ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Master Thesis
in
Computer Science

Conversational Context in Toxicity Detection

Alexandros Xenos

Supervisors: John Pavlopoulos
Ion Androutsopoulos

September 2021

Alexandros Xenos
Conversational Context in Toxicity Detection
September 2021
Supervisors: John Pavlopoulos and Ion Androutsopoulos

Athens University of Economics and Business
School of Information Sciences and Technology
Department of Informatics
AUEB NLP Group
Athens, Greece

Abstract

This thesis focuses on context-aware toxicity detection. Most existing work ignores the conversational context, focusing on context-unaware datasets training toxicity detectors that learn to disregard the conversational context. This makes the detection of posts whose perceived toxicity depends on the conversational context a lot harder. This work utilizes the CCC dataset, a new context-aware dataset of 10,000 posts, created by Google Jigsaw and the AUEB NLP group, containing both context-aware and context-unaware annotations. In the first case, annotators had access to the previous post of the conversation. Based on this dataset, this work introduces a new task, *context sensitivity estimation*, which aims to identify posts whose perceived toxicity changes if the context (previous post) is also considered. Then, traditional machine learning algorithms along with more complex NLP models are evaluated on this task achieving low error. These systems can further be improved by using data augmentation with knowledge distillation. Context sensitivity estimation systems could be used in order to enrich toxicity datasets with more context-sensitive posts. Moreover they can be used as a suggestion tool to moderators, advising them when to consider the parent posts, which may often be unnecessary and may otherwise introduce significant additional cost.

Acknowledgements

First of all I would like to thank my supervisor Prof. Ion Androutsopoulos for giving me the opportunity to work with him. I would also like to praise his mentoring through the whole process. Secondly I would like to thank my co-supervisor Prof. John Pavlopoulos, who supported my e□orts through my whole research and always provided me with the best advice and tips, while being always available when I needed advice. It was a pleasure to work and cooperate with both of them. Last but not least I would like to thank my family, my parents, my two sisters and my very good friend Joanna who support me unconditionally all these years.

Contents

1 Introduction

1.1 Motivation and Problem Statement

Social Media are platforms where millions of people interact and communicate with each other on a daily basis. However, social media platforms are not always civil, with multiple users being toxic towards others, causing them to leave a conversation or stopping them from sharing their perspective. Thus, moderation is crucial in order to promote healthy conversations online. Artificial intelligence and more specifically Natural Language Processing (NLP) can be used in order to automate moderation. A lot of work has been done towards this area in the previous years, but most of this work ignores the context of the conversation of the target post that is going to be classified as toxic or not, making the detection of context-sensitive toxicity a lot harder when it occurs.

In this work, context is considered to be any information relevant in order to decode the meaning and intention of a post. When the context is not present, the interpretation of a post can be more ambiguous. Therefore context can be very useful in order to create datasets with high-quality context-aware annotations with high inter-annotator agreement, or when more information is needed in order to decode the intention of a post regarding its perceived toxicity.

Given this approach, this work focuses on the past conversational context, specifically the previous post in a discussion. For instance, a post "Of course, they should!!" is likely to be considered as non-toxic by a moderator who has not seen that the parent post was "Do you believe that all minorities should die?".

Although toxicity datasets that include conversational context have recently started to appear, in previous work Pavlopoulos, Sorensen, et al. (2020a) showed that context-sensitive posts (i.e. posts whose perceived toxicity depends on the conversational context) seem to be rare and this makes it hard for models to learn to detect context-dependent toxicity when it occurs. In this work, to study this problem, a context-aware dataset of 10,000 posts was used. Each of these posts was annotated by raters who (i) had access to the previous (*parent*) post as context during the annotation, apart from the post being annotated (the *target* post), and by raters who (ii) did not have any context during the annotation process.[1]

[1] The dataset is released under a CC0 licence. See `http://nlp.cs.aueb.gr/publications.html` for the link to download it.

As a first step to study the role of conversational context in toxicity detection, in this work, the context is limited to the previous post of the thread as in (Pavlopoulos, Sorensen, et al., 2020a). In their work, the authors mentioned some basic challenges of studying context. First of all, it is expensive and can be time consuming to consider it on crowd-sourcing platforms, because ensuring that a person has in fact considered the context is hard. Secondly providing the annotators with more context and more subtle kinds of context in general, makes it a lot harder for the annotators to account for it. Moreover context sensitive toxicity in posts is also rare.

Then, the context-aware dataset is used to study the nature and the role of context sensitivity in toxicity detection, and a new task is introduced, *context sensitivity estimation*, which aims to identify posts whose perceived toxicity changes if the context (previous post) is also considered. Using the dataset, it is also shown that it is possible to develop systems that can achieve low error on this new task. Such systems could be used to enhance toxicity detection datasets with more context-sensitive posts. Moreover they can be used as a suggestion tool to moderators, advising them when to consider the parent posts, which may often be unnecessary and may otherwise introduce significant additional cost.

1.2 Thesis Structure

The rest of the thesis is organised as follows:

1. Chapter 2 discusses background and related work.

2. Chapter 3 presents the new dataset and the new task (context-sensitivity estimation).

3. Chapter 4 presents the models and their architectures.

4. Chapter 5 contains the experimental results.

5. Chapter 6 discusses the conclusions and the future work.

2 Background and Related Work

2.1 Background

In this chapter we remind the reader of some basic concepts that they should know in order to attend the thesis. First, we remind the reader what machine learning is and more speci cally what deep learning is. Then we give some basic background knowledge for the toxicity detection task. Finally, we explain what data augmentation is.

2.1.1 Machine Learning

Machine learning is a sub- eld of arti cial intelligence. It studies algorithms and systems that can make decisions with minimal human intervention. It is based on the idea that data can be used in order to help systems learn. There are four subcategories of machine learning; supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Supervised Learning relies on learning from labeled data. In this type of learning, datasets are used, containing data points (instances) whose ground truth has been labeled by one or more humans. This means that in supervised learning a training dataset includes inputs and correct outputs, which allow the model to learn over time.

Unsupervised Learning relies on learning from unlabeled data. It uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

Semi-supervised Learning involves a small number of labeled data and a large number of unlabeled data. Unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in a model's performance.

Reinforcement Learning is a machine learning method where a model learns how to make a sequence of decisions, and it is often rewarded (or penalized) only at the end of the sequence of decisions. The model, often called an agent, learns to achieve a goal in an uncertain, potentially complex environment.

2.1.2 Deep Learning

Deep learning (Goodfellow et al., 2016; Goldberg et al., 2017) is a sub- eld of machine learning, where deep neural network architectures are developed in order to solve the same problems that traditional machine learning is trying to solve. In this thesis both traditional machine learning algorithms (e.g., random forests (Ho, 1995)) and deep learning models were used but this thesis focus more on the deep learning models as they are most often the state-of-the-art methods to use.

2.1.3 Toxicity Detection

Toxicity detection is a well-known NLP task, where machine learning techniques are employed, in order to create systems that automatically detect toxic content in online platforms, social media or more generally in platforms where people interact. In this thesis we focus on detecting toxic posts. Some people de ne a post as toxic, if it is a very hateful, aggressive, or disrespectful comment that is very likely to make someone leave a discussion or give up on sharing his perspective (Borkan et al., 2019). Following the work of Pavlopoulos, Sorensen, et al. (2020b) this work uses the term `toxic' as an umbrella term, but the literature uses several terms for di erent kinds of toxic language or related phenomena: `o ensive' (Zampieri et al., 2019), `abusive' (Pavlopoulos, Malakasiotis, and Androutsopoulos, 2017a), `hateful' (Djuric et al., 2015; ElSherief et al., 2018; Zhang et al., 2018), etc. As Waseem, Davidson, et al. (2017) observed, there are also taxonomies for these phenomena based on their directness (e.g., whether the abuse was unambiguously implied/denoted or not), and their target (e.g., whether it was a general comment or targeting an individual/group).

2.1.4 Data Augmentation

Data augmentation in data analysis includes various techniques used to increase the amount of training data. This can be achieved by adding slightly modi ed copies of already existing training instances or newly created synthetic data from existing data or in a semi-supervised fashion by adding new machine-labeled data to the training set (Feng et al., 2021; Bayer et al., 2021). It sometimes acts as a regularizer and helps reduce over tting when training a machine learning model (Shorten et al., 2019).

2.2 Related Work

This section describes related work by following three dimensions. First, it describes work regarding toxicity detection. Second, it focus on context-aware natural language

processing approaches. Finally, it describes work that tackles classi cation tasks with regression-based approaches.

Toxicity detection

Abusive language detection is not an easy task due to its subjective nature. Victims are getting attacked by cyberbullies on di erent topics such as gender, race, and religion across multiple Social Media Platforms (Agrawal et al., 2018). Thus, in a di erent context, the vocabulary used and the perceived meaning of words may vary when abusive language occurs. In order to tackle the problem of abusive language detection, researchers have been experimenting with several approaches. Initially machine learning techniques using hand crafted features such as lexical features, syntactic features, etc. (Davidson et al., 2017; Waseem and Hovy, 2016; Djuric et al., 2015) were used. Then, deep learning techniques were employed, operating on word embeddings (Park et al., 2017; Pavlopoulos, Malaka-siotis, and Androutsopoulos, 2017b; Pavlopoulos, Malakasiotis, Bakagianni, et al., 2017; Chakrabarty et al., 2019; Badjatiya et al., 2017; Haddad et al., 2020; Ozler et al., 2020). These techniques seem to work better for this task than the traditional machine learning methods based on handcrafted features (Badjatiya et al., 2017).

Researchers have publicly released a lot of datasets containing di erent types of toxicity, to help expand the research on this eld. The rst corpus annotated for abusive language was developed by Nobata et al. (2016). This dataset comprises user comments posted on Yahoo Finance and News. Wulczyn et al. (2017) created and experimented with three new datasets; the Personal Attack dataset where 115K comments from Wikipedia Talk pages were annotated as containing personal attack or not, the Aggression dataset where the same comments were annotated as being aggressive or not, and the Toxicity dataset that includes 159K comments again from Wikipedia Talk pages that were annotated as being toxic or not. Waseem and Hovy (2016) annotated by themselves a corpus for hate speech detection of more than 16k tweets, containing sexist, racist and non-toxic posts. Most of the published toxicity datasets contain posts in English, but datasets in other languages also exist, such as French (Chiril et al., 2020), Greek (Pavlopoulos, Malakasiotis, and Androutsopoulos, 2017a), German (Ross et al., 2016; Wiegand et al., 2018), Arabic (Mubarak et al., 2017) and Indonesian (Ibrohim et al., 2018).

Context-aware NLP

The integration of context into human language technology has been successfully applied to various applications and domains. Context plays a central role in text/word representation (Mikolov et al., 2013; Pennington et al., 2014; Melamud et al., 2016; Peters et al., 2018; Devlin et al., 2019). Integrating context is important in the sentiment analysis task too,

where the semantic orientation of a word changes according to the domain or the context in which that word is being used (Agarwal et al., 2015). Vanzo et al. (2014) examined the role of incorporating contextual information in supervised Sentiment Analysis over Twitter. They experimented with two di erent types of contexts, a conversation-based context and a topic-based context, which includes several tweets in the history stream that contain overlapping hashtags. In their work, each tweet and its context were modeled as a sequence of tweets and they used a sequence labeling model, SVM^{HMM}, to predict their sentiment labels jointly. The authors found that these employed contexts provide bene ts in sentiment classi cation. Ren et al. (2016) proposed a context-based neural network model for Twitter sentiment analysis, incorporating contextualized features from relevant tweets into the model in the form of word embedding vectors. They experimented with three types of context, a conversation-based context, an author-based context and a topic-based context. They found that integrating contextual information about the target tweet in their neural model o ers improved performance compared with the state-of-the-art, discrete and continuous word representation models. They also reported that topic-based context features were the most e ective for this task.

While context is widely used in other Natural Language Processing (NLP) tasks, such as dialogue systems (Lowe et al., 2015; Du²ek et al., 2016), informational bias detection (Berg et al., 2020) etc., very limited work has focused on context-aware toxic language detection. Gao et al. (2017) provided a corpus for hate speech detection. The dataset contains posts under Fox News articles obtained from full threads of online discussion. The authors proposed two types of hate speech detection models, that incorporate context information, a logistic regression model with context features and a neural network model with learning components for context. They reported performance gains in F1 score when incorporating context and that, combining these two models further improved the performance by another 7% in F1 score. Mubarak et al. (2017) provided the annotators with the title of the respective news article, but they ignored parent comments since they did not have the entire thread. As Pavlopoulos, Sorensen, et al. (2020a) already observed, this presents the following problem: new comments may change the topic of the conversation and replies may require the previous posts to be assessed correctly. Pavlopoulos, Malakasiotis, and Androutsopoulos (2017a) provided the annotators with the whole conversation thread for each target comment as context during the annotation process. The plain text of the comments for this dataset is not available, which makes further analysis di cult. In later work Pavlopoulos, Sorensen, et al. (2020a) published two new toxicity datasets containing posts from the Wikipedia Talk pages, where during the annotation process, annotators were provided with the previous post in the thread and the discussion title. The authors found that providing annotators with context can result both in ampli cation or mitigation of the perceived toxicity of posts. Moreover, they found no evidence that context actually improves the performance of toxicity classi ers. In a similar work that was conducted by Menini et al. (2021), the authors investigated the role of textual context in abusive language detection on Twitter. They rst re-annotated the tweets in the dataset from (Founta et al.,

2018) in two conditions, i.e. with and without context. After comparing the two datasets (with and without context-aware annotations) they found that the context is sometimes necessary to understand the real intent of the user, and that it is more likely to mitigate the abusiveness of a tweet even if it contains profanities. Finally they experimented with several classi ers, using both context-aware and non-context-aware architectures. Their experimental results showed that when context is given as input to the classi ers and they are evaluated on context-aware datasets, their performance drops dramatically as opposed to when context is not given as input and they are evaluated on non-context-aware datasets.

Regression as classi cation in NLP

Although unusual, approaching a text classi cation problem as a regression-based problem has been tested by researchers in various Natural Language Processing tasks, such as sentiment analysis (Wang et al., 2016), emotional analysis (Buechel et al., 2016), metaphor detection (Parde et al., 2018), toxicity detection, etc. Wulczyn et al. (2017) proposed a regression-based evaluation method for a classi er, in terms of the aggregated number of crowd-workers it can approximate for personal attacks detection. The authors observed that using the empirical distribution of human-ratings, instead of the majority vote when estimating the likelihood of a post to be personal attack or not, produces a better classi er. D'Sa et al. (2020) experimented on the toxicity detection task, using the English Wikipedia Detox corpus. They designed both binary classi cation and regression-based approaches aiming to predict whether a comment is toxic or not. They examined and compared di erent unsupervised word representations and di erent deep learning based classi ers. In most of their experiments, they found that the regression-based approach showed slightly better performance than the classi cation setting which is consistent with the ndings of Wulczyn et al. (2017).

3 Dataset

3.1 Collecting the Data

This work utilizes the Civil Comments in Context (CCC) dataset that was created by Google Jigsaw[1] and the AUEB NLP group[2]. CCC was created by randomly sampling 10,000 posts from the Civil Comments (CC) dataset (Borkan et al., 2019). While in CC, each post was annotated by ten annotators who did not have access to any conversational context, in CCC each post was annotated by 5 annotators who had access to the previous post in the discussion (parent post) as context. Each CCC post was rated either as "not toxic", "unsure", "toxic", or "very toxic", as in the original CC dataset. To simplify the problem the latter two labels were uni ed in both CC and CCC annotations. To obtain the new in-context labels of CCC, the APPEN platform and ve high accuracy annotators per post (annotators from zone 3, allowing adult and warned for explicit content) were used, selected from 7 English speaking countries, namely: UK, Ireland, USA, Canada, New Zealand, South Africa, and Australia.[3] The constructing cost of CCC was 2,865 euros.

3.2 Explanatory Data Analysis

The inter-annotator agreement of this dataset was measured with the free-marginal kappa and the average (mean pairwise) percentage agreement. They were found to be 83.93% and 92% respectively. In only 71 posts (0.07%) an annotator used "unsure", meaning annotators were con dent in their decisions most of the time. These 71 posts were excluded from our study, as there are too few to generalize about. When counting the average length (in characters) of the target and the parent posts in CCC, the rst is only slightly lower than the latter (see gure 3.1). The same holds when counting in words: 56.5 vs. 68.8 words on average (see gure 3.2 and 3.3). To obtain a single toxicity score per post, the percentage of the annotators who found the post to be insulting, profane, identity-attack, hateful, or toxic in another way (all toxicity sub-types provided by the annotators were collapsed to a single toxicity label) was calculated and used as a continuous toxicity score. This is similar to arrangements in the work of Wulczyn et al. (2017), who also found that training using

[1] https://jigsaw.google.com/
[2] http://nlp.cs.aueb.gr/
[3] Populous majority English-speaking countries were chosen. The most common country of origin was the USA.

the empirical distribution (over annotators) of the toxic labels (a continuous score per post) leads to better toxicity detection performance, compared to using labels re ecting the majority opinion of the raters (a binary label per post). See also Fornaciari et al. (2021).

Fig. 3.1.: Length of parent/target posts in characters.

Fig. 3.2.: Distribution of length (in words) of target posts.

Fig. 3.3.: Distribution of length (in words) of parent posts.

CCC contains 10,000 posts for which both context-aware annotations (in-context or ic) and context-unaware annotations (out-of-context or oc) are available. Figure 3.4 shows the number of posts (Y axis) per ground truth toxicity score (X axis). Orange (dashed) represents the ground truth obtained by annotators who were provided with context (the parent post) when rating ic), while blue is for annotators who rated the post without context (oc). The vast majority of the posts were unanimously perceived as non-toxic (0.0 toxicity), both by the oc and the ic coders, showing that CCC is a heavily imbalanced dataset. However, when context was provided to the annotators (ic coders), fewer posts were found with toxicity greater than 0.2, compared to annotators who did not have access to any context (oc coders). This is consistent with the findings of Pavlopoulos, Sorensen, et al. (2020a), where the authors observed that when the parent post is provided, the majority of the annotators perceive fewer posts as toxic, compared to showing no context to the annotators. To study this further, in this work we compared the two annotation scores ic, oc) per post, as discussed in the next section.

3.3 Defining Context Sensitivity

For each post p, we define $s^{ic}(p)$ to be the toxicity score (fraction of coders who perceived the post as toxic) derived from the ic coders and $s^{oc}(p)$ to be the toxicity derived from the oc coders. Then, their difference is $(p) = s^{oc}(p) \quad s^{ic}(p)$. A positive means practically that context mitigated the perceived toxicity of the annotators for this post, while a negative means that the context amplified the perceived toxicity of the annotators. Figure 3.5 shows that is most often 0, but when the toxicity score changes, is most often positive. In numbers, in 66.1% of the posts the toxicity score remained unchanged while out of the remaining 33.9%, in 9.6% it increased (960 posts) and in 24.2% it decreased (2,408) when

Fig. 3.4.: Histogram (converted to curve) showing the distribution of toxicity scores according to annotators who were (c) or were not (oc) given the parent posts.

context was provided. By binarizing the ground truth we get a similar trend, but with the toxicity of more posts remaining unchanged (i.e., 94.7%).

Fig. 3.5.: Histogram of context sensitivity. Negative (positive) sensitivity means the toxicity increased (decreased) when context was shown to the annotators.

Figure 3.6 shows that when counting the number of posts for which $|s|$ exceeds a given threshold t, called *context-sensitive posts*, as t increases, the number of context sensitive posts decreases. This means that clearly context sensitive posts (e.g., in an edge case, ones that all oc coders found as toxic while all c coders found as non toxic) are rare. Some examples of target posts, along with their parent posts and s, are shown in Table 3.1.

Fig. 3.6.: Number of context-sensitive posts (| t), when varying the context-sensitivity threshold t.

parent of post p	post p	$s^{oc}(p)$	$s^{ic}(p)$	
Oh Don..... you are soooo predictable.	oh Chuckie you are such a tattle tale.	36.6%	80%	-43.4%
Oh Why would you wish them well? They've destroyed the environment in their country and now they are coming here to do the same.	They"? Who is they? Do all Chinese look alike to you? Or are you just revealing your innate bigotry and racism?	70%	0%	70%

Tab. 3.1.: Examples of context-sensitive posts in CC. Here $s^{oc}(p)$ and $s^{ic}(p)$ are the fractions of out-of-context or in-context annotators, respectively, who found the target post p to be toxic; and = $s^{oc}(p)$ $s^{ic}(p)$.

4 System Design and Implementation

In this section we describe the deep neural network architectures that were used for the experiments of this work. These systems can be divided into two categories. The first category comprises the systems that are context-unaware (i.e. they do not have any mechanism to integrate contextual information about the conversation). On the other hand we have the systems that have a context-aware mechanism in order to integrate context. As we discuss in section 5, some traditional machine learning algorithms were also used (e.g. Linear Regression, Support Vector Regression etc.) but in this chapter we do not discuss them, as they are very well known.

4.1 Context-Unaware Architectures

BERTr Our first context-unaware model is a BERT (Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) regressor. BERT is pre-trained (to predict masked tokens and whether or not two sentences are consecutive or not) on a large corpus of text documents from Wikipedia, and uses Transformers (Vaswani et al., 2017). Using transfer learning, typically adding just a task-specific dense layer on top of the pre-trained model and training (" ne-tuning") on task-specific training instances, BERT is able to get state-of-the-art results on many NLP tasks. We ne-tuned BERT on the training subset of each experiment, with a task-specific regressor on top, fed with BERT's top-level embedding of the [CLS] token, which is intended to represent the entire input text. We used BERT-BASE pre-trained on cased data, with 12 layers and 768 hidden units and 110M parameters in total. We only unfreezed the top three layers during ne-tuning, with a small learning rate (2e-05) to avoid catastrophic forgetting. The task-specific regressor is a feed-forward neural network (FFNN) that consists of a dense layer (128 neurons) and a tanh activation function, followed by another dense layer (see figure 4.1). The last dense layer has a single output neuron, with no activation function, that produces the context sensitivity score.

PERSPECTIVE Our second context-unaware model is PERSPECTIVE. It is a CNN-based model created by Jigsaw and Google's Counter Abuse Technology team for toxicity detec-

tion. It is trained on millions of user comments from online publishers and conversations. It is publicly available through the PERSPECTIVE API.[1]

Fig. 4.1.: BERTr architecture.

4.2 Context-Aware Architectures

CA SEP BERT Our first context-aware model is CA SEP BERT. It is a BERT-based model with a simple context-aware mechanism added, and the same task-specific regressor as in the simple BERTr model. This model does not use a separate encoder for the parent post (context), however it concatenates the text of the parent and target posts, separated by BERT's [SEP] token, as in BERT's next sentence prediction pre-training task (see Fig. 4.2). We used the training subset to fine-tune the model.

PCT BERT Our second context-aware model is PCT BERT. It is a BERT-based model with a more complex context-aware mechanism added, and the same task-specific regressor as in the simple BERTr model. This model uses a separate encoder for the parent post (context), then the 2 representations of the [CLS] tokens (parent and target) are concatenated and passed to a similar FFNN as in the simple BERTr model (see Figure 4.3). We used the training subset to fine-tune the model.

[1] https://www.perspectiveapi.com/

Fig. 4.2.: Ca-SEP-BERT architecture.

Fig. 4.3.: PcT BERT architecture.

5 Experiments

5.1 Evaluation Metrics

In this section we describe the evaluation metrics that were used in our experiments. Our systems are regressors but we evaluate them also as classi ers, therefore we use both regression and evaluation metrics that we describe below.

5.1.1 Regression Metrics

Mean Squared Error Our rst regression metric is the mean squared error (MSE). This measure indicates how close a regression line is to a set of points. It takes the distances of the points to the regression line (these distances are the errors) and squares them. Because of this squaring term, MSE is sensitive to outliers (i.e. it gives more weight to larger di erences). The lower the MSE, the better the predictions of the regressor.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i \quad y_i)^2 \tag{5.1}$$

Equation 5.1 presents the mathematical formula of MSE, where y_i is the ground truth and y_i is the model's prediction.

Mean Absolute Error Our second regression metric is the mean absolute error (MAE). This measure also indicates how close a regression line is to a set of points. It takes the absolute distances of the points to the regression line (these distances are the errors) without squaring them. MAE is not sensitive to the outliers but is a metric more comprehensible and easier to interpret. The lower the MAE, the better the predictions of the regressor.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} jy_i \quad y_i j \tag{5.2}$$

Equation 5.2 presents the mathematical formula of MSE, where y_i is the ground truth and y_i is the model's prediction.

5.1.2 Classi cation Metrics

Area Under Precision-Recall Curve (AU PR) Our rst classi cation metric is the area under the precision-recall curve (AU PR) score. Here we used the total area under the precision-recall curve (Davis et al., 2006). The precision-recall curve shows the trade-o between precision and recall for di erent threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

Precision The precision of the positive class is a measure to evaluate how the model actually performs on predicting the positive class. It is the fraction of the number of comments that the model classi ed as positive and they actually were (also known as true positives), divided by the number of the total comments that the model classi ed as positives.

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

Recall also known as sensitivity quanti es the number of true positive class predictions made by the model out of all positive examples in the dataset. It is the fraction of the total amount of relevant instances that were actually retrieved.

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

ROC AUC Our second classi cation metric is the ROC area under the curve (AUC) score. Here we used the total area under the ROC curve (AUC) (Bradley, 1997). This is a standard classi cation metric that gives the performance of a binary classi er averaged over all possible trade-o s between true positive predictions and false positive predictions. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection. The false-positive rate is also known as probability of false alarm and can be calculated as 1 - speci city, where speci city is the recall of the negative class.

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \tag{5.5}$$

$$\text{False Positive Rate } (FPR) = \frac{FP}{FP + TN} \tag{5.6}$$

Figure 9: Illustration of the ROC curve. (Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc).

5.2 Experimental Study

Initially, CCC was used to experiment with existing toxicity detection systems, to investigate if context-sensitive posts are more difficult to automatically classify correctly as toxic or non-toxic. Then, new deep learning and machine learning systems were trained trying to solve a different task, that of estimating the context sensitivity of a target post (i.e. estimating how sensitive the toxicity score of each post is to its parent post).

5.2.1 Toxicity Detection

Figure 5.1 shows the MAE (see section 5.1) as a function of the context-sensitivity threshold, when employing the Perspective API toxicity detection system (see section 4.1), as is and with no further fine-tuning, to classify CCC posts as toxic or not.[1] Along with the Perspective model, a context-aware version of it was also employed, allowing it to see the parent post by concatenating it to the target post. In this experiment the context-aware (ic) gold labels were used as the ground truth since they are more reliable. Remember that the greater the sensitivity threshold, the smaller the sample (see figure 3.6).

[1] https://www.perspectiveapi.com

Fig. 5.1.: Mean Absolute Error (Y-axis) when predicting toxicity for different context-sensitivity thresholds (t; X-axis). We applied Perspective to target posts alone (w/o) or concatenating the parent posts (w).

A first observation is that when evaluating on all the CCC posts ($t = 0$) where the non context-sensitive posts dominate, the context-unaware Perspective (w/o curve) achieves lower error than the context-aware Perspective (w curve). On the other hand, when evaluating on smaller subsets with increasingly context-sensitive posts ($posts(t), t > 0$), and more specifically when $t > 0.2$, the context-aware variant of Perspective achieves lower error. Hence, the benefits of integrating context in toxicity detection systems may be visible only in subsets enriched with context-sensitive posts, like the ones we would obtain by evaluating (and training) on posts with $t \geq 0.2$. This explains related observations in previous work of Pavlopoulos, Sorensen, et al. (2020a), where the authors found that context-sensitive posts are too rare and, thus, context-aware models do not perform better on existing toxicity datasets. Another interesting observation is that the more we move to the right of Figure 5.1, the higher the error for both the context-unaware and the context-aware variants of Perspective. This is explained by the fact that Perspective is trained on posts that have been rated by annotators who were not provided with the parent post (out of context; oc), whereas here the in-context (ic) annotations were used as ground truth. Moreover, the greater the t in Figure 5.1, the larger the difference between the toxicity scores of oc and ic annotators, hence the larger the difference between the (oc) ground truth that Perspective saw during its training and the ground truth that was used here (ic).

The solution to the problem of the increasing error as context sensitivity increases (Figure 5.1) would be to train toxicity detectors on datasets that are sufficiently rich in context-sensitive posts. However, such posts are rare (Figure 3.6) and thus, they are hard to collect and annotate. This observation motivated the experiments of the next section, where context-sensitivity detectors are trained, which allow one to collect posts that are likely

to be context-sensitive. These posts can then be used to train toxicity detectors on datasets richer in context-sensitive posts.

5.2.2 Context Sensitivity Estimation

In this experiment, four regressors were trained and assessed on the CCC dataset, to predict the context-sensitivity (see section 3.3). Three traditional machine learning algorithms were used: Linear Regression, Support Vector Regression (Drucker et al., 1996), and a Random Forest regressor (Ho, 1995). Moreover a BERT-based (deep learning) regression model (BERTr see section 4.1) was used. The first three regressors used tf-idf features (Manning et al., 2008). All the models of this section use only the target post, because preliminary experiments showed that adding simplistic and naive context-mechanisms (e.g., concatenating the parent post) to the context sensitivity regressors does not lead to improvements. This may be due to the fact that it is often possible to decide if a post is context-sensitive or not by considering only the target post without its parent (e.g., in responses like YES!!). Future work will investigate this hypothesis further by experimenting with more elaborate context-mechanisms. If the hypothesis is verified, manually annotating context-sensitivity (not toxicity) may also require only the target post.

	MSE#	MAE #	AUPR"	AUC "
B1	2.3 (0.1)	11.56 (0.2)	12.69 (0.7)	50.00 (0.0)
B2	4.6 (0.0)	13.22 (0.1)	13.39 (0.8)	50.01 (1.6)
LR	2.1 (0.1)	11.0 (0.3)	30.11 (1.2)	71.67 (0.8)
SVR	2.3 (0.1)	12.8 (0.1)	28.66 (1.7)	71.56 (1.0)
RFs	2.2 (0.1)	11.2 (0.2)	21.57 (1.0)	59.67 (0.3)
BERTr	1.8 (0.1)	9.2 (0.3)	42.01 (4.3)	80.46 (1.3)

Tab. 5.1.: Mean Squared Error (MSE), Mean Absolute Error (MAE), Area Under Precision-Recall curve (AUPR), and ROC AUC of all context sensitivity estimation models. An average (B1) and a random (B2) baseline have been included. All results averaged over three random splits, standard error of mean in brackets.

For the experiments a train/validation/test split of 80/10/10 percent, respectively was used, and a Monte Carlo 3-fold Cross Validation was performed. MSE was used as the loss function and early stopping with patience of 5 epochs was used. Table 5.1 presents the MSE and MAE of all the models on the test set. Unsurprisingly, all the traditional machine learning models were outperformed by BERTr in MSE and MAE. Previous work (Wulczyn et al., 2017) reported that training toxicity regressors (based on the empirical distribution of codes) instead of classifiers (based on the majority of the codes) leads to improved classification results too, so we also computed classification results. For the latter results, the ground truth probabilities of the test instances were turned to binary labels by setting a threshold t (section 3.3) and assigning the label 1 if $\geq t$ and 0 otherwise. In

this experiment, t was set to the sum of the standard error of mean $sem(\cdot)$ of the oc and ic raters for that specific post p, i.e., $t(p) = sem^{oc}(p) + sem^{ic}(p)$. By binarizing the ground truth, AUPR and AUC (Table 5.1) verified that BERTr outperforms the other models, even when the models are used as classifiers.

5.3 Collecting Context Sensitive Posts

In Section 3 we saw that context sensitive posts can be very rare in toxicity datasets (Figure 3.6) and thus it can be hard to collect and annotate them. Section 5.2 described how the integration of a simple context-aware mechanism (concatenating the parent post to the target post) to an existing toxicity detection system can reduce the system's error when evaluating on context-sensitive posts (Figure 5.1). However, the error remains at a high level for context-sensitive posts. This problem can potentially be addressed by augmenting the current datasets with more context-sensitive posts. As shown in Section 5.2, a regressor trained to predict the context sensitivity of a post can achieve low error (Table 5.1). Hence, the scenario where a context sensitivity regressor is employed to obtain a dataset richer in context-sensitive posts was assessed.

In this scenario, the best context-sensitivity regressor (BERTr) was used in order to retrieve the 250 most likely context-sensitive posts from the 2M CC posts, excluding the 10,000 CCC posts. Then, these posts were crowd-annotated by annotators who had access to the parent post (c), but the out-of-context (oc) annotations were also kept. Table 5.2 shows examples of the 250 target posts obtained, along with their parent posts and Δ. Then the same experiment was repeated, this time using 250 randomly selected posts from the 2M CC posts, excluding the 10,000 CCC posts and the 250 posts that were selected using BERTr. Figure 5.2 is the same as Figure 3.6, but now we consider the 250 randomly selected posts (dashed line) and the 250 posts that were selected using BERTr (solid line). As in Figure 3.6, the context-sensitivity threshold t is varied on the horizontal axis. It is clearly shown that the 250 posts that were sampled using BERTr include more context-sensitive posts than the 250 random ones, with the threshold (t) in the range $0.1 \leq t < 0.7$. This indicates that BERTr can be successfully used to obtain datasets richer in context-sensitive posts. As in Figure 3.6, there are very few context-sensitive posts for $t \geq 0.7$.

In this experiment, the crowd-annotators were also asked to indicate whether the parent post was helpful or not, during the annotation process regarding the toxicity of each target post. Figure 5.3 shows for how many of the 250 target posts (sampled using BERTr or random) the majority of the annotators responded that the parent post was useful. The sensitivity threshold (t) is varied on the horizontal axis up to $t = 0.7$, since no posts are context-sensitive for $t > 0.7$ (Fig. 5.2). The perceived utility of the parent posts is clearly higher for the 250 posts sampled with BERTr, compared to the 250 random ones, for all

sensitivity thresholds. This again indicates that BERTr can be used to obtain datasets richer in context-sensitive posts.

parent of post p	post p	$s^{OC}(p)$	$s^{IC}(p)$	
And since Thomas Aquinas never observed animals having gay sex in the wild homosexuality never made it into the annals of natural law theory.	Animals having "gay' sex? You mean there are "gay" animals. So, when they're not "doing it" they do other things like go to Madonna concerts?	60%	0%	60%
Making a cake is MUCH di erent then selling gasoline or renting hotel rooms. Making a cake is a form of artistry and requires the cake maker to artistically express him/herself which means the cake maker is actively participating. Owning a gas station where random people pump their own gas does not require active participation.	Oh, ok. So the if the gas guy had to pump gas for that gay man, he should be able to refuse that, right?	60%	0%	60%
And SCOTUS will slap Watson & Chin down yet again....these Odummy Sock Puppets never learn. That threesome they shared back in the day must have been amazing.	Is the post implying that the judge is gay? I don't understand the comment, please explain? Are gays involved in this and not Muslims and their relatives? *	83.3%	20%	63.3%
The appeal courts have one thing to do, ask is it legal or not, thats it, that is what appeals judges do, and they didnt, they coward away cause they knew then could not rule it illegal. sorry for your ignorance	The case has not yet been adjudicated on its merits (whether the Executive Order is illegal or not). Both the trial decision and the appeal decision were about staying the EO *until the trial on its merits* - ie, an injuction. I'd think about nding out some facts before calling someone else ignorant, Rex.	80%	20%	60%
"...."marriage," by de nition, meant one man, one woman ..." Actually no. The de nition restricting it to one man one woman unions was only introduced into USA law 2004/5/6 across numerous states in a frantic attempt to avoid courts making similar ndings to those of the Massachusetts Supreme court ruling. Prior to that it had always been expressly de ned as between "two people", which is what triggered the Massachusets challenge. The fact that only opposite sex marriages were performed in the past, does not mean marriage was de ned as only between opposite sex couples, it simply illustrates that couples who were not opposite sex were being denied a fundamental right. The evidence of the existence of discrimination is not proof that the discrimination was justi ed or justi able.	The de nitive dictionary of the English language, the OED, does not contain a single instance in which "modern" civilized society has included gay marriage. It does mention instances of "group" marriage in small, primitive societies, where all the men in a village are married to all the women. But those, as you know, are by far the exception. Actually, your argument bolsters my point. It was so universally understood at the founding of the Nation that marriage meant man-woman that marriage did not need to be de ned. Indeed, in most States, marriage could not have been de ned so as to allow gay marriage, because until 1961, ALL 50 STATES outlawed sahdemy. Do you begin to get at least part of the point?	0%	60%	-60%
May be Trudeau should do a double apology just to one up Harper and then apologize for Papa Trudeau and no himself ruining the Canadian economy.	What has this got to do with the rape and abuse of boys and girls in residential schools?	30%	60%	-30%

Tab. 5.2.: Examples of context-sensitive posts in the sampled dataset. Here, $s^{OC}(p)$ and $s^{IC}(p)$ are the fractions of out-of-context or in-context annotators, respectively, who found the target post p to be toxic; and $= s^{OC}(p) \quad s^{IC}(p)$.

Fig. 5.2.: Number of context-sensitive posts (t), for di erent context-sensitivity thresholds (t), using 250 likely context-sensitive posts sampled with BERTr (solid) or 250 randomly selected posts (dashed line).

Fig. 5.3.: Percentage of the 250 target posts, sampled with BERTr (solid) or random (dashed line), for which the majority of annotators found the parent post useful when assessing the toxicity of the target post.

To verify the statistical signi cance of the nding that the annotators nd the parent post useful more often in posts sampled with BERTr than in random posts, a paired bootstrap resampling was performed, following the experimental setting of Koehn (2004). 100 posts from the 250 random posts, and 100 posts from the 250 posts obtained by using BERTr were sampled, and the percentage of posts where the majority of annotators found the parent post helpful, for random posts and BERTr posts was computed. By resampling 1,000 times, it was found that this percentage is greater for BERTr posts than for random posts, with a P -value of 0.05.

Finally, by turning the ground truth toxicity probabilities (for IC and OC annotation) into binary labels as in Section 5.2, a context sensitivity class ratio (fraction of context-sensitive posts out of all 250 posts) was estimated, for the BERTr-sampled and the randomly sampled posts. By using this class ratio, it was found that 99 out of the 250 BERTr-sampled posts (39.6 %) were context sensitive, while only 43 out of the 250 randomly sampled posts (17.2 %) were context-sensitive (22 percent points lower; i.e., 57% decrease). The statistical signi cance of this nding (lower fraction) was veri ed by using bootstrapping with a P -value of 0.05, as in the previous paragraph. Therefore, sampling with BERTr leads to a higher context-sensitivity class ratio than random sampling.

5.4 Improving the Context-Sensitivity Regressor with Data Augmentation

In the previous section it was shown that a context sensitivity regressor (BERTr was the best one) can be employed in order to sample new sets of posts (e.g., from the 2M CC posts) that are richer in context-sensitive posts (by 22 percent points in our previous experiments). Adding such richer (in context-sensitive posts) sets to an existing context sensitivity dataset (e.g., the CCC dataset), can lead to increment of the ratio of context-sensitive posts (which is low in CCC, see Fig. 3.6). A logical question then is if the context-sensitivity regressor can further be improved by re-training it on the augmented dataset, which is less dominated by context-insensitive posts (more balanced in terms of context-sensitivity). Ideally the newly sampled (and overall more context-sensitive) posts would be crowd-annotated for context-sensitivity (by IC and OC raters) to obtain ground truth (gold context-sensitivity scores). To avoid this additional annotation cost, however, in this section a teacher-student approach (Hinton et al., 2015) is explored. The teacher is the initial BERTr (experiments with CA SEP BERT and PcT BERT as teachers were also performed, see appendix A) context-sensitivity regressor (section 5.2.2), which provides silver context-sensitivity scores for the newly sampled posts. The student is another BERTr instance, which is trained on the augmented dataset (the data with gold sensitivity scores the teacher was trained on, plus the newly sampled posts with silver sensitivity scores).

Following this teacher-student approach, experiments with data augmentation to improve the context-sensitivity estimator, using two di erent settings were performed. In both settings, the teacher silver-scores the newly sampled additional training posts. In the setting discussed rst, the teacher is also used to sample the new training posts. By contrast, in the second setting the new posts are randomly sampled, and the teacher is only used to silver-score them.

Teacher-student with teacher sampling: In this setting, 20,000 posts were randomly sampled from the Civil Comments (CC) dataset and were used as a pool to select (and silver-score) new training instances from, as follows:

1. Train a BERTr teacher on the gold-scored (by crowd-annotators) training instances of our CCC dataset (Section 3).

2. Use the BERTr teacher to silver-score for context-sensitivity all the posts of the pool (initially 20,000).

3. Select from the pool the 1,000 posts with the highest silver sensitivity scores, remove them from the pool, and add them (with their silver sensitivity scores) to the training set.

4. Train a BERTr student on the new training set (augmented by 1,000 silver-scored posts).

5. Evaluate the student using exactly the same splits as in Section 5.2.2.

6. (Optional) Go back to step 2, using the student as a new teacher in a new cycle.

This process was repeated for ve cycles. Thus, the training set was augmented by 5,000 likely context-sensitive posts. Experimental results (Fig. 5.4, blue solid line) show performance gains in MSE even from the rst cycle. Sampling using bootstrapping was also compared against using a single cycle with 5,000 new posts added at once (blue dashed line), instead of adding only 1,000 posts per cycle and re-training the teacher. Performing cycles and re-training the teacher clearly leads to lower MSE, but with diminishing returns after cycle 4.

Teacher-student with random sampling: This setting is the same as the previous one, but in step 3 we randomly select 1,000 posts from the pool, instead of selecting the 1,000 posts with the highest silver sensitivity scores. Again, ve cycles were used (Fig. 5.4, orange solid line) and also sampling using bootstrapping was compared to a single cycle that adds 5,000 silver-scored training instances at once (orange dashed line). Sampling with the teacher's scores (blue solid line) is clearly better than random sampling (orange lines).

Fig. 5.4.: Data augmentation with knowledge distillation to improve BERTr context-sensitivity regressor. Blue solid line: the teacher model is used both to silver-score the new training instances and to sample them. Orange solid line: the teacher model is used only to silver-score the new training instances, which are randomly selected. Dashed lines: same as the solid ones, but only one cycle is performed, which adds 5,000 silver-scored new training instances at once.

6 Conclusions and Future Work

This thesis presented a new context-aware dataset that was created by Google Jigsaw and the AUEB NLP group, containing both context-aware and context unaware annotations. It was also shown that existing toxicity detection systems perform worse on context-sensitive posts, but by integrating even a simple and naive context-aware mechanism (e.g. concatenating the parent post to the target), one can reduce the error. Moreover, this work introduced a new task, that of estimating the context-sensitivity of posts in toxicity detection, i.e., estimating the extent to which the perceived toxicity of a post depends on the conversational context. Experiments on this task using traditional machine learning algorithms and deep learning models, showed that context-sensitivity estimation systems of practical quality can be developed, achieving low error. Moreover, it was also shown that context-sensitivity estimation systems can be used to collect larger samples of context-sensitive posts, which is a prerequisite to train toxicity detectors to better handle context-sensitive posts. Furthermore, the best performing system BERTr and 2 other context-aware systems (CA SEP BERT and PcT BERT), can further be improved by augmenting the training dataset with knowledge distillation using a teacher-student model with teacher sampling. Furthermore, context-sensitivity estimators can also be used as a suggestion tool to advise when moderators should consider the context of a post (parent), which is more costly and may not always be necessary.

In future work we will investigate if it is often possible to decide if a post is context-sensitive or not by considering only the target post without its parent (e.g. in responses like YES!!). We leave for future work the implementation of more complex context-aware models, as well as training (and evaluate) them on datasets su ciently rich in context-sensitive posts. We also leave for future work the study of other types of context, such as the title of the thread, the entire thread or personal information about the author of the target post, or a combination of them. Finally, it would be interesting to see if toxicity detectors trained on more context-sensitive posts indeed perform better on unseen context-sensitive posts.

Bibliography

[AA18] Sweta Agrawal and Amit Awekar. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms . In: ArXiv abs/1801.06482 (2018).

[Aga+15] Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg. Sentiment Analysis Using Common-Sense and Context Information . In: Computational intelligence and neuroscience 2015 (Apr. 2015), p. 715730.

[Bad+17] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep Learning for Hate Speech Detection in Tweets . In: Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion (2017).

[BH16] Sven Buechel and Udo Hahn. Emotion Analysis as a Regression Problem Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation . In: Proceedings of the Twenty-Second European Conference on Arti cial Intelligence. ECAI'16. The Hague, The Netherlands: IOS Press, 2016, pp. 1114 1122.

[BKR21] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. A Survey on Data Augmentation for Text Classi cation . 2021. arXiv:2107.03158 [cs.CL] .

[BM20] Esther van den Berg and Katja Markert. Context in Informational Bias Detection . In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6315 6326.

[Bor+19] Daniel Borkan, Lucas Dixon, Je rey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classi cation . In: WWW. San Francisco, USA, 2019, pp. 491 500.

[Bra97] A. P. Bradley. The use of area under the ROC curve in the evaluation of machine learning algorithms . In: Pattern Recognition 30.7 (1997), pp. 1145 1159.

[CGM19] Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. Pay Attention to your Context when Classifying Abusive Language . In: Proceedings of the Third Workshop on Abusive Language Online. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 70 79.

[Chi+20] Patricia Chiril, Véronique Moriceau, Farah Benamara, et al. An Annotated Corpus for Sexism Detection in French Tweets . English. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 1397 1403.

[Dav+17] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of O ensive Language . In: Proceedings of the International AAAI Conference on Web and Social Media 11.1 (May 2017).

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . 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). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171 4186.

[DG06] Jesse Davis and Mark Goadrich. The Relationship between Precision-Recall and ROC Curves . In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 233 240.

[DIF20] Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN . English. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 21 25.

[DJ16] Ond°ej Du²ek and Filip Jur£ £ek. A Context-aware Natural Language Generator for Dialogue Systems . In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles: Association for Computational Linguistics, Sept. 2016, pp. 185 190.

[Dju+15] Nemanja Djuric, Jing Zhou, Robin Morris, et al. Hate Speech Detection with Comment Embeddings . In: Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, 2015, pp. 29 30.

[Dru+96] Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support Vector Regression Machines . In: Proceedings of the 9th International Conference on Neural Information Processing Systems. NIPS'96. Denver, Colorado: MIT Press, 1996, pp. 155 161.

[ElS+18] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. 2018. arXiv: 1804.04257 [cs.CL] .

[Fen+21] Steven Y. Feng, Varun Gangal, Jason Wei, et al. A Survey of Data Augmentation Approaches for NLP. 2021. arXiv:2105.03075 [cs.CL] .

[For+21] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, et al. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning . In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021, pp. 2591 2597.

[Fou+18] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, et al. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. 2018.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. http://www.deeplearningbook.org . MIT Press, 2016.

[GH17a] Lei Gao and Ruihong Huang. Detecting Online Hate Speech Using Context Aware Models . In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. Varna, Bulgaria: INCOMA Ltd., Sept. 2017, pp. 260 266.

[GH17b] Yoav Goldberg and Graeme Hirst. Neural Network Methods in Natural Language Processing. Morgan Claypool Publishers, 2017.

[Had+20] Bushr Haddad, Zoher Orabe, Anas Al-Abood, and Nada Ghneim. Arabic O ensive Language Detection with Attention-based Deep Neural Networks . English. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on O ensive Language Detection. Marseille, France: European Language Resource Association, May 2020, pp. 76 81.

[Ho95] Tin Kam Ho. Random decision forests . In: Proceedings of 3rd International Conference on Document Analysis and Recognition. Vol. 1. 1995, 278 282 vol.1.

[HVD15] Geo rey Hinton, Oriol Vinyals, and Je Dean. Distilling the Knowledge in a Neural Network. 2015. arXiv: 1503.02531 [stat.ML] .

[IB18] Muhammad Okky Ibrohim and Indra Budi. A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media . In: Procedia Computer Science 135 (2018). The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life, pp. 222 229.

[Koe04] Philipp Koehn. Statistical signi cance tests for machine translation evaluation . In: Proceedings of the 2004 conference on empirical methods in natural language processing. 2004, pp. 388 395.

[Low+15] R. Lowe, Nissan Pow, Laurent Charlin, and Joelle Pineau. Incorporating Unstructured Textual Knowledge Sources into Neural Dialogue Systems . In: 2015.

[MAT21] Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection. 2021. arXiv: 2103.14916 [cs.CL] .

[MDM17] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive Language Detection on Arabic Social Media . In: Proceedings of the First Workshop on Abusive Language Online Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 52 56.

[MGD16] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning Generic Context Embedding with Bidirectional LSTM . In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 51 61.

[Mik+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Je rey Dean. Distributed Representations of Words and Phrases and Their Compositionality . In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111 3119.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval Cambridge University Press, 2008.

[Nob+16] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive Language Detection in Online User Content . In: Proceedings of the 25th International Conference on World Wide Web WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 145 153.

[Ozl+20] Kadir Bulut Ozler, Kate Kenski, Steve Rains, et al. Fine-tuning for multi-domain and multi-label uncivil language detection . In: Proceedings of the Fourth Workshop on Online Abuse and Harms Online: Association for Computational Linguistics, Nov. 2020, pp. 28 33.

[Pav+17] John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. Improved Abusive Comment Moderation with User Embeddings . In: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 51 55.

[Pav+20a] John Pavlopoulos, Je rey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity Detection: Does Context Really Matter? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics Online: Association for Computational Linguistics, July 2020, pp. 4296 4305.

[Pav+20b] John Pavlopoulos, Je rey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity Detection: Does Context Really Matter? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics Online: Association for Computational Linguistics, July 2020, pp. 4296 4305.

[Pet+18] Matthew Peters, Mark Neumann, Mohit Iyyer, et al. Deep Contextualized Word Representations . In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227 2237.

[PF17] Ji Ho Park and Pascale Fung. One-step and Two-step Classi cation for Abusive Language Detection on Twitter . In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 41 45.

[PMA17a] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deep Learning for User Comment Moderation . In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 25 35.

[PMA17b] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper Attention to Abusive User Content Moderation . In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1125 1135.

[PN18] Natalie Parde and Rodney Nielsen. Exploring the Terrain of Metaphor Novelty: A Regression-Based Approach for Automatically Scoring Metaphors . In: Proceedings of the AAAI Conference on Arti cial Intelligence 32.1 (Apr. 2018).

[PSM14] Je rey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation . In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532 1543.

[Ren+16] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. Context-Sensitive Twitter Sentiment Classi cation Using Neural Network . In: Proceedings of the Thirtieth AAAI Conference on Arti cial Intelligence. AAAI'16. Phoenix, Arizona: AAAI Press, 2016, pp. 215 221.

[Ros+16] Björn Ross, Michael Rist, Guillermo Carbonell, et al. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis . In: Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication. Ed. by Michael Beiÿwenger, Michael Wojatzki, and Torsten Zesch. 2016, pp. 6 9.

[SK19] Connor Shorten and Taghi Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning . In: Journal of Big Data 6 (July 2019).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need . 2017. arXiv: 1706.03762 [cs.CL] .

[VCB14] Andrea Vanzo, Danilo Croce, and Roberto Basili. A context-based model for Sentiment Analysis in Twitter . In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 2345 2354.

[Wan+16] Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model . In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 225 230.

[Was+17] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding Abuse: A Typology of Abusive Language Detection Subtasks . In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 78 84.

[WH16] Zeerak Waseem and Dirk Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter . In: Proceedings of the NAACL Student Research Workshop. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88 93.

[WSR18] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the GermEval 2018 Shared Task on the Identi cation of O ensive Language . In: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria September 21, 2018. Sept. 2018.

[WTD17] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale . In: Proceedings of the 26th International Conference on World Wide Web - WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1391 1399.

[Zam+19] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, et al. Predicting the Type and Target of O ensive 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). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1415 1420.

[ZRT18] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network . In: The Semantic Web. Ed. by Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, et al. Cham: Springer International Publishing, 2018, pp. 745 760.

A

Additional Details Of The Experiments with Data Augmentation

This appendix presents some additional results regarding the experiments with data augmentation (section 5.4). Experiments when using as a teacher (and student) a CA-SEP-BERT as well as a PcT BERT model (see section 4.1), that are context-aware models were performed. Moreover, the performance of the models when evaluating with MSE, MAE, AUPR and ROC AUC (as in section 5.2.2) is examined.

A rst observation is that the CA SEP BERT model has the worst performance among all the other models when evaluating with all four metrics. This is probably due to the fact that this model uses a max sequence length of 128 tokens. This practically means that it uses 64 tokens for the parent post and 64 tokens for the target post, while BERTr and PcT BERT use 128 and 256 tokens (128 for the target and for 128 for the parent post) respectively. Therefore, we cannot make a fair comparison between these three models. A second observation is that BERTr achieves the best performance (in most of the cases) indicating that often it may be possible to decide if a post is context-sensitive or not by considering only the target post without its parent. In order to verify this, more complex context-aware models need to be developed and to get evaluated. A more interesting observation is that data augmentation with knowledge distillation leads to performance gains in all four metrics even from the rst cycle regardless which model plays the role of the teacher. Moreover, performing cycles and re-training the teacher clearly leads to better performance in most of the cases. Finally, sampling with the teacher's scores (blue, green and purple solid lines) is clearly better than random sampling (orange, red and brown lines) when evaluating with MSE, AUPR and ROC AUC but this does not apply when evaluating with MAE.

Fig. A.1.: MSE scores for data augmentation with knowledge distillation to improve all context-sensitivity regressors. Blue, green and purple solid lines: the teacher model is used both to silver-score the new training instances and to sample them. Orange, red and brown solid lines: the teacher model is used only to silver-score the new training instances, which are randomly selected. Dashed lines: same as the solid ones, but only one cycle is performed, which adds 5,000 silver-scored new training instances at once.

Fig. A.2.: MAE scores for data augmentation with knowledge distillation to improve all context-sensitivity regressors. Blue, green and purple solid lines: the teacher model is used both to silver-score the new training instances and to sample them. Orange, red and brown solid lines: the teacher model is used only to silver-score the new training instances, which are randomly selected. Dashed lines: same as the solid ones, but only one cycle is performed, which adds 5,000 silver-scored new training instances at once.