Universität Regensburg MaxS at GermEval 2021 Task 1: Synthetic Data in Toxic Comment Classification

Maximilian Schmidhuber
Universität Regensburg / Fakultät für Sprache, Literatur und Kultur
maximilian.schmidhuber@stud.uni-regensburg.de

Abstract

We report on our submission to Task **1** of the GermEval 2021 challenge – toxic comment classification. We investigate different ways of bolstering scarce training data to improve off-the-shelf model performance on a toxic comment classification task. To help address the limitations of a small dataset, we use data synthetically generated by a German GPT-2 model.

The use of *synthetic data* has only recently been taking off as a possible solution to addressing training data sparseness in NLP, and initial results are promising. However, our model did not see measurable improvement through the use of synthetic data. We discuss possible reasons for this finding and explore future works in the field.

1 Introduction

In recent years, social media platforms have become an integral part of our everyday lives. Together with their enormous rise in use and popularity, they have also faced several troubles. These range from PR problems due to privacy concerns[1] to Fake News (Wells et al., 2019). There have also been incidents related to deplatforming controversial individuals of public interest[2].
In 2018, for instance, Facebook was used to incite a Genocide against the Rohingya people of Myanmar[3]. The ongoing global pandemic has seen an increase in xenophobic and antisemitic hate (Greenblatt, 2020). In the light of these and other developments, the task of detecting toxicity on the internet has seen increased attention in recent years. By now, the legislature of multiple countries is getting modified to accommodate laws counteracting the incitement of hatred online.
The German *Netzwerkdurchsetzungsgesetz* (network enforcement act), for instance, requires social media providers with over 2 million users registered in Germany to report hate speech on their platform to legal authorities[4]. These steps try to prevent the marginalization of populations. Users exposed to hate online may no longer take part in debates or discourse. This work aims to advance the state of the art in the field of toxicity detection by providing an additional avenue of working with scarce data. Any code used for this work is available on GitHub[5].

2 Related Work

There have been numerous developments in the space of toxicity detection since GermEval in 2019. Struß et al. (2019) give a well-composed overview of the state of affairs in 2019. We will therefore focus on recent work and concepts closely related to the challenge.

2.1 Recent Developments in Toxicity Detection

In the scope of the Shared Task, the concept of toxicity includes *"uncivil forms of communication that can violate the rules of polite behaviour, such as insulting discussion participants, using vulgar or sarcastic language or implied volume via capital letters"* (Risch et al., 2021). A more thorough definition of the annotation guidelines is provided by Risch et al. (2021).

In terms of 'traditional' hate speech detection, Mathew et al. (2020) proposed HateXplain. HateXplain is a new benchmark dataset for hate speech

[1] https://www.wired.com/story/facebook-privacy-ftc-changes/
[2] https://www.theatlantic.com/ideas/archive/2021/05/facebooks-trump-ban-effects/618818/
[3] https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html
[4] https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html
[5] https://github.com/khaliso/GermEval2021_submission

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

detection, which tries to factor in hate speech bias and interpretability aspects.
Another notable contribution was made by Rosenthal et al. (2020). They created a new dataset called *SOLID* using a semi-supervised learning approach using an ensemble of four different models. It is the largest available dataset in the field right now.
A recent work provided by Sheth et al. (2021) notes that context is of key importance for toxicity detection. As surrounding conversation would mitigate a potentially toxic comment, exchange history would inform the determination of toxicity. Therefore, an exchange history surrounding the potentially toxic comment is needed in the corpus. Sheth et al. (2021) also note a potential problem with current transformer-based state-of-the-art systems such as BERT and GPT-2/GPT-3 (**G**enerative **P**retrained **T**ransformer). These models are designed to predict the next token given previous tokens from the dataset they were trained on. As these datasets have been collected from the web, corpus bias and incidentally confounded features can result in models that may cause harm to individuals or society (Kursuncu et al., 2020; McGuffie and Newhouse, 2020). There are indications that BERT embeddings may have racist or toxic tendencies (Zhang et al., 2020). Solaiman et al. (2019) of *OpenAI* note that GPT-2 is capable of producing extremist text if trained on suitable data. However, machine-generated content detection tools such as *Grover* by Zellers et al. (2019) can spot GPT-2 generated content in most cases. They also note that *the skills and resources required for using language models, both beneficially and maliciously, will decrease over time* (Solaiman et al., 2019). GPT-3 by Brown et al. (2020) does develop in this direction.

2.2 Synthetic Data

Shu et al. (2020) note that limited labelled data is becoming the largest bottleneck for supervised learning systems. This is especially the case for many real-world tasks where large scale annotated examples can be too expensive to acquire. Therefore, they proposed a technique using semi-supervised learning; however, there have also been different approaches to face this task. For example, GPT-2 (Radford et al., 2019) is a Text Generation model created by *OpenAI* using the transformers architecture. Both GPT-2 and, more prominently, its successor GPT-3 by Brown et al. (2020) are most well-known for their ability to create text that is almost indistinguishable from text written by humans. Moreover, as Budzianowski and Vulić (2019) found, the model also *holds promise to mitigate the data scarcity problem*. With these recent advancements on the horizon, interest in Synthetic Data Generation has grown in many areas of research, including NLP. Works include generating synthetic data for Lexical Normalization (Dekker and van der Goot, 2020) or Neural Grammatical Error Correction Systems (Grundkiewicz et al., 2019). Recent works in the field of Toxic Comment Classification appear to be very promising, but this particular field of research is still very young (Juuti et al., 2020; Whitfield, 2021).
Synthetic data has implications besides its potential use in bolstering datasets: This approach poses a possible solution to ethical, security and privacy concerns with real datasets (Surendra and Mohan, 2017).

2.3 Related Challenges

In terms of state-of-the-art systems, other recent Shared Tasks can give a good overview:

1. GermEval-2019 Task 2: Identification of offensive language (Struß et al., 2019)

2. Kaggle, 2020: Jigsaw Multilingual Toxic Comment Classification[6]

3. SemEval-2020 Task 12: Multilingual offensive language identification in social media (Zampieri et al., 2020; Ranasinghe and Hettiarachchi, 2020)

4. SemEval 2021 Task 5: Toxic Spans Detection was more fine-grained than previous tasks, as participants in this shared task were asked to determine which span(s) of text in a post were responsible for the classification of the entire post (Pavlopoulos et al., 2021)

The best-performing systems in the field of toxicity detection online as found by Zampieri et al. (2020) were mainly based on XLM-RoBERTa, ALBERT or ERNIE 2.0 (Safaya et al., 2020). These are among the newest iterations of BERT-based models.
However, to train and fine-tune such models, large quantities of - preferably labelled - data are required.

[6]https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

3 Methodology

In this work, we investigate whether or not synthetically generated data can improve the baseline of a model solely trained on a scarce dataset.

3.1 Data analysis

The Dataset used for this task was provided by Risch et al. (2021) and consists of:

- 3244 Facebook comments with binary labels for each of the three tasks at hand. This work focusses on **Task 1:** *Toxic Comment Classification*.
- 2122 of the comments were labelled as 'non-toxic', and
- 1122 were labelled 'toxic'.
- The longest toxic comment had a length of 2035 tokens;
- The longest non-toxic comment had 2833 tokens.

The Dataset is drawn from the Facebook page of a political talk show of a German television broadcaster and includes user discussions from February-July 2019.
The Dataset is anonymized by not sharing comment IDs and user information. Furthermore, links to users are replaced by @USER and links to the show replaced with @MEDIUM. Links to the show's moderator are replaced with @MODERATOR.
This raises a couple of challenges:

1. The **first** challenge is the anonymized nature of the dataset, making it impossible to take context into account (Sheth et al., 2021).
2. The **second** challenge we face is the relatively small amount of data available. This issue is well-known (Shu et al., 2020) and not exclusive to the field of NLP.

The **first** challenge, lack of context. A rudimentary form of context can be constructed, as the dataset provides the *@USER*, *@MODERATOR* and *@MEDIUM* tags. It is, however, not possible to extract more fine-grained types of relationships between commenters. For example, a comment determined to be potentially toxic can be both harmless bickering within common groups and a toxic remark to outsiders. To more accurately determine the correct label, relationship data is required (Sheth et al., 2021). However, supplying the data necessary to create these relationships would raise severe privacy implications.

Therefore, the main variable we are ethically and technically able to influence is the **second** challenge, the size of the dataset. There have been several ways in the past to bolster a dataset.

3.2 Bolstering the Dataset

The standard approach, even in the history of the GermEval shared task, is to use additional, related datasets (Paraschiv and Cercel, 2019). There are a number of options:

1. **GermEval 2018 Dataset** by Wiegand et al. (2018) contains 8541 labeled offensive German tweets with an inter-author agreement of k = 0.66.
2. **GermEval 2019 Dataset** by Struß et al. (2019) contains two separate datasets. The training Dataset for Task 1 and 2 (binary and fine-grained classification) consists of 3995 annotated German offensive tweets, while the Dataset provided for Task 3 (explicit or implicit offensive language classification) consists of 1958 annotated German tweets
3. **German Federal Election Dataset** by Kratzke (2017), containing 1.212.220 unlabeled tweets crawled around the German Federal Election 2017
4. **OLID** by Zampieri et al. (2019) contains over 14.000 labeled English tweets
5. **SOLID** by Rosenthal et al. (2020) contains over 9 million English tweets labeled in a semi-supervised manner

The main issue we face is that the most available datasets for toxicity detection online are English.
Another method of bolstering datasets has been explored for the task of object detection on images by augmenting available data by rotating images or similar modification methods (Zoph et al., 2020). A comparable approach for the field of NLP is the [MASK] token used by Devlin et al. (2018) for BERT, the current gold-standard model for a wide variety of NLP tasks.

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

'Ziemlich traurig, das ganze Nachrichten zu einem zweiten Geburtstag.(1) The President of the Federal Republic of Germany shall be elected for a four-year term by the Bundestag on the basis of proportional representation by direct universal suffrage. (2) Die Bundesrat der Bundesregierung ist ein Bundesgesetz über eine gesetzliche Verfassungsgerichtshof, die durch die Bundesversammlung ausgeführt werden, soweit sie in dem Bundesverwaltung des Bundesministeriums und des Landesministers aufgehoben wird. Diese Fähigkeit wurde darüber häufig zur Verwandteilung der Fachberechtigung eines Bundesrates'

Figure 1: Output of GPT-2 fine-tuned on toxic comments from the dataset

With advancements in text generation models, another avenue of research has opened up: Synthetic Data.

3.3 Selecting the Data Generation Model

An initial evaluation, as seen in Figure 1, revealed that text generated by a fine-tuned GPT-2 included both English and German phrases. Fine-tuned German GPT-2 (gGPT-2) by Schweter (2020) on the other hand created German-sounding, yet incoherent sentences (Figure 2). gGPT-2 was fine-tuned on the normalized version of *Faust I and II* by Johann Wolfgang von Goethe. The model has not yet been used in the reviewed literature. However, initial experiments on German recipes[7] and German medical reviews appeared promising[8]. The initial reasoning was that the more German-sounding text created by gGPT-2 could benefit the trained system, as coherence might be less relevant on the token level if the readable text was also part of the finished dataset.
GPT-3 was not an option for this work, as we did not get access to the API in time.

3.4 Data Generation Model Description

The German GPT-2 Model is part of Huggingface's *transformers* library. We selected a batch size of 16, and set the maximum sequence length to 1024. Similar to the approach used by Whitfield (2021),

[7]https://towardsdatascience.com/fine-tune-a-non-english-gpt-2-model-with-huggingface-9acc2dc7635b

[8]https://data-dive.com/finetune-German-gpt2-on-tpu-transformers-tensorflow-for-text-generation-of-reviews

'Woche in die Sozial für Deutschen müssig wicht in eingeword und dann sich abende sind kontraum den Zu vällen. Diese vorlose Ihre wir die Viel einschauen dann nicht geworden. Und wurde vollwerk viel von Menschen die Grünen darf und dahler Ihr einmal das einmal in wollen in um dann nicht, noch mal schleiner ist die'

Figure 2: Output of German GPT-2 (Schweter, 2020) fine-tuned on toxic comments from the dataset

we fine-tuned two distinct models. Model 1 was fine-tuned on the data labelled as 'toxic' and model 2 on the 'non-toxic' data. Similar to the training data, we generated 2000 non-toxic synthetic comments and 1000 synthetic toxic comments. The synthetic comments were then merged with the original dataset. The train/test split was set to 80/20.

3.5 Classification Model Description

For the binary classification task at hand here, *BERT Multilingual Cased* was selected, as it is the gold standard for non-English NLP tasks (Miranda-Escalada et al., 2020; Keung et al., 2020). The focus of this work was not to create a top-performing system but to investigate the effects of fine-tuning using synthetic data. It is likely that more robust results could be achieved using one of the models mentioned previously.
A traint/test split of 80/20 was applied, and the model was trained over 5 epochs. We set the batch size to 6 and the maximum sequence length to 192. The *Adam* optimizer was used. As seen in Table 1, an initial test comparing an mBERT model trained solely on original data and another mBERT model trained on the merged dataset appeared promising.

4 Results and Discussion

As seen in Table 2, the validation results were not replicable on the test data. Therefore, the test results imply no measurable impact on the system's effectiveness through synthetic data generated by the used methodology. In the light of these results, a couple of methodological mistakes need to be addressed.
First, the discrepancy in the validation and testing performance of the model using synthetic data is possibly due to the initial validation results being achieved on data that included synthetic data. This could be problematic, as the data generated by

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

Model	F1	Precision	Recall
mBERT[9]	0.651	0.667	0.645
mBERT - synthetic[10]	0.766	0.772	0.761

Table 1: Initial validation results of an mBERT model fine-tuned using the base dataset and another fine-tuned on the merged dataset

Model	F1	Precision	Recall
mBERT[11]	0.618	0.635	0.599
mBERT - synthetic[12]	0.615	0.623	0.608

Table 2: Results of both models when applied on the test data

gGPT-2 could have caused artificially high testing results. The testing data must be composed solely of original data to avoid potential impacts created by gGPT - 2 on testing results Therefore, future investigations will only use original data for validation testing.

Another issue is the selected Data Generation Model, gGPT-2. The output of GPT-2, as seen in Figure 1, is composed of both English and German sentences. This composition is not the desired outcome, but English and German are sister languages. The output generated by gGPT-2, as seen in Figure 2, on the other hand, appears to be effective on the token level, but in some cases not capable of generating coherent sentences or words. Therefore, we should have selected GPT-2 over gGPT-2.

In light of these mishaps, we still deem the approach of using synthetic data to be successful. Synthetic Data is comparatively easy, cheap and fast to use. It did not negatively affect the baseline approach.

Furthermore, ethical and practical implications of Synthetic Data are major issues. Privacy concerns of real datasets can be mitigated if Synthetic Data can be generated using real-world datasets that can not be published themselves.

5 Conclusion and Future Work

We conclude that synthetic data does seem to be a promising avenue of research. However, this particular work did not find a measurable improvement over the baseline model using synthetic data. Future work will use either GPT-3 or GPT-2, and we will rework the methodology. Once a more robust method is formulated, we will test it on several different datasets. We suggest further research on Synthetic Data in classification tasks, as the comparatively poor performance of our model may be due to the limitations of our chosen methodology.

Other possible future avenues of research include the training and evaluation of a model solely trained on synthetic data.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2–how can I help you? towards the use of pre-trained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Kelly Dekker and Rob van der Goot. 2020. Synthetic data for English lexical normalization: How close can we get to manually annotated data? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6300–6309, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jonathan A Greenblatt. 2020. Fighting Hate in the Era of Coronavirus. *Horizons: Journal of International Relations and Sustainable Development*, (17):208–221.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263.

Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N Asokan. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. *arXiv preprint arXiv:2009.12344*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. 2020. The multilingual Amazon reviews corpus. *arXiv preprint arXiv:2010.02573*.

Nane Kratzke. 2017. The# btw17 Twitter dataset–recorded tweets of the federal election campaigns of 2017 for the 19th German Bundestag. *Data*, 2(4):34.

Ugur Kursuncu, Yelena Mejova, Jeremy Blackburn, and Amit Sheth. 2020. Cyber social threats 2020 workshop meta-report: Covid-19, challenges, methodological and ethical considerations.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. In *CLEF (Working Notes)*.

Andrei Paraschiv and Dumitru-Clementin Cercel. 2019. UPB at GermEval-2019 Task 2: BERT-Based Offensive Language Classification of German Tweets. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. *Proceedings of SemEval*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Tharindu Ranasinghe and Hansi Hettiarachchi. 2020. BRUMS at SemEval-2020 task 12: Transformer based multilingual offensive language identification in social media. *arXiv preprint arXiv:2010.06278*.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*, pages 1–12.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

Stefan Schweter. 2020. German GPT-2 model.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2021. Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key. *arXiv preprint arXiv:2104.10788*.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining disinformation and fake news: concepts, methods, and recent advancements. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 1–19. Springer.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365. sa.

HMHS Surendra and HS Mohan. 2017. A review of synthetic data generation methods for privacy preserving data publishing. *International Journal of Scientific & Technology Research*, 6(3):95–101.

John R Wells, Carola A Winkler, and Carole A Winkler. 2019. *Facebook fake news in the post-truth world*. Harvard Business Publishing Education, September 14.

Dewayne Whitfield. 2021. Using gpt-2 to create synthetic data to improve the prediction performance of nlp machine learning classification models. *arXiv preprint arXiv:2104.10658*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–10.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. 2020. Learning data augmentation strategies for object detection. In *European Conference on Computer Vision*, pages 566–583. Springer.