculated with Pearson standard correlation coefficient [8].

the fact that toxicity is an umbrella term it does not fully correlate with the other labels, meaning it is not always used in conjunction with the other labels and the other labels can also occur by themselves. This could be interpreted in the sense that not every subtype of toxicity is inherently toxic, for example, swearwords may not always be toxic in itself and can be used positively. Another explanation could be that there is a problem with the annotations but that is impossible for me to conclude with certainty.

For the purposes of fine-tuning a model in the paper by Eskelinen, Silvala, Ginter, *et al.* [8] the dataset was translated to Finnish using the DeepL service and 20% of the train set was taken with stratified splitting to make a development / validation set. A backtranslation of the data was also made, meaning the data was first translated to English, then Finnish, and back to English. In this thesis I use the model introduced in the aforementioned paper which was trained on the Finnish translated version of this data. I also use the full Jigsaw data to train new models in conjunction with the other translated datasets introduced in the following Chapter 3.2.

3.2 Other Translated Datasets

To address generalization and cross-corpus exploration, many different toxicity and related task datasets were translated, of which nine were chosen for experiments and are introduced below. The datasets were chosen to include many different sources and a few different languages which were available in the DeepL translation service. As a basis for finding and getting the datasets, the toxic comment collection paper and code by Risch, Schmidt, and Krestel [31] was used. All datasets except the Jigsaw dataset introduced above are multiclass where a text can only have one label. As can be seen in Table 3.3 and as mentioned in Chapter 2.4, most of the datasets are sourced from Twitter and are originally in English. The translated datasets are available in GitHub[1].

Regarding tweets, some problems were already introduced in Chapter 2.4.3 and for the datasets in this thesis, this means that many of them are not freely available anymore, as academic access has been limited and Twitter's terms of service regarding content redistribution prohibited keeping (non-anonymized) tweets for long periods of time. Many of the toxicity datasets had to be downloaded through their API through only the use of ids [31] and thankfully the data for the purposes of this thesis was taken for translation before this change.

Twitter Dataset by Davidson, Warmsley, Macy, *et al.* [5] is a hate speech dataset, that tries to combat the problem of including general offensive language in the definition. The data was gathered from Twitter using the API to search for tweets containing words from the lexicon in Hatebase.org. The data is annotated into three categories: hate speech, offensive or neither. The dataset includes 76%

[1] https://github.com/anniesk/masters-thesis

Dataset	Source	Lang	Size	Label(s)
Jigsaw Toxicity [36]	Wikipedia	en	223k	toxicity, severe toxicity, obscene, threat, insult, identity attack
Davidson et. al [5]	Twitter	en	25k	hate speech, offensive
Ousidhoum et. al [22]	Twitter	fr, en	9k	abusive, hateful, offensive, disrespectful, fearful
Qian et. al [27]	Reddit	en	22k	hate speech
Waseem and Hovy [41]	Twitter	en	16k	racism, sexism
Zampieri et. al [45]	Twitter	en	13k	offensive
Founta et. al [9]	Twitter	en	80k	abusive, hate, spam
Bretschneider and Peters [2]	LOL[2] Forum	en	17k	offense (many labels)
Bretschneider and Peters [2]	WOW[3] Forum	en	17k	offense (many labels)
Novak	Twitter	sl	60k	inappropriate, offensive, violent

Table 3.3: General information about all the datasets.

of examples labeled as offensive, 5% labeled as hate and 16.6% labeled as neither meaning the dataset is imbalanced and skewed towards offensive comments. The results of their best performing model on the dataset with 5-fold cross-validation is shown in Table 3.4. Most of the misclassifications are suggested to be from the model classifying tweets as less hateful or offensive than what they really are according to the human annotators.

Twitter dataset by Ousidhoum, Lin, Zhang, *et al.* [22] is also about hate speech. The dataset included subsets in English, French and Arabic but only English and French were taken for this thesis. The dataset includes many different labels: the directness label, meaning whether the target is explicit or not in the tweet, the type of hostility from a set of labels (abusive, hateful, offensive or disrespectful, fearful, normal), the target attribute which means "whether the tweet insults or discriminates against people based on their (1) origin, (2) religious affiliation, (3)

[2]League Of Legends

[3]World of Warcraft

Dataset	Precision	Recall	F1	ROC AUC
Jigsaw Toxicity [36]	-	-	-	0.99[4]
Davidson et. al [5]	0.91	0.90	0.90	-
Ousidhoum et. al [22]	-	-	0.31 (Macro) 0.54 (Micro)	-
Qian et. al [27]	-	-	0.78	0.92
Waseem and Hovy [41]	0.74	0.73	0.78	-
Zampieri et. al [45]	0.82	0.82	0.81	-
Founta et. al [9]	-	-	-	-
Bretschneider and Peters [2] LOL	0.74	0.60	0.66	-
Bretschneider and Peters [2] WOW	0.59	0.52	0.55	-
Novak [16]	-	-	-	-

Table 3.4: Results on the datasets from previous work when available.

gender, (4) sexual orientation, (5) special needs or (6) other." [22], the target group of 16 common target groups tagged by annotators and finally the sentiment of the annotator. Because there are many different labels there are also many classification tasks that can be done on this dataset but for the purposes of this thesis, the type of hostility is the point of interest as a multilabel task. The experimental results are shown in Table 3.4 and the scores are clearly very low. The average Krippendorff scores for IAA were only 0.153 and 0.244 for English and French respectively [22] which can explain why the model scores are so low.

A Reddit dataset by Qian, Bethke, Liu, *et al.* [27] was collected from known toxic subreddits by using hate keywords "to identify potentially hateful comments and then reconstructed the conversational context of each comment". These were then labeled for hate speech by crowd-sourced workers and about 23% of the comments were labeled as hate speech meaning the dataset is imbalanced like many other toxicity related datasets. The results for the binary hate speech detection task is in Table 3.4.

[4]From leaderboard https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/leaderboard

Waseem and Hovy [41] built a Twitter dataset for hate speech detection by first making a list of common slurs and terms used for minorities. Twitter API was then used to find tweets that contained these words, be they offensive or non-offensive. The authors of the paper manually annotated this dataset themselves, after which the annotations were reviewed by a woman who studied gender studies. The texts were labeled as having racism, sexism, or none. Results on the dataset are included in Table 3.4.

OLID by Zampieri, Malmasi, Nakov, *et al.* [45] is a dataset for offensive language detection and includes three subtasks for offensive language classification: offensive language detection, categorization and target identification. The posts for OLID were gathered from Twitter using keywords often present in offensive messages and the dataset includes around 30% of offensive posts. The annotations were done by crowdsourced human-raters and the annotations include three levels of hierarchy for modelling offensiveness: offensive or not, its type (targeted insult/threat or not), and target (individual, group, other). For the thesis the offensive or not is the pointo of interest. Results on the datasets are shown in Table 3.4.

A twitter dataset by Founta, Djouvas, Chatzakou, *et al.* [9] presented a random set of tweets collected in 2017. During the making of this dataset different labeling schemes were explored: first they had seven different labels but after exploration, the tweets were annotated by crowdsourced human-raters to have one of four labels: abusive, hateful, normal or spam making the task multiclass. The dataset includes boosted sampling so that minority classes were properly represented in the data instead of most of the examples belonging to the "normal" label. Regarding agreement, more than half of the tweets achieve overwhelming agreement where "at least 4 out of 5 annotators agreed on the label" [9] and around 36% reach an agreement of three out of five and few achieve majority with only two annotators. This means that agreement was strong and annotation results are robust. The paper only included the making of the dataset and no results on any trained model.

Bretschneider and Peters [2] introduced two datasets in their paper. Comments were taken from two gaming forums, League of Legends (LOL) forum and World of Warcraft (WOW) forum. The datasets were labeled for 20 topics which were preselected from noswearing.com and the datasets can be used for online harassment classification or cyberbully and victim classification which together help detect cyberbullying. For my purposes, the online harassment classification task is the point of interest as it includes labels that can be turned to "toxic" or "non-toxic". IAA was measured only on examples that were judged as online harassment and for the LOL dataset the Fleiss' Kappa value was 0.72 which indicated substantial agreement and for the WOW dataset it was 0.51 which indicated moderate agreement. Results of the datasets for the relevant task are shown in Table 3.4.

A Slovenian Twitter dataset by Kralj Novak, Mozetič, and Ljubešić [16] is a hand-labeled dataset that has 50,000 tweets for training and 10,000 tweets for evaluation. The labeling includes the hate speech type: appropriate, inappropriate, offensive and violent, and hate speech target which has 12 different possible labels. The training set includes tweets from a sample collected between the end of 2017 and beginning of 2020 whereas the evaluation set is a sample from data collected between beginning of 2020 and late 2020. The training set sampling was intentionally biased to contain as much hate speech as possible whereas the test set is a random unbiased sample. The IAA with Krippendorff Alpha is around 0.6 for the dataset. No experimental results on the dataset are available with the original release of the dataset.

3.3 Manually Annotated Finnish Datasets

In this section two Finnish datasets are introduced. One which was introduced in Eskelinen, Silvala, Ginter, *et al.* [8] and another which was made for the purposes of the experiments in this thesis.

3.3.1 Original Dataset

A manually annotated Finnish dataset consisting of 2,260 examples to be used as a test set, was introduced in Eskelinen, Silvala, Ginter, *et al.* [8]. The data was sampled from Suomi24, a Finnish discussion forum. The sampling was done by first classifying 945,867 comments taken from Suomi24 with a model that was at the time their best fine-tuned model, and then taking comments from 10 probability brackets (0.0-0.1, 0.1-0.2, ..., 0.9-1.0) for the six labels to "ensure a representative set of comments featuring varying degrees of toxicity and the six toxicity classes" [8]. The classifier was extremely certain about most of its predictions, so the distribution of labels was not uniform and most of the comments were in the 0.0-0.1 bin and there were high peaks with higher probabilities as well. 50 comments were taken from each bin for annotation meaning for each of the six labels there were 500 comments with "broadly varying degrees of predicted toxicity" [8]. The comments were annotated for only the label from which label bin they were taken, meaning the task was turned into multiclass instead of multilabel like the Jigsaw dataset [36] which worked as the basis for the annotations in terms of labels and definitions of those labels. This also unfortunately means that a comment annotated could have some other type of toxicity. This I will talk about more below with the new Finnish dataset.

The annotations were done by three native Finnish speakers and ambiguous borderline cases were jointly resolved and documented which resulted in general guidelines for the labels which can be found with the data [38]. The full process included the initial annotation process of 100-200 comments which used the same definitions of labels as introduced in Chapter 3.1, after which a discussion was had where some specifications were added to the guidelines. The rest of the comments were then annotated according to those guidelines.

Label	**Initial**	**After discussion**
Toxicity	58%	54%
Severe Toxicity	63%	66%
Threat	82%	80.3%
Obscene	69%	62%
Insult	47.5%	49.6%
Identity Attack	54.5%	66.6%
Mean	62.3%	63%

Table 3.5: Unanimous inter-annotator agreement (IAA) for the native Finnish toxicity dataset [8].

	Label	**No label**
Toxicity	158	193
Severe Toxicity	25	328
Threat	40	391
Obscene	170	239
Insult	145	219
Identity Attack	131	221
Total	669	1591

Table 3.6: Label distribution in the native Finnish annotations according to the label for which a text was annotated for [8].

The inter-annotator agreement (IAA) for the initial annotation and the annotations that were done after discussions can be found in Table 3.5. The unanimous agreement was very low for most of the classes, which is very common for toxicity datasets as mentioned in Chapter 2, and unfortunately even after discussion the scores remained low which highlights the difficulty of the task with the mean IAA being only around 63%. The final dataset only included comments that were either already initially unanimously labeled or comments that were resolved in a later discussion. By only taking the unanimously labeled comments, the validity of the dataset was insured. As mentioned, 2,260 comment comments were in the final dataset and the label distribution is described in Table 3.6.

Original label	**Toxic**	**Non-toxic**
Toxicity	41	4
Not-Toxicity	0	55
Severe Toxicity	8	0
Not-Severe Toxicity	58	34
Threat	12	0
Not-Threat	43	45
Obscene	19	23
Not-Obscene	24	34
Insult	42	0
Not-Insult	7	51
Identity Attack	27	0
Not-Identity Attack	20	53
Total	301	299

Table 3.7: Distribution of the re-annotated Finnish toxicity Suomi24 dataset.

3.3.2 Re-annotated Dataset

For the purposes of this thesis, of the 2,260 examples in the dataset introduced above, 600 were re-annotated with only the generic toxicity definition given in Section 2.3 as a guideline. The 600 examples were taken by randomly sampling 100 from each of the six classes meaning there is a random amount of that type of toxicity. This is due to the fact that some of the labels had significantly less labeled as the type of toxicity but for all of the classes there was mostly no toxicity. This re-annotation was done because the dataset was not annotated for all of the six toxicity classes (technically 12 classes as for each a "not" class is present) in the original annotation process and thus some types of toxicity were definitely missed and the dataset was not fully ready for mapping to a binary task. Unfortunately, the re-annotation was only done by me, so the decisions are subjective and someone else might have different interpretations of the same data meaning the quality is not confirmed as it was with the original annotations. This might introduce some sort of bias.

The distribution of the labels after the re-annotation was 299 for the "non-toxic" label and 301 for the "toxic" label which is surprisingly balanced. The more specific distribution according to original labels can be seen in Table 3.7. The data is available in GitHub[5]. During the annotation process, although I mostly tried to use the general toxicity definition, the guidelines made during the original annotation process regarding political content, sexual content, self-harm, and positive swear-words were helpful. Mostly labeling went as expected but the obscene label ended up having 23 comments that were obscene but not toxic. It must be noted that there was in my opinion too much sensitivity to swearwords in the original dataset, meaning not all comments that were marked as obscene ended up being given the toxic label. The same in a smaller scale happened with the general toxicity label, where four comments were in my opinion unjustly marked as there was something really minor which can be up to interpretation.

3.4 Unifying Datasets

One solution to the problem of datasets having different labels and thus being incompatible with each other out of the box, was introduced in Risch, Schmidt, and Krestel [31]. Their goal was to unify the labeling of toxicity and similar tasks' datasets by making a mapping to 126 different class labels (out of original 162) for over 40 different datasets. These labels can then be further mapped to a binary view to have the class labels **toxic** and **non-toxic** by following the instructions given in the paper. Risch, Schmidt, and Krestel [31] also introduced an easy way to download these over 40 datasets in their repository[6], which was used to get the datasets wanted for translation.

I mapped the labels of the small datasets with the following strategy based on the strategy introduced in Risch, Schmidt, and Krestel [31]: if a label is missing the example is mapped as non-toxic, if a label is either "none", "normal", "other", "positive" or "appropriate" the new label is non-toxic, and if a label is anything

[5]https://github.com/anniesk/masters-thesis

[6]https://github.com/julian-risch/toxic-comment-collection

Dataset	**Non-toxic**	**Toxic**	**Total**
Jigsaw [37]	201,081	22,468	223,549
Davidson et. al [5]	4,163	20,620	24,783
Ousidhoum et. al [22]	1,357	8,304	9,661
Qian et. al [27]	17,062	5,255	22,317
Waseem and Hovy [41]	7,393	2,663	10,056
Zampieri et. al [45]	8,840	4,400	13,240
Founta et. al [9]	33,191	18,263	51,454
Bretschneider and Peters LOL	1,739	84	1,823
Bretschneider and Peters WOW	1,115	40	1,155
Novak	52,606	25,928	78,534
Total	328,547	108,025	436,572

Table 3.8: Amount of toxic or non-toxic examples in the dataset after translation and unified mapping.

other than empty or the ones listed above, the text is mapped as toxic. If an example has the label "idk/skip" that example is discarded. In the mapping process I additionally got rid of examples where the text was empty for some reason, be it because of a fault with the original dataset or because something happened to it during the translation process. For the Jigsaw dataset, the full dataset including train and test is used together, and if an example has any of the six labels the example is mapped as "toxic" or if the example has no labels it is mapped as "non-toxic". This same principle is later also used with the existing toxicity model when testing with new datasets as it also uses the same six labels.

Note that due to the process explained above, the amount of examples may differ for the datasets when comparing Tables 3.3 and 3.8. The size of the datasets and the label distribution after translating and unifying the datasets is shown in Table 3.8 and in Figure 3.2. The table and figure show that there are differences between the datasets in how the labels distribute. For the "ultimate" model where I fine-tune using all the translated datasets used in this thesis the distribution seems to be a bit better in the sense that there is almost 25% of toxicity, while in the training of the existing model with the Jigsaw dataset [36], there was only 11% of toxicity. The datasets by Davidson, Warmsley, Macy, *et al.* [5] and Ousidhoum, Lin, Zhang, *et al.*

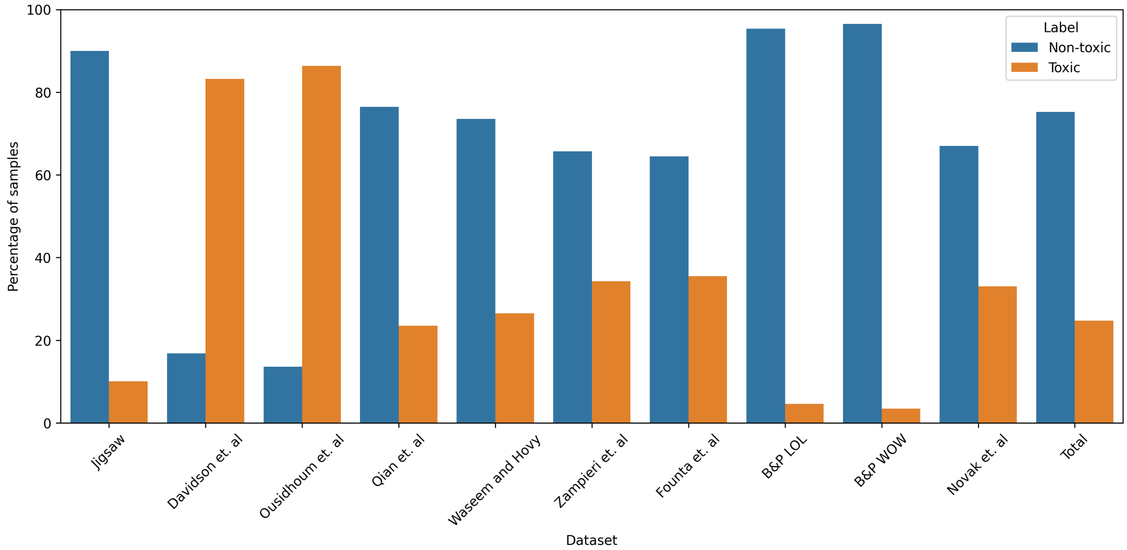

Figure 3.2: Distribution of the labels in the datasets as a bar plot of percentages after translation and unified mapping.

[22] contain a lot of toxicity compared to the other datasets because of the collection strategy. The full unified dataset contains 436,122 examples, of which 223,549 are Wikipedia editor comments, 187,278 are tweets from Twitter, and 25,115 are from other sources. This means that the Jigsaw dataset and the Twitter datasets clearly dominate the full dataset.

Three examples of the original texts and their translations with the original label and new label are displayed in Table 3.9. The translations are seemingly generally good, but clearly translated to written language and not colloquial language that is spoken and often seen in social media contexts. For the LOL dataset, it can be seen that HTML and links are lost during translation, which coincidentally works as a supplementary preprocessing step, but no other preprocessing for the texts is done before tokenization and training on the dataset(s).

Dataset	Original text and Translation	Original Label(s)	New Label
Jigsaw	It's not about citations within an article, but citations of an article. Sure, it's mainly Semantic Web people who currently talk about DBpedia, but it's mainly people interested in anime who talk about Tokyo Mew Mew. I don't understand your line of argument there at all. Kyse ei ole artikkelin sisäisistä viittauksista vaan artikkelin viittauksista. Toki DBpediasta puhuvat tällä hetkellä lähinnä semanttisen webin ihmiset, mutta Tokyo Mew Mewista puhuvat lähinnä animen parissa työskentelevät ihmiset. En ymmärrä lainkaan argumenttiasi.	No label	Non-toxic
Jigsaw	Singing is one thing, but you also need to show that you're a smart person and not some illiterate redneck from the south." Laulaminen on yksi asia, mutta sinun on myös näytettävä, että olet fiksu ihminen etkä mikään lukutaidoton punaniska etelästä.	Identity Attack, Insult, Toxicity	Toxic
B&P LOL	Welcome to F2P game guys. Enjoy your free rooms, free meals, and free lag Tervetuloa F2P-peliin. Nauttikaa ilmaisista huoneista, ilmaisista aterioista ja ilmaisesta viiveestä.	No label	Non-toxic

Table 3.9: Examples of texts and their translations from the datasets.

4 Results

In this Chapter the experiments with their results will be explained and the safety, usefulness and usage of the model will be discussed with examples of predictions where the best model makes mistakes and examples of what the model sees in texts before making a prediction.

4.1 Results with the Existing Fine-tuned Model

In this Chapter I present results of using the existing model by Eskelinen, Silvala, Ginter, *et al.* [8] on the new translated datasets, the translated Jigsaw dataset, the manually annotated Finnish dataset as well as the new the re-annotated Finnish dataset which is based on a sample from the manually annotated Finnish dataset.

4.1.1 Results on the Machine Translated Small Datasets

To get an answer to the research question regarding whether a model trained on one type of corpus works on a different corpus I test the existing toxicity model [8] on the small translated datasets introduced in Section 3.2. I compare the results to those reported in the papers of the original datasets when available. Unfortunately, the comparison is not fully equal as the translation and the unified mapping to binary (non-toxic, toxic) might skew the results which is why this research question is also the topic of the next Chapter 4.2.

Dataset	**Accuracy**	**Precision**	**Recall**	**F1**
Davidson et. al [5]	0.85	0.75	0.85	0.78
Ousidhoum et. al [22]	0.53	0.54	0.58	0.47
Qian et. al [27]	0.77	0.71	0.77	0.72
Waseem and Hovy [41]	0.71	0.58	0.56	0.56
Zampieri et. al [45]	0.78	0.78	0.70	0.72
Founta et. al [9]	0.78	0.81	0.70	0.72
Bretschneider and Peters LOL [2]	0.90	0.63	0.83	0.68
Bretschneider and Peters WOW [2]	0.85	0.56	0.75	0.58
Novak [16]	0.71	0.67	0.60	0.59

Table 4.1: Results on the small datasets with the existing fine-tuned model (macro average).

Because the existing fine-tuned model is multilabel, I needed to map the predictions it gives to binary as well. To do this I had to decide on a threshold for the labels, where everything above the threshold is considered as toxic and below as non-toxic. This means that if any of the six labels have a probability greater than the threshold, the text is considered toxic. I decided on the simple threshold of 0.5 for the sake of simplicity in the experiments, but it is possible to use thresholds such as 0.4 or 0.6 as well. As metrics, macro average was used for precision, recall and F1. ROC AUC is not used for this as it is impossible to get the probabilities from the existing model because of the mapping of the labels.

Original results on the used datasets were introduced in Table 3.4 when available. Results with the existing model are shown in Table 4.1. Comparison of these two tables shows that results on the Davidson et. al [5], Zampieri et. al [45] and Qian et. al [27] datasets are slighly lower with the existing Finnish model than those reported in the original papers. Results on the Waseem & Hovy dataset [41] are much worse in terms of both precision and recall, which also brings the F1-score down. On the other hand, results on both of the Bretschneider and Peters datasets [2] are slightly better with the existing Finnish model than those reported in the original paper.

Worse results could be explained by the mapping of the labels to binary as some information might be lost, the different domain of the texts as the existing model was

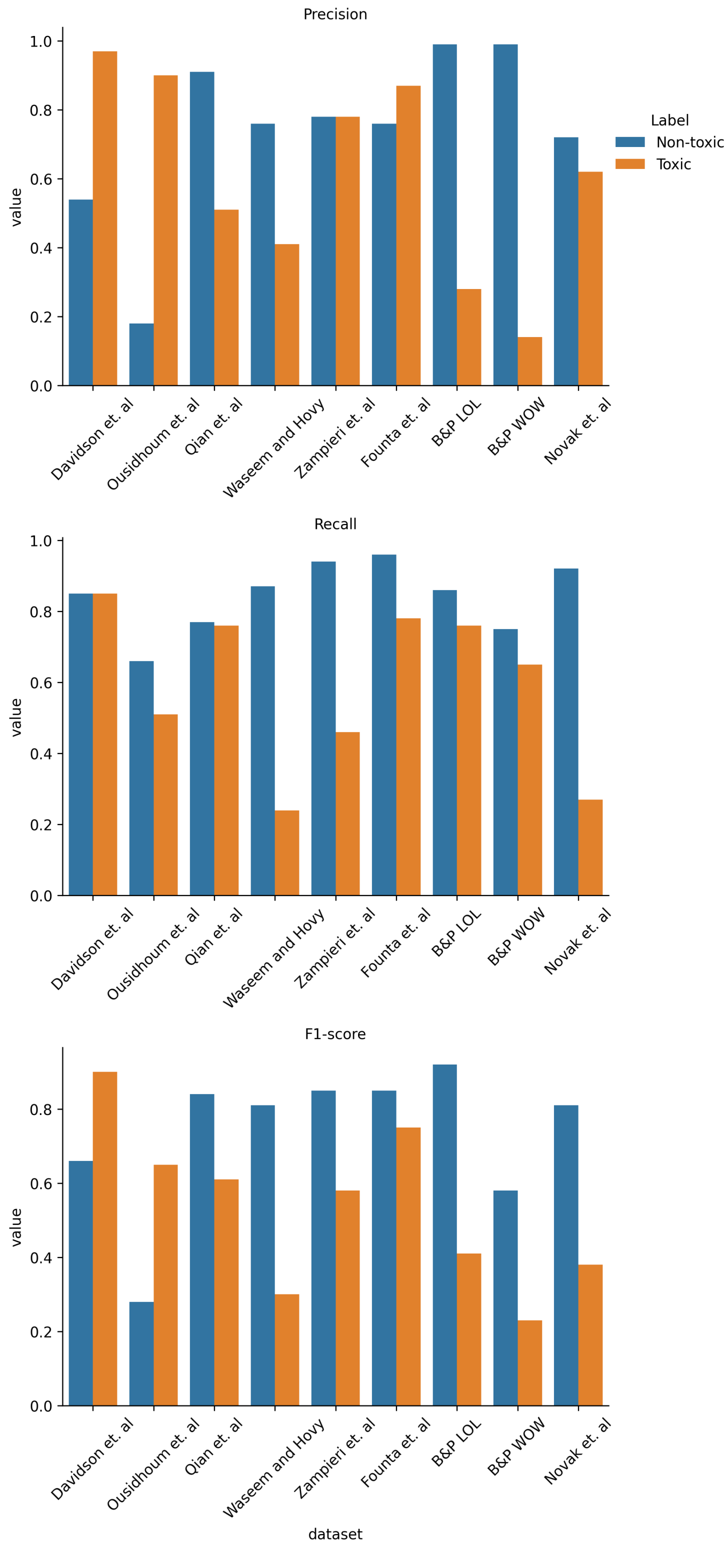

Figure 4.1: Class-wise metrics for the small datasets with the existing toxicity model.

trained with the translated Jigsaw dataset [37] which might have different language than in other social media domains, or some toxicity might have been lost in the translation process.

Precision, recall and F1-score can vary between the two labels, and they are visualized in Figure 4.1. Some datasets have better results for the toxic label and others have better results for the non-toxic label and this seems to reasonably correspond to the distribution of labels in the datasets shown in Table 3.2, i.e. how many toxic examples there are compared to non-toxic.

4.1.2 Results on the Translated Jigsaw and Both Finnish Manually Annotated Datasets

To answer the research question about how a corpus from another culture works in the Finnish context I discuss the test set results of the translated Jigsaw test set and both the manually annotated Finnish test set and the re-annotated sample. The results are gained with the existing toxicity model fine-tuned on the translated version of the Jigsaw dataset [37], based on Wikipedia editor comments.

For the original manually annotated Finnish dataset, a modified prediction script was made [8] to only look at the prediction competence on the one label it was annotated for. The script mapped the probability and the prediction of all other labels as 0 except the one annotated to match those of the true label. For the ROC AUC metric, in the paper [8] the results were mistakenly given with the thresholded labels instead of probabilities which resulted in slightly worse numbers. The updated number is shown in Table 4.2.

The results of testing the existing model on the different test sets are introduced in Table 4.2. The model gets the best performance in terms of F1-score (micro, takes label imbalance into account) on the new re-annotated test set. This could be due to the fact that perhaps the binary task is easier or because the new annotations were

Test Set	**Acc.**	**Precision**	**Recall**	**F1**	**ROC AUC**
Jigsaw DeepL (6 labels)	0.87	0.52 (micro) 0.47 (macro)	0.82 (micro) 0.73 (macro)	0.63 (micro) 0.56 (macro)	0.97 (micro)
Jigsaw DeepL (binarized, 2 labels)	0.91	0.76 (macro)	0.89 (macro)	0.81 (macro)	-
Annotated (1 label)	0.75	0.58 (micro) 0.51 (macro)	0.58 (micro) 0.49 (macro)	0.58 (micro) 0.48 (macro)	0.97 (macro)
Re-annotated (2 labels)	0.64	0.70 (macro)	0.64 (micro)	0.61 (macro)	-

Table 4.2: Results with the existing fine-tuned model on the translated Jigsaw test and the manually annotated test sets.

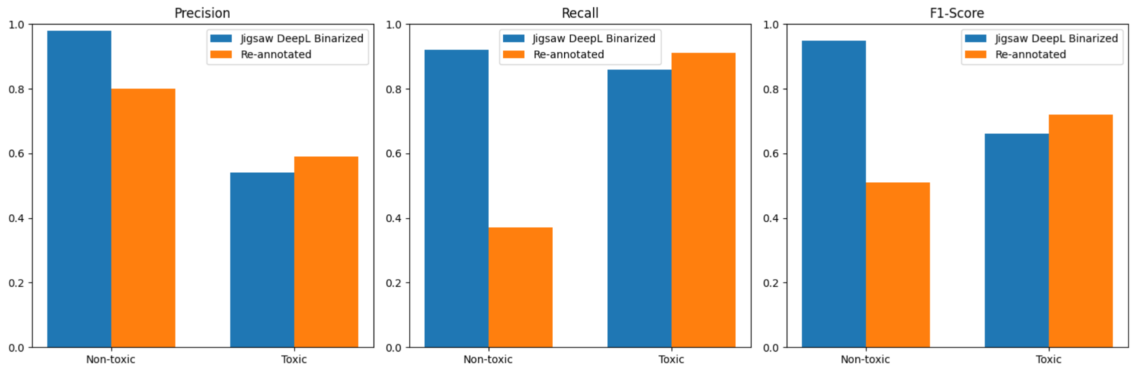

Figure 4.2: Class-wise metrics for the Jigsaw DeepL binarized and the re-annotated dataset with the existing toxicity model.

less sensitive as mentioned in Chapter 3.3.2. When binarizing the Jigsaw toxicity test set with the method specified in Chapter 3.4 and turning the predictions to binary as we do for the re-annotated dataset, the results in terms of all metrics are higher than those of the multi-label Jigsaw test set. This reiterates the possibility that the binary task might be slightly easier than the multi-label task.

Precision and recall are balanced only for the original manually annotated Finnish dataset and they are extremely imbalanced especially with the translated Jigsaw dataset with the original labels [37]. For the translated Jigsaw, the ROC_AUC is slightly lower than what was the highest result on the Kaggle competition on a subset of the original English dataset 3.4. For the two labels, "toxic" and "non-toxic", Figure 4.2 shows the results on the Jigsaw DeepL binarized version and the re-annotated dataset. The toxic label shows similar performance on both models, but for the non-toxic label, the results on the re-annotated dataset are worse than on the binarized Jigsaw. For binarized Jigsaw, performance on the non-toxic label is similar across all metrics, but for the toxic label recall is higher than precision which results in the F1-score being higher for the non-toxic label. The differences on the results of the datasets for the two labels is possibly due to the distribution of the labels in training data and how it matches the test data, as the distribution for the Jigsaw is similar to the training data with significantly more non-toxic examples whereas the re-annotated dataset has a balanced distribution between the two labels.

4.2 Results with the New Models

To answer the research questions about how a corpus from another culture works in the Finnish context and if a model trained on different corpora works on another corpus, I do more experiments and use the nine small datasets and the bigger Jigsaw dataset introduced in Chapter 3 to test how the different datasets from different social media platforms, with different labels and annotation schemes can work together by mapping the labels to toxic or non-toxic. I use these datasets to train new models by using all but one of the datasets, changing the one left out each time, and then compare results on the left-out translated dataset and on the re-annotated Finnish dataset. I also do experiments by training only on Twitter data or no Twitter data as Twitter corpora dominate the translated datasets.

The intentions with training many models is to see whether the Jigsaw dataset or Twitter datasets override everything as they dominate the datasets, to see if the corpora are compatible after the mapping or if they are too dissimilar making the model confused with way worse results, and to see how much numbers fluctuate between different combinations of the datasets and which datasets are more compatible than others. I also train a model based on all of the data and test on the re-annotated manually annotated Finnish data introduced in Chapter 3.3 to make one ultimate model for toxicity prediction in Finnish. A development set of 20% of the comments is taken from the full data after shuffling with a fixed seed (42) so that at least in theory both the train and development sets contain examples from all the current datasets in use, except of course the test set which remains the same.

In this thesis the cased base version of FinBERT is fine-tuned instead of the large version which was the best performing model in the paper by Eskelinen, Silvala, Ginter, *et al.* [8]. This change is due to limited resources and the fact that there is a lot of data to train on. As a bonus, running inference with the fine-tuned base model is faster which brings benefits in the long run if the best model is used. For the training of the ultimate model with all of the data I used the following hyperparameters: Learning rate 5e-5, 3e-5, 2e-5, 1e-5, batch size 4, 8, 12, sequence length 512 and training for 5 epochs (not 10 because there is so much data and training would take forever) with early stopping and evaluation every 25,000 steps. For predictions, to keep the experiments simple, I take the label with highest probability as the probabilities sum to 1 after softmax. In the binary case this means that one label always has a probability equal to or greater than 0.5.

The best model according to the development set for every hyperparameter is saved and of all the different models, the best model is chosen to be tested on the re-annotated data. For training the models with varying datasets, I used the hyperparameters that gave the best development set results for the ultimate model. The best hyperparameters according to the results gained on the development set of the ultimate model trained on all of the data were: learning rate 1e-05, batch size 8, sequence length 512, 5 epochs with early stopping and evaluation every 25,000 steps. Training of the models was done with a single GPU on the supercomputer Puhti provided by CSC - IT Centre for Science.

Training Data	**Test data**	**Accuracy**	**Precision**	**Recall**	**F1**
Excluding Jigsaw [37]	Jigsaw [37]	0.87	0.69	0.82	0.73
Excluding Davidson et. al [5]	Davidson et. al [5]	0.84	0.73	0.81	0.76
Excluding Ousidhoum et. al [22]	Ousidhoum et. al [22]	0.61	0.54	0.58	0.51
Excluding Qian et. al [27]	Qian et. al [27]	0.76	0.71	0.77	0.72
Excluding Waseem and Hovy [41]	Waseem and Hovy [41]	0.71	0.62	0.61	0.61
Excluding Zampieri et. al [45]	Zampieri et. al [45]	0.71	0.69	0.68	0.69
Excluding Founta et. al [9]	Founta et. al [9]	0.74	0.76	0.66	0.67
Excluding Bretschneider and Peters LOL [2]	Bretschneider and Peters LOL [2]	0.91	0.63	0.75	0.66
Excluding Bretschneider and Peters WOW [2]	Bretschneider and Peters WOW [2]	0.86	0.56	0.73	0.58
Excluding Novak [16]	Novak [16]	0.69	0.64	0.60	0.61

Table 4.3: Results of changing datasets tested on the left-out dataset (macro average).

4.2.1 Results on the Machine Translated Datasets

As introduced above, 10 models were trained with changing the dataset that is left-out and then tested with the left-out dataset. Results on changing datasets tested on the left-out dataset are reported in Table 4.3. These results can be compared to the results on the small datasets reported earlier in Table 4.1 and to the results for the Jigsaw dataset reported in Table 4.2.

The results when evaluating on the full Jigsaw dataset (including the train and test set) [37] are better with the unified dataset which excluded the dataset from training when compared to the results with the existing model tested on just the

Jigsaw test set. The performance increase is due to the precision score rising from 0.52 to 0.69 which increases the F1-score as well. The test set and the full dataset have a similar distribution in toxic and non-toxic examples meaning that performance gain from that is not expected in this comparison. The performance increase is instead most likely due to a more diverse training dataset but the mapping of the labels to a binary task could possibly also help, as the results on the Jigsaw test set were also better with the binarized labels.

Results on the small datasets are varied. Some datasets get better results whereas some are doing slightly worse and a few get similar results compared to the existing model. Specifically, the datasets by Davidson et. al [5], and the LOL dataset by Bretschneider and Peters [2] do slightly worse and the datasets by Zampieri et. al [45] and Founta et. al [9] do significantly worse. The reason could be that these datasets have very different language and they need examples from their own dataset. On the other hand, the datasets by Ousidhoum et. al [22] and Waseem and Hovy [41] get better results and actually benefit from the varying datasets, and the datasets by Qian et. al [27], Novak [16] and the WOW dataset [2] get similar performance. None of the datasets that perform better or get similar performance need examples from their own datasets meaning they are highly compatible with other datasets.

The precision, recall and F1-score are visualized in Figure 4.3. The results are similar to those of the existing model reported in Table 4.1. There are only a few notable differences, specifically with the datasets by Zampieri et. al [45] and Founta et. al [9] as their recall for the toxic label is worse but for the LOL and WOW datasets [2] the recall is better. Regarding F1-score, WOW gets better results for the non-toxic label. Precision remains seemingly similar for all datasets.

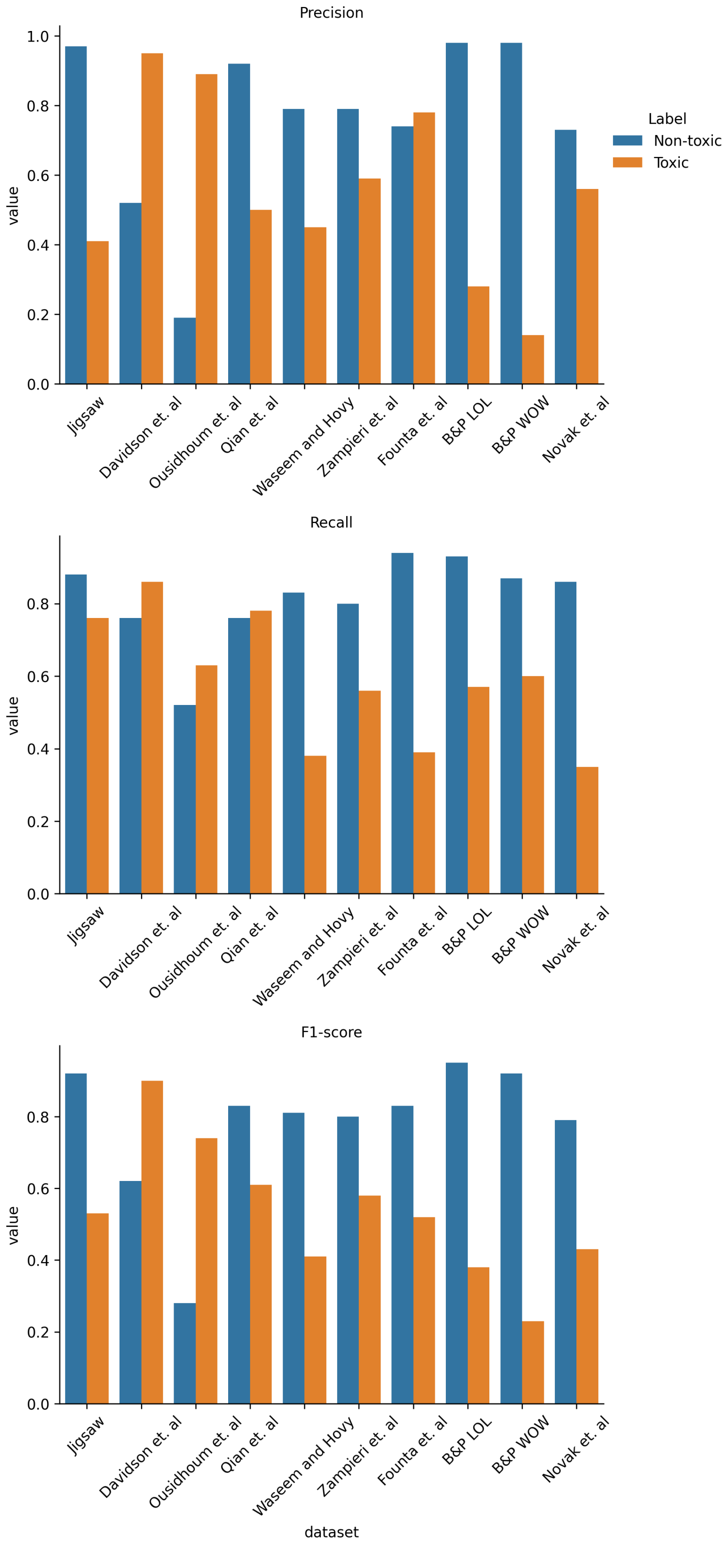

Figure 4.3: Class-wise metrics on the small datasets with the new models which exclude the test set from training data.

Training Data	Accuracy	Precision	Recall	F1
Excluding Jigsaw [37]	0.69	0.71	0.69	0.69
Excluding Davidson et. al [5]	**0.71**	0.70	**0.73**	**0.71**
Excluding Ousidhoum et. al [22]	0.68	0.70	0.68	0.67
Excluding Qian et. al [27]	0.68	0.71	0.68	0.67
Excluding Waseem and Hovy [41]	0.70	0.72	0.70	0.70
Excluding Zampieri et. al [45]	0.70	0.72	0.70	0.69
Excluding Founta et. al [9]	**0.71**	**0.73**	**0.71**	**0.71**
Excluding Bretschneider and Peters LOL [2]	0.69	0.72	0.69	0.68
Excluding Bretschneider and Peters WOW [2]	0.70	0.73	0.70	0.69
Excluding Novak [16]	0.68	0.71	0.68	0.67
No Twitter Data	0.70	0.72	0.70	0.70
Only Twitter Data	0.69	0.72	0.69	0.67
All	**0.71**	**0.73**	**0.71**	**0.71**

Table 4.4: Results of changing datasets tested on the re-annotated Finnish dataset (macro average). Best results bolded.

4.2.2 Results on the Re-Annotated Dataset

As introduced at the beginning of this Chapter and in the previous section, 10 models were trained while changing the small translated dataset that was left-out and then tested on the re-annotated Finnish dataset. Two additional models were also trained due to the fact that many of the smaller datasets were from Twitter, one with only Twitter data and one without any Twitter data. Finally, a model trained on all of the small translated datasets was tested on the re-annotated data resulting in a total of 13 models for comparison.

Results of models trained with varying datasets, and a model trained with all of the datasets and tested on the manually re-annotated Finnish dataset are reported in Table 4.4. There are surprisingly no huge differences in the results between models that had one dataset excluded, or were trained only with or completely without any Twitter data meaning even huge datasets can be excluded and the results are still comparable implying they are not necessarily needed. All results have an F1-score between 0.67 and 0.71. The best results were gained with the model that was trained on all of the data and with excluding the dataset by Founta et. al [9]. Almost as good results were gained by excluding the Davidson et. al dataset [5] but the precision is lower than for the aforementioned models. Because almost all the datasets are necessary to make the best performing model, the datasets seem to mesh together well despite having different gathering techniques, domains, labels and even original languages.

The results for F1-score are better than with the existing model and also better than what was reported for the original manually annotated Finnish dataset [8] and this is due to the fact that the precision of the model is higher. This is most likely due to the fact that both the training dataset and the test set are balanced, with about 50% being toxic and 50% being non-toxic. There are small differences with precision, recall and F1 with the re-annotated test set between all of the different models and the variation of the numbers can be seen in Table 4.5. Despite the fact that the dataset is balanced, there is a disparity with the two labels, with the toxic label having higher recall and lower precision whereas the opposite is true for the non-toxic label. The trade-off between precision and recall will be discussed more below in Chapter 4.3.

Label	**Precision**	**Recall**	**F1-Score**
Non-toxic	0.77 – 0.81	0.52 – 0.58	0.61 – 0.67
Toxic	0.63 – 0.67	0.84 – 0.87	0.73 – 0.75

Table 4.5: Class-wise metrics with the new models (macro average).

Because the best results were gained with two different models, the model with all of the training data was taken for further examination (see Chapter 4.3) and chosen as the best model, since for my purposes, as much data as possible that brings variety to the training data is desired. Meanwhile, the least amount of training data that brings the best results is possibly a dataset that just excludes the dataset by Founta et. al [9], although tests could be made with all combinations to find the best model trained on the least amount of data but unfortunately that is out of scope for this thesis. More training data in my case seems to make a better model, although this might not always be the case if the quality of the data is poor.

4.3 Safety and Usefulness of the Model

In this Chapter I answer the research question about how safe and useful the model is in actual use. I explore the safety of the toxicity model by discussing the precision and recall of the best new model on the manually annotated Finnish dataset and by looking at the misclassifications of the model and what the model sees before making a prediction. For the usefulness part, I discuss how the existing model has been used in other works already and how it and the new model could be used in the future.

4.3.1 The Trade-Off Between Precision and Recall

Precision and recall were introduced as metrics in Chapter 2.5 and the results on different models in general and of the two labels were introduced in Chapter 4. A precision-recall curve for the two labels on the best model was plotted and can be seen in Figure 4.4. As mentioned before, even though the dataset is balanced, there are differences between the labels even though when combining both labels, the model has similar precision and recall with precision being slightly higher. From the curves it can be seen that for the non-toxic label there is a very steep high precision at a very low recall meaning that when the model makes very few positive predictions for the non-toxic label, they are usually correct. On the other hand, when recall increases, precision drops initially and only slightly increases, never reaching very

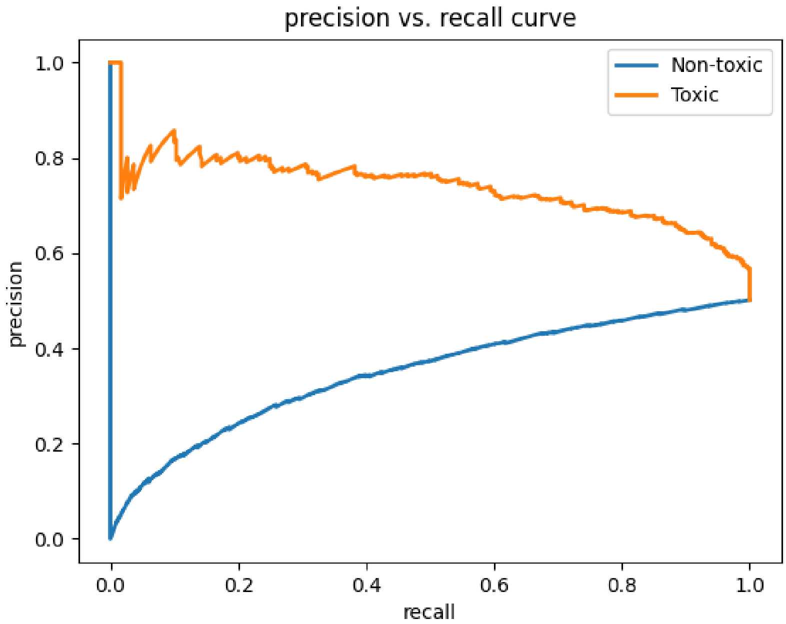

Figure 4.4: Precision-Recall Curve of the best model.

high precision. There is a lower precision level at all recall levels compared to the toxic label.

For the toxic label, the results are much more stable. Based on the class-wise metrics and the precision-recall curve, the model is better at detecting toxic comments than non-toxic ones. The reason for the non-toxic label performing so poorly could be due to the fact that there is now too little non-toxic data and the dataset is too balanced making the model more biased towards toxicity. Another reason could be that as there are different labeling strategies and different ideas on what toxicity is, the signal for what is toxic is not clear to the model.

Changing the threshold of the labels, or even implementing a different threshold for each label, e.g., mapping the toxic label to non-toxic if the probability is less than 0.7, could be helpful. Nevertheless, either precision, recall or F1-score of the model could be maximized by changing the threshold to have a different focus, but generally F1-score is the most important one as it takes the harmonic mean of both precision and recall. From a content moderation perspective for example, higher

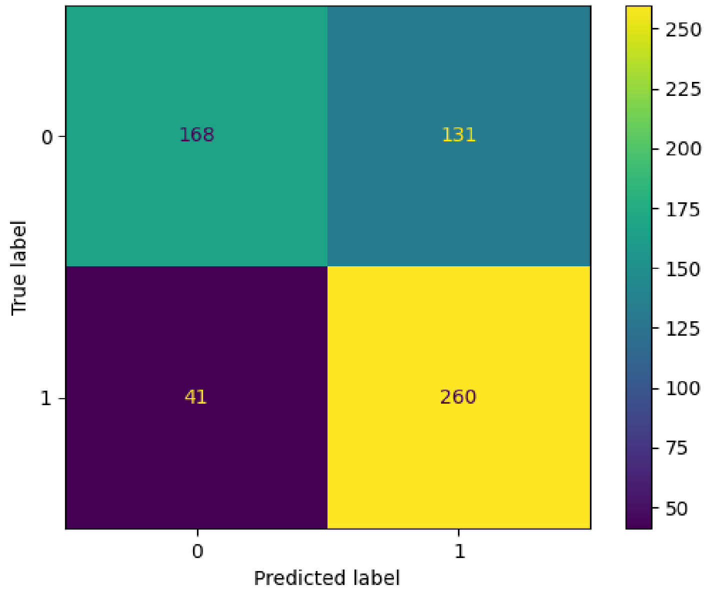

Figure 4.5: Confusion matrix of the unified model trained on all of the data and tested on the re-annotated Finnish dataset.

recall is more important and that is clearly higher for the toxicity label which is a good thing, although some precision is sacrificed.

4.3.2 Misclassifications

In this Chapter misclassifications in the form of a confusion matrix, examples of classifications, and visualizations of what the model sees on the new re-annotated Finnish dataset with the new best unified model are introduced. Model explainability, i.e. what the black box model sees before making predictions was visualized with the integrated gradients method [35] where no modification is made to the original model and the method is simply implemented by making "a few calls to the standard gradient operator".

The confusion matrix in Figure 4.5 shows and reiterates that the model has problems where it predicts a non-toxic text to have toxicity as 131 comments out of 299 are misclassified in that way, whereas the model does not find toxicity in only

41 comments out of the 301 that are annotated as toxic. This means that the model is highly sensitive to toxicity which can either be a good or a bad thing depending on the use case. As mentioned above, if the use case is e.g., content moderation, a high recall for the toxic label is more important than precision.

The toxicity model is generally very certain of its predictions. When looking at the probabilities, only 12 texts of the 600 (2%) get probabilities between 0.5 and 0.8, four of them being close to 0.5. This means that 98% of the texts are predicted with a certainty of more than 0.8, and almost 97% are predicted with a certainty of at least 0.9. Optimization of the threshold for labels was discussed before but as most of the predicted probabilities are so high, the matter of where to put the threshold becomes irrelevant and justifies the simple approach in the thesis of just using the bigger probability to choose the label.

The texts that get probabilities closer to 0.5 are usually thought of as more difficult, and this seems to be true as almost all predictions in the probability range 0.5-0.8 are wrong. A possible reason for this could be that the text contains elements that point to both classes. Three examples are shown in Figure 4.6, and all of the texts have true labels of "non-toxic" with the first two texts being predicted wrong. When looking at the visualizations, the model clearly sees elements that it thinks could point to both classes, hence the uncertainty that is reflected on the probability. Interesting notes about these misclassifications are that "feministi" (Translation: "feminist") is considered toxic by the model, possibly due to it being used as an insult in some contexts. Same thing happens with the word "blond".

Examples of misclassifications can be seen in Table 4.6 of both types of misclassifications, texts that were labeled as toxic but were predicted as non-toxic and texts that were labeled as non-toxic but were predicted as toxic. A visualization of what the model sees in these misclassifications is shown in Figure 4.7. An examination of the examples and their visualization shows that there are a few possible reasons for

```
Prediction: toxic Weights: [0.3194977939128876, 0.3732896149158478]
kuitenkaan mikään feministi oo..
ATTRIBUTION WITH RESPECT TO not-toxic
[CLS] kuitenkaan mikään feministi oo . . [SEP]

kuitenkaan mikään feministi oo..
ATTRIBUTION WITH RESPECT TO toxic
[CLS] kuitenkaan mikään feministi oo . . [SEP]

Prediction: toxic Weights: [0.18053768575191498, 0.26253074407577515]
Asiasta: Blondit ja äly. Kaikki blondit eivät ole kauniita mielestäni. Vain älykkäät blondit ovat kauniita,
äly näkyy silmissä. -Tyhmä "Gro magno" ei ole mielestäni koskaan kaunis.
ATTRIBUTION WITH RESPECT TO not-toxic
[CLS] Asiasta : Blo ##n ##dit ja äly . Kaikki blondi ##t eivät ole kauniita mielestäni . Vain älykkää ##t blondi ##t ovat kauniita , äly näkyy
silmissä . - Tyh ##mä " Gro magn ##o " ei ole mielestäni koskaan kaunis . [SEP]

Asiasta: Blondit ja äly. Kaikki blondit eivät ole kauniita mielestäni. Vain älykkäät blondit ovat kauniita,
äly näkyy silmissä. -Tyhmä "Gro magno" ei ole mielestäni koskaan kaunis.
ATTRIBUTION WITH RESPECT TO toxic
[CLS] Asiasta : Blo ##n ##dit ja äly . Kaikki blondi ##t eivät ole kauniita mielestäni . Vain älykkää ##t blondi ##t ovat kauniita , äly näkyy
silmissä . - Tyh ##mä " Gro magn ##o " ei ole mielestäni koskaan kaunis . [SEP]

Prediction: not-toxic Weights: [0.30260327458381653, 0.055026646703481674]
Pillu, perse, kyrpä/mulkku, tissit/buubsit, pano tms. Kuuluvat sujuvasti sanavalikoimaani myös miesten
kanssa jutellessa. Mitä sitä turhia keksimään uusia sanoja.
ATTRIBUTION WITH RESPECT TO not-toxic
[CLS] Pill ##u , perse , kyr ##pä / mul ##kku , tissit / buu ##bs ##it , pano tms . Kuul ##uvat sujuvasti sana ##valikoima ##ani myös
miesten kanssa jut ##ellessa . Mitä sitä turhia keksimään uusia sanoja . [SEP]

Pillu, perse, kyrpä/mulkku, tissit/buubsit, pano tms. Kuuluvat sujuvasti sanavalikoimaani myös miesten
kanssa jutellessa. Mitä sitä turhia keksimään uusia sanoja.
ATTRIBUTION WITH RESPECT TO toxic
[CLS] Pill ##u , perse , kyr ##pä / mul ##kku , tissit / buu ##bs ##it , pano tms . Kuul ##uvat sujuvasti sana ##valikoima ##ani myös
miesten kanssa jut ##ellessa . Mitä sitä turhia keksimään uusia sanoja . [SEP]
```

Figure 4.6: Visualization of what the model sees in texts predicted with a probability between 0.5-0.6.

Text	**English Translation**	**True Label**	**Pred. Label**	**Probability of Pred. Label**
on tää typerä koira-hommakin mennyt monimutkaiseksi siis jollakin on tai mieluumin ei ole koiraa.Mun puolesta kaikki paitsi sokeitten koirat voi hyvinkin ampua.	this stupid dog thing has gotten complicated too, so someone has a dog, or rather doesn't have a dog.As far as I'm concerned, everyone except blind dogs can be shot.	Toxic	Non-toxic	0.684
Sairasta touhua. Joutas sanonko minne.... hyi helvetti....	That is sick. They should go should I say where.... Holy crap....	Non-toxic	Toxic	0.998
Tää on aika hyvä: Mrs Weasley: NOT MY DAUGHTER, YOU BITCH!	This is pretty good: Mrs. Wealey: NOT MY DAUGHTER, YOU BITCH!	Non-toxic	Toxic	0.999

Table 4.6: Examples of misclassifications on the re-annotated Finnish dataset.

why some texts are misclassified. These include: swears and e.g., demographic mentions, which the model is sensitive to. This might be due to the training datasets, as many were collected based on keywords and can thus be sensitive to those contents or texts in which the model does not understand that something was not said in a malicious way. That is why the last example is an acceptable mistake, but the first maybe not so much.

From the last example, it can also be noted that the text in all caps is tokenized in a weird way. Often no preprocessing is done when fine-tuning transformer models on text classification tasks as it can make performance worse depending on the task [34]. However, some preprocessing to change sentences that are all caps could be done because tokenizing purely on an all caps text might affect the results, as can be seen from the slightly different predicted probability and visualization of the sentences in Figure 4.8. The sentence was taken from part of the re-annotated data

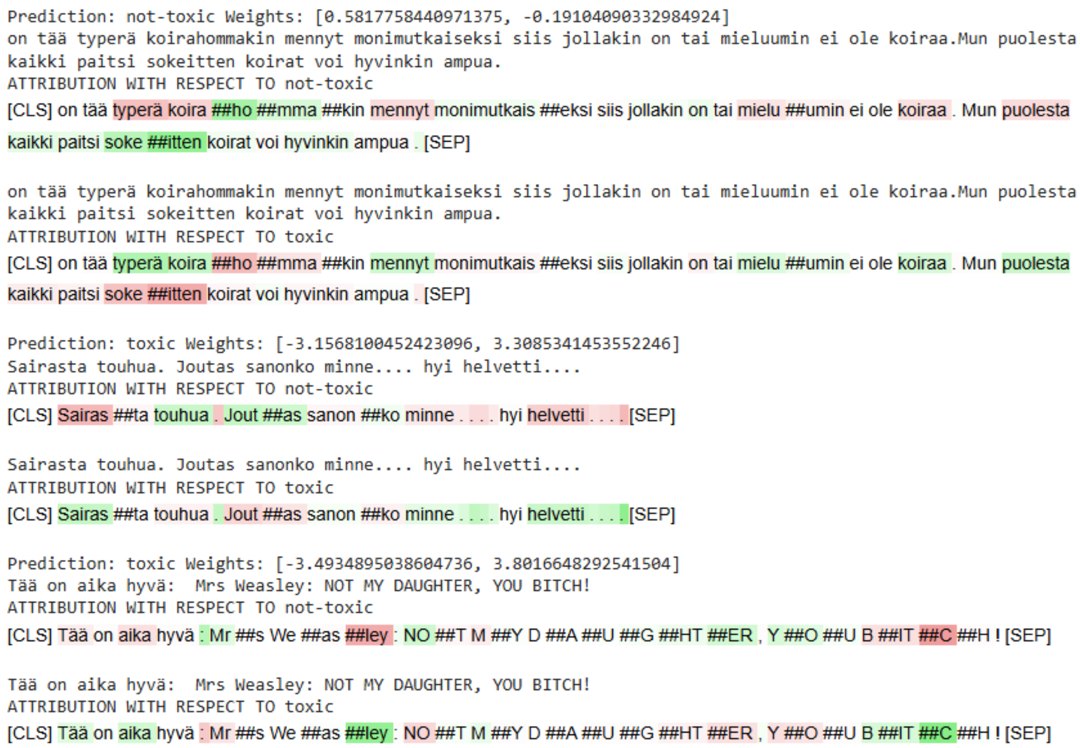

Figure 4.7: Visualization of what the toxicity model focuses on in texts while making predictions.

and is translated to "The president should unify the people". Yet, a text in all caps could imply that the text is "yelling" and aggressive, meaning that it could in itself be a sign of toxicity and any noisy thing that text pre-processing usually takes away could be a signal for toxicity that the model needs. Regarding non-toxic content, what the model sees as non-toxic can be difficult, as that could be basically anything and there might be no real signal to what is non-toxic and only a signal for toxic content exists. For many texts, there seems to be no clear signal for the model and it seems to think some random words could be toxic or non-toxic.

One cause for the misclassifications is the mismatch on how sensitive the annotator is to toxicity compared to the model. Annotation is not an easy task and toxicity annotation specifically can be highly subjective so the labels could vary person to person. Another possible reason is how carefully the full text is read, because if the annotations are done fast, there may also be plain mistakes which might also be

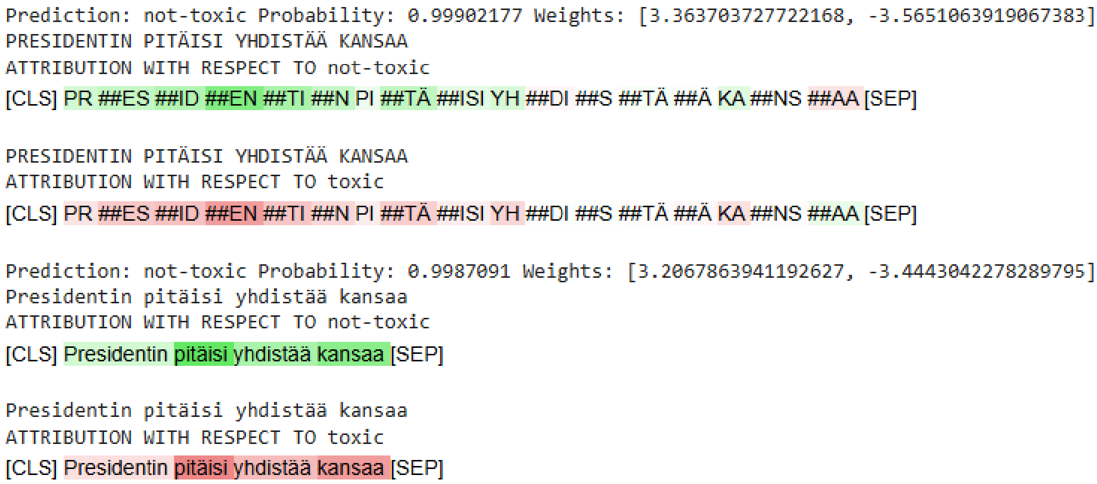

Figure 4.8: Sentence in all caps vs. regular sentence visualized.

the case in some of my own annotations. A problem with this re-annotation that was already mentioned in Chapter 3.3, is that the re-annotation was done only by one person, whereas the original Finnish dataset was annotated by three people. This makes the dataset susceptible to mistakes or bias, as the annotations were not double checked and jointly resolved as the original comments were for the different types of toxicity.

4.3.3 Model Usage

The toxicity detection model is relatively easy to use with ready-made code for inference and even going through big chunks of data is fairly easy by scaling to multiple GPUs. Tutorials for the use of the models can be found from e.g., Huggingface and a simple example of a pipeline is shown on the models Huggingface page[1].

The model introduced in Eskelinen, Silvala, Ginter, *et al.* [8] was used in the paper by Luukkonen, Komulainen, Luoma, *et al.* [19] to clean large Finnish datasets for the training of a generative LLM. This process is especially important for generative models, as they should not produce any toxicity. The model filtered as much

[1]https://huggingface.co/annieske/bert-base-finnish-cased-toxicity

as 23% of the CC-Fi dataset as that dataset had not been filtered for e.g., obscenity previously like the others from which only 1-5% of text was removed. The toxicity of the generations of the trained model was analyzed and they found that toxicity was generated less, than for models which were not filtered for toxicity. Nevertheless, toxicity was produced unprompted 2% of the time leaving still room for improvement. The model has also been used to label Finnish Reddit content[2] alongside some other datasets.

A potential use case for a toxicity model is, for example, automatic content moderation on social media sites, although recently some sites have cut down on content moderation e.g., Meta (includes e.g., Instagram, Facebook, Threads) has lifted restrictions on certain topics in their content moderation policies[3] leaving the platform vulnerable to hate speech and other forms of toxicity.

As a content moderation tool, the model built during the creation of this thesis can help flag comments for moderation, but in the current state, the model cannot fully on its own do content moderation, as the model is too sensitive and mistakes can happen in both directions. But because recall is high for the toxic label, the model can be used quite safely and be mostly trusted to not let through toxic texts. As the model is trained on the unified labels and has a binary mapping of the original labels, fine-grained information about the type of toxicity is lost and cannot be gained from the predictions. A multilabel model could thus perhaps be more informative in certain cases, especially in content moderation where some types of toxicity are more dangerous than others, but this would require more data with many labels, which is a process that takes a lot of resources. The model's effectiveness on real Finnish data is proven, but the formal language used in the translations and colloquial Finnish seen in real data create a mismatch and perhaps better results

[2]https://huggingface.co/datasets/Finnish-NLP/Reddit_fi_2006_2022_cleaned_v2

[3]https://theintercept.com/2025/01/09/facebook-instagram-meta-hate-speech-content-moderation/

could be obtained if a bigger dataset for training was annotated on real Finnish, but this is as mentioned, a process that would require a lot of resources.

Technically, the model can also be used to rank how toxic a text is, i.e. worse stuff should have a higher probability, although that is not inherently correct as the prediction is just a probability of how certain the model is of its predictions. Nevertheless, examples of the "most toxic" texts are included in Appendix A. In the 10 most toxic texts as predicted by the model, there were two that were labeled as non-toxic by me.

5 Limitations & Future Work

Although the research done on this thesis has developed a new, better toxicity model and investigated the unification of datasets, machine translation as a form of cross-lingual transfer and the use of the FinBERT model on the toxicity detection task, the limitations of the work should be acknowledged and noted for potential future work.

FinBERT was chosen as the model for fine-tuning as it was still state-of-the-art when the work on this topic began. In the meantime, technology such as ChatGPT and other generative LLMs started trending and became the default to use on many tasks by fine-tuning or prompting. Thus, for future work, generative large language models could be used for toxicity detection, as ChatGPT has proved to be quite the annotator. For example, Henriksson, Tarkka, and Ginter [11] used a generative LLM (GPT4-o mini) to annotate texts on the line-level to get rid of bad quality content for training new LLMs, including toxic and explicit adult content. This use case for toxicity detection was already mentioned briefly in Chapter 4.3.3. They note that GPT4-o mini can sometimes be too sensitive, labeling content such as mild expletives as toxic, which matches the results of this thesis as my model was also too sensitive. This is an additional avenue on which future work can be done to make sure models are not too sensitive, but also not too lax.

OpenAI has a content moderation model available for free[1] for classifying toxicity

[1]https://platform.openai.com/docs/guides/moderation?example=text

in generated texts to help people using the GPT models not output toxic content to the public. Even though ChatGPT and other huge LLMs with millions of parameters can be great tools, for large scale annotation it can become quite costly, meaning smaller models should be built for faster and less computationally intensive inference.

As the focus of the thesis was on machine translation as a form of cross-lingual transfer, multilingual models trained on the original data were out of scope for the thesis. This choice was due to the fact that translation and using a monolingual model was deemed the better approach in previous work [8]. A multilingual model such as XLM-R [4] could be tested in future work on both the original datasets and their translations to see if they behave similarly. The use of XLM-R for cross-lingual transfer is also more straightforward and less expensive than translation, as one can just use the model on data in almost any language.

The BERT family of models is limited to a sequence length of 512, which normally means that the rest of the text is truncated, although methods to circumvent this also exist. This is a limitation of the models and thus of this work, as the end of the text might also contain relevant information. Thankfully newer models can support longer contexts so no information is left out.

Unfortunately, toxicity datasets are possibly a dying breed, as for example X (previously Twitter), cut down academic access and building similar datasets is not as easy as before. In addition, content moderation makes it hard to find these toxic comments for future training unless you own the platform and gather the comments while doing moderation, although as mentioned in model usage, some platforms are cutting down on content moderation. Besides this, large language models cannot make synthetic data for toxicity identification, as usually models are aligned so that they should not produce any toxic content, prompted or unprompted.

6 Conclusion

The purpose of this thesis was to answer the research question *"To what extent is it possible to use machine translated unified datasets to build good quality models for toxicity detection in Finnish?"* and answer three subquestions:

1. How well does a corpus from another culture fair in the Finnish context?
2. Does a model trained on one or multiple corpora work on a different corpus?
3. How safe and useful can the model be in actual use?

Chapter 2 laid the foundations for the thesis by explaining the transformer architecture, the process of fine-tuning and FinBERT, which was the pre-trained model chosen for training on the text classification task. Machine translation was also introduced as a form of cross-lingual transfer. Toxicity was specifically defined as "rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion", and related work on similar tasks was introduced. Lastly, the metrics used to evaluate the fine-tuned language models were established.

Chapter 3 introduced the many datasets used in the thesis, starting from the Jigsaw dataset which was machine translated previously [8], continuing with the other smaller machine translated datasets and two manually annotated Finnish datasets, the second being a re-annotated subset of the first one. Finally, the process of unifying the many datasets into one large dataset was explained.

The first two subquestions were explored through various experiments which tested the existing model against different datasets and then trained new models which were also tested on different datasets, notably the re-annotated Finnish dataset, which was manually annotated for the purposes of this thesis. The results showed that the machine translated datasets work well on the re-annotated Finnish dataset and that even though the datasets come from different cultures and languages, the model still works well. Testing different models on the small datasets highlighted differences in the datasets through the differences in results and some datasets were deemed more compatible with each other than others. The differences in results and the reasons for compatibility in some datasets were hypothesized to be due to different distributions of labels, different original languages, and different domains.

Based on the research described in the Background in Chapter 2, I was skeptical of how well the unified dataset would do as training data for a new toxicity model due to the fact the datasets were from different social media platforms (domain), had different methods for collecting and annotating the datasets, different definitions for the task, and different labels. Nevertheless, in the end according to the results reported in Chapter 4, the model that was trained on all the different datasets was deemed the best performing, with only one dataset left out performing equally well. Regardless, the dataset was preserved for the best model to keep the training as variable as possible and because more data is often beneficial to have.

The safety of the best-performing model, as well as discussion about the trade-off between precision and recall, was further examined in Chapter 4.3, where the thresholds for the label and how the model is sensitive were explored. The model had higher recall than precision, especially for the toxic label meaning that some non-toxic content can be flagged by the model as toxic. Misclassifications were further examined in Chapter 4.3, where possible reasons for the wrong predictions were discussed and what the model sees before making predictions was visualized with integrated gradients and discussed.

The usefulness and use of the model was discussed last, with examples of how the previous model [8] has been used to filter training data for generative models and to augment datasets with information about toxicity in texts. Automatic content moderation was also mentioned as a possible use case, although the model cannot fully replace humans just yet, instead it can act as a tool.

Lastly, to answer the overarching research question, according to the results, it seems to be possible to build high-quality models for toxicity detection in Finnish using many different machine translated datasets that are unified. The toxicity detection task is difficult and can be subjective, making it hard for even humans to annotate texts consistently, as outlined in Chapter 2 when exploring previous work. For a text classification model, there is of course always room for improvement, in this case especially in terms of precision. Because of this, possible future work and limitations of the work were considered in Chapter 5. The work on the thesis produced several artifacts, a new toxicity model for Finnish, a small manually annotated test set for toxicity of real Finnish from the Suomi24 corpus and the machine translated datasets alongside code for unifying the datasets and training an LLM for toxicity detection.

ChatGPT was used as a tool to assist with summarization, rephrasing, and improving the clarity of written content throughout parts of the thesis. All text provided by the tool was carefully checked and edited. It did not influence the design, analysis, or interpretation of the research itself.

References

[1] E. Blakey, "The day data transparency died: How twitter/x cut off access for social research", *Contexts*, vol. 23, pp. 30–35, May 2024. DOI: 10.1177/15365042241252125.

[2] U. Bretschneider and R. Peters, "Detecting cyberbullying in online communities", Jun. 2016.

[3] S. Chaudhury, A. Banerjee, and J. Bhawalkar, "Hypothesis testing, type i and type ii errors", *Industrial psychiatry journal*, vol. 18, p. 127, Jul. 2009. DOI: 10.4103/0972-6748.62274.

[4] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, "Unsupervised cross-lingual representation learning at scale", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. [Online]. Available: https://aclanthology.org/2020.acl-main.747/.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language", in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ser. ICWSM '17, Montreal, Canada, 2017, pp. 512–515.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: `1810.04805 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/1810.04805`.

[7] J. Eronen, M. Ptaszynski, F. Masui, M. Arata, G. Leliwa, and M. Wroczynski, "Transfer language selection for zero-shot cross-lingual abusive language detection", *Information Processing & Management*, vol. 59, no. 4, p. 102 981, 2022, ISSN: 0306-4573. DOI: `https://doi.org/10.1016/j.ipm.2022.102981`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0306457322000978`.

[8] A. Eskelinen, L. Silvala, F. Ginter, S. Pyysalo, and V. Laippala, "Toxicity detection in Finnish using machine translation", in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, T. Alumäe and M. Fishel, Eds., Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 685–697. [Online]. Available: `https://aclanthology.org/2023.nodalida-1.68`.

[9] A.-M. Founta, C. Djouvas, D. Chatzakou, *et al.*, "Large scale crowdsourcing and characterization of twitter abusive behavior", in *11th International Conference on Web and Social Media, ICWSM 2018*, AAAI Press, 2018.

[10] Geetanjali and M. Kumar, "Exploring hate speech detection: Challenges, resources, current research and future directions", *Multimedia Tools and Applications*, pp. 1–37, Mar. 2025. DOI: `10.1007/s11042-025-20716-2`.

[11] E. Henriksson, O. Tarkka, and F. Ginter, "FinerWeb-10BT: Refining web data with LLM-based line-level filtering", in *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, R. Johansson and S. Stymne, Eds., Tallinn, Estonia: University of Tartu Library, Mar.

2025, pp. 258–268. [Online]. Available: https://aclanthology.org/2025.nodalida-1.27/.

[12] M. Huzaifah, W. Zheng, N. Chanpaisit, and K. Wu, "Evaluating code-switching translation with large language models", in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 6381–6394. [Online]. Available: https://aclanthology.org/2024.lrec-main.565/.

[13] S. Joseph, A. G. Parthi, D. Maruthavanan, V. Jayaram, P. K. Veerapaneni, and V. Parlapalli, "Transfer learning in natural language processing", in *2024 7th International Conference on Information and Communications Technology (ICOIACT)*, 2024, pp. 30–36. DOI: 10.1109/ICOIACT64819.2024.10912895.

[14] J. K. Kobellarz and T. H. Silva, "Should we translate? evaluating toxicity in online comments when translating from portuguese to english", in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, ser. WebMedia '22, Curitiba, Brazil: Association for Computing Machinery, 2022, pp. 89–98, ISBN: 9781450394093.

[15] J. K. Kobellarz and T. H. Silva, "Should we translate? evaluating toxicity in online comments when translating from portuguese to english", ser. WebMedia '22, Curitiba, Brazil: Association for Computing Machinery, 2022, pp. 89–98, ISBN: 9781450394093. DOI: 10.1145/3539637.3556892. [Online]. Available: https://doi.org/10.1145/3539637.3556892.

[16] P. Kralj Novak, I. Mozetič, and N. Ljubešić, *Slovenian twitter hate speech dataset IMSyPP-sl*, Slovenian language resource repository CLARIN.SI, 2021. [Online]. Available: http://hdl.handle.net/11356/1398.

[17] V. Laippala, R. Kyllönen, J. Egbert, D. Biber, and S. Pyysalo, "Toward multilingual identification of online registers", in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, M. Hartmann and B. Plank, Eds., Turku, Finland: Linköping University Electronic Press, 2019, pp. 292–297. [Online]. Available: `https://aclanthology.org/W19-6130/`.

[18] J. A. Leite, D. Silva, K. Bontcheva, and C. Scarton, "Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis", in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 914–924.

[19] R. Luukkonen, V. Komulainen, J. Luoma, *et al.*, "FinGPT: Large generative models for a small language", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2710–2726. DOI: `10.18653/v1/2023.emnlp-main.164`. [Online]. Available: `https://aclanthology.org/2023.emnlp-main.164`.

[20] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 14 867–14 875.

[21] D. Nozza, "Exposing the limits of zero-shot cross-lingual hate speech detection", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 907–914.

[22] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4675–4684. DOI: `10.18653/v1/D19-1474`. [Online]. Available: `https://aclanthology.org/D19-1474/`.

[23] A. Pelicon, R. Shekhar, M. Martinc, B. Škrlj, M. Purver, and S. Pollak, "Zero-shot cross-lingual content filtering: Offensive language and hate speech detection", in *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, Online: Association for Computational Linguistics, Apr. 2021, pp. 30–34. [Online]. Available: `https://aclanthology.org/2021.hackashop-1.5`.

[24] "Perspective api attributes and languages". (2024), [Online]. Available: `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en%5C_US`.

[25] "Perspective api training data". (2024), [Online]. Available: `https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en%5C_US`.

[26] S. Poplack, "Syntactic structure and social function of code-switching", in Jan. 1981, pp. 169–184.

[27] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong,

China: Association for Computational Linguistics, Nov. 2019, pp. 4755–4764. DOI: 10.18653/v1/D19-1482. [Online]. Available: https://aclanthology.org/D19-1482/.

[28] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training", 2018.

[29] C. Rastogi, N. Mofid, and F. Hsiao, "Can we achieve more with less? Exploring data augmentation for toxic comment classification", 2020. arXiv: 2007.00875.

[30] L. Repo, V. Skantsi, S. Rönnqvist, *et al.*, "Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, I.-T. Sorodoc, M. Sushil, E. Takmaz, and E. Agirre, Eds., Online: Association for Computational Linguistics, Apr. 2021, pp. 183–191. DOI: 10.18653/v1/2021.eacl-srw.24. [Online]. Available: https://aclanthology.org/2021.eacl-srw.24/.

[31] J. Risch, P. Schmidt, and R. Krestel, "Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format", in *Proceedings of the Workshop on Online Abuse and Harms (WOAH)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 157–163. DOI: 10.18653/v1/2021.woah-1.17. [Online]. Available: https://aclanthology.org/2021.woah-1.17.

[32] S. Rönnqvist, V. Skantsi, M. Oinonen, and V. Laippala, "Multilingual and zero-shot is closing in on monolingual web register classification", in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, S. Dobnik and L. Øvrelid, Eds., Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 157–165. [Online]. Available: https://aclanthology.org/2021.nodalida-main.16/.

[33] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "A large-scale semi-supervised dataset for offensive language identification", *arXiv preprint arXiv:2004.14454*, 2020.

[34] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers", *Information Systems*, vol. 121, p. 102 342, Dec. 2023. DOI: 10.1016/j.is.2023.102342.

[35] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks", in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328.

[36] "Toxic comment classification challenge". (2024), [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[37] TurkuNLP, *Jigsaw toxicity dataset finnish translation*, 2023. [Online]. Available: https://huggingface.co/datasets/TurkuNLP/jigsaw%5C_toxicity%5C_pred%5C_fi.

[38] TurkuNLP, *Suomi24 toxicity annotations*, 2023. [Online]. Available: https://huggingface.co/datasets/TurkuNLP/Suomi24-toxicity-annotated.

[39] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper%5C_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[40] A. Virtanen, J. Kanerva, R. Ilo, *et al.*, *Multilingual is not enough: Bert for finnish*, 2019. arXiv: 1912.07076 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1912.07076.

[41] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter", in *Proceedings of the NAACL Student Research Workshop*, J. Andreas, E. Choi, and A. Lazaridou, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93. DOI: 10.18653/v1/N16-2013. [Online]. Available: https://aclanthology.org/N16-2013/.

[42] Wikimedia. "Roc curve". (2018), [Online]. Available: https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg.

[43] G. Winata, A. F. Aji, Z. X. Yong, and T. Solorio, "The decades progress on code-switching research in NLP: A systematic survey on trends and challenges", in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 2936–2978. DOI: 10.18653/v1/2023.findings-acl.185. [Online]. Available: https://aclanthology.org/2023.findings-acl.185/.

[44] Y. Wu, M. Schuster, Z. Chen, *et al.*, *Google's neural machine translation system: Bridging the gap between human and machine translation*, 2016. arXiv: 1609.08144 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1609.08144.

[45] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1415–1420. DOI: 10.18653/v1/N19-1144. [Online]. Available: https://aclanthology.org/N19-1144/.

[46] R. Zhang, S. Cahyawijaya, J. C. B. Cruz, G. Winata, and A. F. Aji, "Multilingual large language models are not (yet) code-switchers", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 567–12 582. DOI: 10.18653/v1/2023.emnlp-main.774. [Online]. Available: https://aclanthology.org/2023.emnlp-main.774/.

[47] Y. Zhu, R. Kiros, R. Zemel, *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books", ser. ICCV '15, USA: IEEE Computer Society, 2015, pp. 19–27, ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.11. [Online]. Available: https://doi.org/10.1109/ICCV.2015.11.

[48] F. Zhuang, Z. Qi, K. Duan, *et al.*, "A comprehensive survey on transfer learning", English (US), *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, Publisher Copyright: © 2020 IEEE., ISSN: 0018-9219. DOI: 10.1109/JPROC.2020.3004555.

Appendix A Most Toxic Texts

Text (Finnish)	**Translation (English)**	**True**	**Pred**	**Prob.**
Ime sä kuule sykkivääs.. Mullo muskelit!	You suck that pulsing.. I've got muscles!	1	1	0.9998
Haista Vittu Fuck You.	Fuck you / Fuck you.	1	1	0.9998
Vit-tu, että osaakin olla ruma ämmä!!! :-):-):-)	Fu-ck, she really can be one ugly bitch!!! :-):-):-)	1	1	0.9998
Antonio=Korkkaa teininaisen omistukseensa Aisuri Ti**=Nuolee vaimonsa huoranvitusta Antonion spermanjämät.	Antonio = Claims a teenage girl for himself. Cuck Ti** = Licks his wife's slut pussy clean of Antonio's leftover sperm.	1	1	0.9998
minulla oikeus pitää saastaisena huorana naista joka jakelee vittuaan ympäri kyliä.	I have the right to consider as a filthy whore any woman who spreads her pussy all over the villages.	1	1	0.9998
Millä tavalla äitisi murskasi unessa munasarjasi?	How did your mother crush your ovaries in your dream?	0	1	0.9998
Haluuksä nussii mun kans? Tehdään se heti!	Wanna fuck with me? Let's do it now!	0	1	0.9998
Oletko koskaan ollut miesten kanssa? Oletko saanut koskaan kullia pillusi? Tykkäätkö vain naisista ja miksi? Eihän pillua voi työntää pilluun vai voiko? Minkälaisista naisista sitten pidät ja miksi?	Have you ever been with men? Have you ever had a dick in your pussy? Do you only like women and why? You can't stick a pussy into a pussy, or can you? What kind of women do you like and why?	1	1	0.9998
..hyi helvetti. vanha ämmä.	..ugh, eww fucking hell. old bitch.	1	1	0.9998

Table A.1: Texts that were predicted as most (certainly) toxic by the best toxicity detection model.