IRCologne at GermEval 2021: Toxicity Classification

Fabian Haak **Björn Engelmann**
TH Köln
Gustav-Heinemann-Ufer 54
50968 Köln
fabian.haak@th-koeln.de, bjoern.engelmann@th-koeln.de

Abstract

In this paper, we describe the TH Köln's submission for the "*Shared Task on the Identification of Toxic Comments*" at GermEval 2021. Toxicity is a severe and latent problem in comments in online discussions. Complex language model based methods have shown the most success in identifying toxicity. However, these approaches lack explainability and might be insensitive to domain-specific renditions of toxicity. In the scope of the GermEval 2021 toxic comment classification task (Risch et al., 2021), we employed a simple but promising combination of term-frequency-based classification and rule-based labeling to produce effective but to no lesser degree explainable toxicity predictions.

1 Introduction

Toxic language in online comments and discussions is an increasingly relevant problem (Mathew et al., 2019). However, toxicity classification is a challenging task (Paasch-Colberg et al., 2021). There is no universally agreed on complete definition of toxicity. Instead, toxicity is an umbrella term for a variety of problematic, negative phenomena (Malik et al., 2021). Since, even for human annotators, it often is hard to explain why precisely a comment is or is not toxic, language models trained on an extensive dataset of labeled comments usually performed best in the task of identifying toxicity (Zhao et al., 2021). However, these approaches lack explainability, and when the systems are employed to filter user-composed comments, the system should be able to indicate what aspect(s) of the comment lead to it being declared as toxic. More traditional approaches like support vector machines, linear models, or naive Bayes classification do not perform as well, generally. Nevertheless, TFIDF-based classification has clear advantages in terms of ease of domain adaptability and explainability. Our approach employed in the GermEval 2021 toxicity classification task (Risch et al., 2021) combines these traditional classification methods with the Snorkel framework (Ratner et al., 2017). Using labeling functions (LFs), we identify indicators for aspects of toxicity to enable explainable toxicity judgments and improve classification performance.

2 Toxicity: Definition, Aspects, Classification.

Depending on the definition, a wide range of aspects defines what constitutes toxic comments (Wulczyn et al., 2017). As described by Georgakopoulos et al. (2018), a toxic comment, aside from exhibiting verbal violence, can also be "a comment that is rude, disrespectful or otherwise likely to make someone leave a discussion". This definition is consistent with the definition of toxicity given in the overview paper, where toxic comments are deemed problematic because they discourage and hamper participation in discussions. (Risch et al., 2021). To better tackle the issue of toxic language in comments, we categorized the aspects of toxicity in comments into three categories. The first category of toxicity-defining aspects can be characterized as language aspects. These describe features like particular vocabularies or attributions that carry a toxic tonality. They range from different forms of hate speech such as racism (Kwok and Wang, 2013), sexism (Jha and Mamidi, 2017), fanaticism and identity hate, to profane, offensive and aggressive language and incivility (Risch et al., 2019, 2021). Of the toxic features listed in the task description, vulgar language, screaming, and insults are listed as toxic comment features (Risch et al., 2021). The second category can be described as toxic behavioral aspects, defined by their communicative intentions. They are composed of cyberbullying (Chavan and Shylaja, 2015), sexual

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

predation (McGhee et al., 2011), threats, and so-called spam messages (Founta et al., 2018). Sarcasm and zynism, discrimination, discrediting, accusations, and threats are the aspects of toxic comments mentioned in Risch et al. (2021) that best fit this category. Finally, the most latent type of toxic aspects can be grouped as inappropriate language. These highly context- or domain-specific aspects span from age-inappropriateness (Alshamrani et al., 2021) to general undesirable topics or off-topic messages.

In one of the rare approaches to more explainable toxicity classification, Xiang et al. (2021) tried to address the issue of poor explainability of language model classification techniques. However, their approach is decidedly different from ours. By assuming that a text is at least as toxic as its most toxic part, they focused their work on increasing the explainability of transformer-based classification. As previously mentioned, this is a rare exception since most recent approaches utilize deep learning and attention-based language models like BERT (Devlin et al., 2018).

3 Methodological Approach

Our approach applies binary classification (BC) and data programming on a preprocessed version of the provided corpus. In this section, we give a brief overview of all aspects of this approach.

3.1 Preprocessing

Before any classifier is trained or other potential biased patterns in the text are addressed, the comment texts are preprocessed. Since the original texts are needed for labeling functions, the cleaned texts are saved separately. The preprocessing consists of the following steps:

1. Removing single-character-words
2. Deleting any html snippets
3. Discard all characters that are not European ASCII characters (f.e. digits)
4. Removing any white space characters
5. Tokenization using the TweetTokenizer provided by the NLTK (Bird et al., 2009)
6. Excluding all tokens from NLTK's list of German stopwords
7. Stemming using the Cistem German Stemmer (Weißweiler, 2017)

3.2 Binary Classification

Especially deep learning models with a large number of training parameters require an extensive data set (Feng et al., 2021). Although datasets for toxic comments for pre-training would have existed for the classification of English texts, for German texts, these do not exist. We decided that translating texts from English to German or using datasets with labels for toxic aspects such as sexism or hate speech would induce too much noise. Therefore, we chose to base our model solely on the training data provided by GermEval. Since the training dataset is relatively small with 3244 labeled comments, we considered four different simple model types (see subsection 4.1). This model then serves as a baseline to ensure complete coverage across all comments.

3.3 Data Programming

Ratner et al. have presented a data programming framework (Snorkel) that produces noisy labels using user-defined labeling functions (Ratner et al., 2017). These labeling functions can express simple rules such as regular expressions or more complex heuristics that use external resources. Snorkel is typically used to solve tasks where no labeled dataset is available by combining these labeling functions to produce provisional labels to train a discriminative model. Snorkel has been successfully used for various NLP tasks, such as named entity recognition (Lison et al., 2020), fake news detection (Shu et al., 2020), and spam classification (Maheshwari et al., 2020).

3.4 Labeling Functions

Labeling functions express simple heuristics that assign either a label, in our case Toxic, OK, or abstain, to label an input comment. An example is shown in Figure 1. Each labeling function should represent either a toxic or normal aspect for a comment as part of a larger set of labeling functions to move the classification in one direction. Our approach to producing labeling functions is to examine incorrect labels from the output of our classification model, namely false positives and false negatives. We only examine the incorrect labels from the training data to prevent overfitting. In this human-in-the-loop approach, new labeling functions can be defined iteratively after every evaluation step to improve the final classification performance (Wu et al., 2018).

```
@labeling_function()
def check_smiley(x):
    smileys = ["🤬", "🤮", "😡"]
    text = x.comment_text
    for char in text:
        if char in smileys:
            return TOXIC
    return ABSTAIN
```

Figure 1: Labeling function for toxic emojis.

We can then use Snorkel to examine the following properties for a set of labeling functions (Ratner et al., 2017):

- **Coverage**: Is a measure of how large the proportion of data is for which this labeling function has not been abstained from.
- **Overlap**: Proportion of data for which at least one other labeling function has also assigned a label.
- **Conflicts**: The fraction of the dataset where this LF and at least one other LF label and disagree.
- **Empirical Accuracy**: The empirical accuracy of this LF (if gold labels are provided).

With these properties, the LFs can be evaluated manually and adjusted if necessary. Snorkel can learn a generative model based on the correlations and accuracies among the LFs. This generative model serves as our final classification model.

4 Results

4.1 Classification Models

We randomly split the provided data into 80% train data and 20% for test data. We performed a parameter search for the following classifiers using scikit-learn's Grid Search (Pedregosa et al., 2011).

- Logistic Regression (LR)
- Support Vector classifier (SVC)
- Linear SVC
- Multi-layer Perceptron classifier(MLP)

For LR, SVC and Linear SVC we evaluated the inverse regulation strength $C \in \{0.8, 1, 1.2\}$. For the MLP, we tested the following sizes for the hidden states $(h_1, h_2, h_3) \in \{(10, 10, 10), (20, 20, 20), (40, 40, 40), (80, 80, 80)\}$.

Method	Accuracy	F1-Score (Macro)
Logistic Regression	67.8	51.78
Linear SVC	65.49	57.38
SVC	57.16	54.28
MLP	62.1	57.62

Table 1: Best Results for different classifiers after grid search parameter optimization.

All classifiers were also evaluated with the following tf-idf parameters for the sklearn-learn TFidfVectorizer:

- word n-gram range $\in \{(1, 1), (1, 2), (1, 3)\}$
- minimum document frequency threshold $min_{df} \in \{0, 0.02, 0.04, 0.06\}$
- maximum document frequency threshold $max_{df} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$

The results of the classifiers can be seen in Table 1. The evaluation metrics in our case are accuracy and F1 Score, as the classes in the training dataset are not balanced. We have chosen to use the Linear SVC model as it has the best balance between accuracy and F1 score.

4.2 Labeling Functions

van Aken et al. (2018) have analyzed the results of machine learning classification for toxicity prediction and identified aspects of toxicity that are hard to detect by these models. Similarly, labeling functions for improving the overall classification results are derived from the false positive and false negative classifications that the linear SVC model produced. We hoped to find indicators in the form of patterns that match aspects for toxic comments we established in 2. The potential patterns we identified are:

- **Quotations**: In some false positives, potentially toxic text is presented by the author in quotation marks to show that it is not their thoughts or opinions but something they comment on. For example: "*Sag mal ...willst du eine Menschen mit einer "Rostlaube" vergleichen?? Wie impertinent !!*". We decided to implement this feature in a second submission, where we included the deletion of quotations as an additional preprocessing step.

- **Sentiment**: Negative sentiment could be a universal indicator for toxic comments. As mentioned by van Aken et al. (2018), this could especially be the case if the toxicity is latent and topic-dependent. In the false negatives produced by our binary classification model, comments such as "*100% der AfD scheißt auf die Menschenrechte*" or "*Was eine dämliche Diskussion . Sind wir jetzt völlig verblödet. Ich muss abschalten .Ich bekomme Kopfschmerzen.*" also show negative sentiment. We implemented labeling functions with different thresholds for BAWL-R (Võ et al., 2009), SentiWS (Goldhahn et al., 2012) and German Polarity Cues (Waltinger, 2010) sentiment lexica, as well as German Sentiment (Guhr et al., 2020), a BERT-based sentiment classification approach. Although there were some indications that negative sentiment could be a signal for toxicity, we could not identify any obvious differences between toxic and non-toxic texts. Since this was supported by poor-performing LFs, we did not include any sentiment LFs into our classification.

- **Capitalization**: Capitalizations of words could indicate aggressive language, for example in "*@MEDIUM Wenn Sie als objektive Presse, die sich an Fakten zu halten haben, da Sie auch einen BILDUNGSAUFTRAG haben, sich Pro-Organspende aussprechen sollten, muss man sie entweder der Organ- und/oder PharmaMafia zuordnen oder aber als Lügenpresse bezeichnen bzw erkennen, dass Sie Ihrem Auftrag nicht gerecht werden können. Ihnen würde dann Unfähigkeit attestiert. Alles nicht wirklich nett!*". We quickly realized, that in many cases, 3-character words that are capitalized are abbreviations. Therefore, the labeling function only checks for words longer than three characters.

- **Sarcasm and Ridiculing**: Some text elements like certain emojis at the end of comments (🤣, 😂, and 👍, especially when used in multiples) and the term "*haha*" in various variations and lengths could indicate toxic language.

- **Punctuation**: The use of multiple exclamation points or question marks at the end of sentences is common amongst the falsely negative classified comments. As implemented in the VADER sentiment analysis tool developed by Hutto and Gilbert (2014), multiple punctuation marks at the end of sentences enforce the sentiment of the sentence. Since we expect this to be the case for German texts and what we find in our dataset matches this assumption, we employ the phenomenon as a toxicity indicator as a labeling function. In addition, some falsely classified toxic comments show multiple ellipses, possibly indicating annoyance, such as in "*Und überhaupt...wenn ich Spahn schon sehe...*🤢".

- **Toxic emojis**: Certain emojis that seem to be used to indicate disgust or anger appear almost exclusively in toxic comments and frequently appear in the list of false negatives. Therefore, we introduced a labeling function that checks for appearances of these emojis (c.f. Figure 1)

- **Insults**: We found a lot of insults in the false negatives, such as "*Moralapostel*", "*Trendlemminge*", or "*Menschenhasser*". A labeling function was created that labels texts based on the appearance of insults from a list of German insults from insult.wiki. However, the LF coverage and accuracy were low, probably due to the complexity and context-relatedness of German insults.

LF	Coverage	Overlap	Acc.
question	0.011	0.005	0.897
exclamation	0.018	0.009	0.957
emojis	0.009	0.003	0.826
caps	0.04	0.011	0.625
haha	0.078	0.015	0.379
short_sens	0.057	0.013	0.265
ellipses	0.055	0.02	0.464

Table 2: Properties of the LFs. Those marked in green are part of the final classification and those marked in red have been discarded. None of the LFs showed any conflicts with other functions.

Table 2 shows the properties of some of the labeling functions, including those used in our classification. The first two LFs check whether there are three question marks or exclamation marks in a row in the comment. The high accuracy scores of the LFs indicate that multiple exclamation points and question marks indicate toxicity. The rela-

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

Method	Accuracy	F1-Score (Macro)
Linear SVC	65.49	57.38
Linear SVC + LF	**67.95**	**62.31**

Table 3: Comparison of the final classification with the baseline model performed on the split-produced test data.

tively higher accuracy for exclamation points could be explained by the fact that German exclamation points are used to signal imperative sentences, which could be perceived as toxic in the context of a discussion. As previously described, capitalized words and the use of particular emojis are also a sign of toxicity. This is also confirmed by our exploration and empirical accuracy of the emojis-LF (c.f. Figure 1). We discarded the remaining LFs marked in red because of the low accuracy since these patterns do not indicate toxicity for texts of the given corpus. The discarded LFs check whether the phrase "*haha*" is included, whether a comment contains at least one ellipsis, and short_sens checks, whether the comment consists of sentences with an average length of two or less words. Of the three categories of toxic aspects of comments described in section 2, mostly language aspects are effectively covered by the LFs. Covering the more latent behavioral aspects like discrimination, sarcasm, or threats indirectly by negative emotions was ineffective. Since inappropriateness is context-dependent and the context of the dataset is unknown, the LFs do not cover inappropriateness aspects of toxic comments.

Table 3 shows that the use of LFs led to an increase in Accuracy of 2.46 points and an increase in the Macro f1 score of 4.93.

4.3 Classification Results

Finally, we used our approach for classifying the provided evaluation test dataset. We produced two almost identical classification runs. The only difference is that text in quotation marks was removed from the training and test data in the second run. As described in subsection 4.2, this aimed at ignoring references to other potentially toxic comments. However, as Table 4 shows, with a F1 score of 0.576 (P: 0.582, R:0.57), the run with no filtering of quotes performed slightly better. Risch et al. (2021) compares the results of all submitted systems.

ID	F1	P	R
1: BC + LFs	0.576	0.582	0.570
2: BC + LFs + Quot.	0.574	0.580	0.568

Table 4: Classification results of the provided test data. Run 1 results were produced only using binary classification and data programming. For run 2 the same methods were used, but as described in subsection 4.2, text in quotation marks was removed from training and test data.

5 Conclusion

To overcome the difficulties posed by GermEval 2021's toxicity classification task, we combined traditional linear SVC classification with labeling functions based on false negative and false positive classifications of the model. This combined approach is able to deliver explainable results and adaptability. Despite the total coverage of the LFs in the training data of about 5% and although we discarded most of the developed labeling functions due to bad classification performance, data programming increased the classification's performance significantly. Including the four best performing labeling functions, our final classification model increased the F1-score by almost 5 points. This increase indicates that the toxic attributes covered by the LFs have not been taken into account by the linear SVC classifier. On the evaluation dataset, our approach reached an F1-score of 0,576. Overall, the approach was a success. However, a more extensive dataset might have benefited our linear SVC model's and our labeling functions' performance.

References

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, abs/1809.07572.

Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. 2021. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Vikas S Chavan and S S Shylaja. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing,*

Communications and Informatics (ICACCI), pages 2354–2358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. *CoRR*, abs/2105.03075.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional neural networks for toxic comment classification.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 759–765, Istanbul, Turkey. European Languages Resources Association (ELRA).

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France. European Language Resources Association.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

insult.wiki. Liste der deutschen schimpfwörter.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 1621–1622. AAAI Press.

Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. *CoRR*, abs/2004.14723.

Ayush Maheshwari, Oishik Chatterjee, KrishnaTeja Killamsetty, Rishabh K. Iyer, and Ganesh Ramakrishnan. 2020. Data programming using semi-supervision and subset selection. *CoRR*, abs/2008.09887.

Pranav Malik, Aditi Aggrawal, and Dinesh K. Vishwakarma. 2021. Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1254–1259.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA. Association for Computing Machinery.

India McGhee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra McBride, and Emma Jakubowski. 2011. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122.

Sünje Paasch-Colberg, Christian Strippel, Joachim Trebbe, and Martin Emmer. 2021. From insult to hate speech: Mapping offensive language in german user comments on immigration. *Media and Communication*, 9:171–180.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

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 SharedTask on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*, pages 1–12.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *KONVENS*.

Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2020. Leveraging multi-source weak social supervision for early detection of fake news. *CoRR*, abs/2004.01732.

Melissa Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus Hofmann, and Arthur Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41:534–8.

Ulli Waltinger. 2010. Germanpolarityclues: A lexical resource for german sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. electronic proceedings.

Leonie Weißweiler. 2017. Developing a stemmer for german based on a comparative analysis of publicly available stemmers. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Berlin, Germany. German Society for Computational Linguistics and Language Technology.

Sen Wu, Luke Hsiao, X. Cheng, Braden Hancock, Theodoros Rekatsinas, P. Levis, and C. Ré. 2018. Fonduer: Knowledge base construction from richly formatted data. *Proceedings of the 2018 International Conference on Management of Data*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale.

Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. 2021. ToxCCIn: Toxic content classification with interpretability. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–12, Online. Association for Computational Linguistics.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. *A Comparative Study of Using Pre-Trained Language Models for Toxic Comment Classification*, page 500–507. Association for Computing Machinery, New York, NY, USA.