

Toxicity Detection in Online Comments with Limited Data: A Comparative Analysis

Max Lübbering^{1,2}, Maren Pielka², Kajaree Das¹, Michael Gebauer³
Rajkumar Ramamurthy^{1,2}, Christian Bauckhage^{1,2} and Rafet Sifa^{1,2} *

1- Universität Bonn - Department for Computer Science

2- Fraunhofer IAIS - Department Media Engineering

3- Technische Universität Berlin - ISE

Abstract. We present a comparative study on toxicity detection, focusing on the problem of identifying toxicity types of low prevalence and possibly even unobserved at training time. For this purpose, we train our models on a dataset that contains only a weak type of toxicity, and test whether they are able to generalize to more severe toxicity types. We find that representation learning and ensembling exceed the classification performance of simple classifiers on toxicity detection, while also providing significantly better generalization and robustness. All models benefit from a larger training set size, which even extends to the toxicity types unseen during training.

1 Introduction

The steadily increasing amount of online communication has been rendering manual moderation almost infeasible. This affirms the importance of automatic detection of toxic content (related to e.g. cyber bullying and harassment[1]) in online conversations. There are different types of toxic comments that are commonly observed, such as threats, insults or attacks based on people's race and sexual orientation. An effective system for toxicity detection should be able to detect all of them with a high accuracy and even generalize to unseen toxicity types, while not wrongly classifying normal comments as toxic.

Toxicity classification poses a supervised classification problem whose existing solutions can be broadly categorized into two categories [2]: manual feature engineering and deep learning methods. While in the first case, features are manually selected and fed to the classifier as input vectors, neural network approaches aim to learn seemingly abstract features present in the text on their own.

A key problem in solving this issue with machine learning (ML), is that there are often no sufficient amounts of data available for all different toxicity types. While conventional ML systems are very accurate in correctly identifying common types of toxicity, such as curse words or obscene language [2], they might lack generalization by failing at detecting other, less obvious attacks.

In order to address this issue of diverse and previously unknown toxicity types, we present a comparative analysis of classification and outlier detection

*First and second author contributed equally.

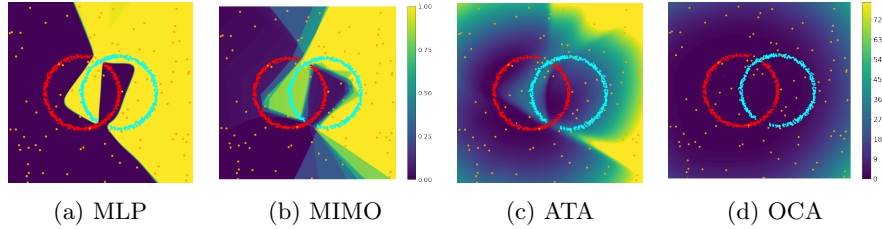


Figure 1: Class probabilities of MLP / MIMO and reconstruction errors of ATA / OCA visualized as contours on the noisy circular segments dataset.

methods. We specifically investigate each system in challenging but very common settings by a) downsizing the training sets and b) constraining the training set to a single type of toxicity and utilizing the remaining classes solely for evaluation. This setup therefore enables us to directly measure the generalization and robustness performance of the algorithms.

In this work, we consider three different types of methods for toxicity detection, namely a) representation learning based outlier detectors, b) ensemble methods and c) traditional deep neural networks. In the first case, a representation of the normal class (here, toxic class) is being learned and any sample that is very dissimilar from this representation is being rejected as an outlier [3, 4, 5, 6]. In practice, this methodology has been successfully applied within a wide spectrum of domains, such as medicine [7], fraud detection [8] or intrusion detection [9]. Building upon these ideas, we selected *adversarially trained autoencoders* (ATA) as a promising supervised outlier detector based on representation learning. Specifically, ATA is composed of an autoencoder that predicts the reconstruction error for a given sample. Due to a custom training approach, that maximizes / minimizes the reconstruction loss for outliers and inliers, respectively, the reconstruction error becomes highly predictive of the outlierness of a sample. As a second baseline based on representation learning, we consider *one class autoencoders* (OCA), a semi-supervised method, which in contrast to ATA only minimizes the reconstruction error of inliers.

Similar to aforementioned outlier detectors, deep learning based ensemble methods have been proven to be more robust than plain *multi-layer perceptrons* (MLPs) [10, 11]. In this work, we specifically consider the MIMO [11] architecture as an ensemble representative, which incorporates the ensembling in a single neural network. Due to this, MIMO makes more efficient usage of parameters and is less overparameterized compared to MLPs [11]. Finally, to put the baseline performances into perspective, we also consider an MLP, one of the most classic methods for binary classification. The different learning behavior of the algorithms is visualized in figure 1, using a 2-dimensional toy dataset. While the MLP splits up the complete feature space with respect to the classes, thereby misclassifying many outliers with high confidence, ATA is more capable in learning a representation for each class.

Our contributions can be summarized as follows:

Test split	#toxic samples	#non-toxic samples
toxic-only	1710	10000
threat	654	10000
insult	10686	10000
identity-hate	1995	10000

Table 1: Number of toxic samples for each of the four test splits. Note, that each test split shares the same 10000 non-toxic samples.

- We present a custom experiment setup by limiting the training set size and constraining the observed toxicity types. This setup enables us to evaluate the models as close to real-world scenarios as possible.
- We compare methods from three different areas, namely representation learning, ensemble methods and deep learning methods solely optimized for classification.
- Our evaluation on the toxicity detection task comprises three different aspects: Classification performance, generalization capabilities and robustness.

2 Toxicity Detection Dataset

In this work, we use the toxicity detection dataset published by Google Jigsaw for the Toxic Comment Classification Challenge [12] on Kaggle. This multi-label dataset originally contains 159,751 training samples and 153,164 independent test samples. The samples have been annotated by 5000 human annotators according to their toxicity level. These annotated comments were categorized into six toxicity classes: *toxic*, *severe toxic*, *insult*, *threat*, *obscene* and *identity hate*.

The categories of the toxic comments overlap. Only 39.2% of them have been categorized with just one label. For better interpretability of the results, we define an additional label *toxic-only*, which is assigned to those samples that have only the *toxic* category annotated (and no other toxicity label).

For our experiments, we consider a strongly reduced version of the original data set. This is done to simulate the common situation, where only limited data is available, and to make it harder for the algorithms to learn general properties of the data. Firstly, we remove all samples from the training set, that have any label other than *non-toxic* and *toxic-only*. Secondly, we apply downsampling to further reduce the overall dataset size. To make the dataset suitable for binary classification, we treat all comments with any toxic label as *toxic*, and all others as *non-toxic*.

The evaluation is done on four separate test sets. They all contain the same, randomly sampled 10000 non-toxic samples from the original test set, and a number of toxic samples with distinct types. They are defined as specified in

Tabl.1. Note that, we allow for overlap with different toxic labels, except for the *toxic-only* test set, which consists of samples with only the *toxic* category (see above).

3 Experiments

For each method, we apply an extensive grid search (GS) over multiple parameter settings. We perform hyperparameter-tuning w.r.t. *learning rate*, *weight decay* for each method and specifically w.r.t. *outlier weighting factor* and *outlier bin start* for ATA.

To achieve a fair comparison, each model is parameterized with comparable complexity. The MLP has four hidden layers of sizes 100, 50, 25 and 12 and binary output. MIMO has an ensemble size of 3 and hidden layers of size 50, 25 and 12. Finally, ATA comprises three hidden layers of sizes 60, 30 and 15 for the encoder and for the decoder in reverse order. All methods have sigmoid activations. In conclusion MIMO, MLP and ATA have 16650, 16765 and 16750 trainable parameters, respectively.

For the representation learning methods (ATA and OCA), we chose to minimize the reconstruction error of *toxic* samples, because we find that the models are able to generalize better using this setup. Intuitively, toxic comments tend to share a rather limited vocabulary and range of topics, which is why they are more homogeneous among each other in comparison to *non-toxic* comments.

4 Results

For the evaluation, we consider the *area under the precision-recall curve* (AUPR) [13] and F1 score, both w.r.t. the *toxic* class. AUPR is a threshold-independent metric which takes the base rate of the positive class into account. Since the toxicity dataset is highly imbalanced, this metric yields more accurate results, compared to imbalance-invariant metrics, such as *area under receiver operating characteristics* (AUROC), which are distortion prone [14, 15]. We also report F1 score to measure the model performance w.r.t. classification and reasonable threshold learning.

As shown in Tabl.2, the MLP which is solely optimized for classification does not generalize well to unseen toxicity types. While the overall classification performance on *toxic-only* is close to MIMO, which is the best classifier on the toxic-only split, MLP shows significantly higher performance degradation on the unseen toxicity classes such as *threat*. MIMO is the most stable method among the four baselines. It provides strong classification performance on the known toxic-only class, but also generalizes well to the three unseen toxicity classes. Nevertheless, similar to MLP, we also find that MIMO tends to fail at times, as seen for the *threat* class on the smallest training set. ATA shows the strongest performance on unseen toxicities, while also providing competitive results on the toxic-only split. Interestingly, ATA never yields any complete failures compared to the other baselines, indicating that the representation learning setup achieves

test split	method	toxic: 250 non-toxic: 1250		toxic: 1000 non-toxic: 5000		toxic: 5000 non-toxic: 60000	
		AUPR	F1 Score	AUPR	F1 Score	AUPR	F1 Score
toxic-only	BASE	14.6	11.3	14.6	11.3	14.6	11.3
	MLP	49.1	46.9	48.5	48.5	51.0	48.6
	MIMO	49.8	50.7	51.0	51.2	53.2	49.7
	ATA	47.7	50.4	44.6	48.0	51.8	52.6
	OCA	14.4	8.5	15.3	9.3	14.4	10.3
threat	BASE	6.1	5.5	6.1	5.5	6.1	5.5
	MLP	50.9	29.5	49.3	48.5	62.9	31.3
	MIMO	47.3	36.3	65.7	36.6	67.9	32.9
	ATA	65.0	38.8	65.2	38.8	67.9	40.4
	OCA	5.7	7.2	7.3	9.8	5.7	7.2
insult	BASE	51.7	25.4	51.7	25.4	51.7	25.4
	MLP	91.9	85.0	91.8	85.6	93.5	86.0
	MIMO	91.9	85.8	92.7	86.7	94.0	86.5
	ATA	92.5	85.6	91.4	83.9	93.4	87.0
	OCA	53.3	14.4	56.7	17.6	53.3	17.6
identity-hate	BASE	16.6	13.5	16.6	13.5	16.6	13.5
	MLP	76.4	55.4	74.6	56.1	81.5	57.3
	MIMO	75.6	62.3	74.5	62.2	82.5	59.1
	ATA	78.4	63.1	76.4	62.0	81.7	64.7
	OCA	16.1	10.4	18.3	12.6	15.9	11.1

Table 2: Performance of MLP, MIMO, ATA and OCA on test splits toxic-only, threat, insult and identity-hate. We consider the Area under the Precision Recall Curve (AUPR) and the F1-Score (both with respect to the "toxic" class). As a reference for the base rate dependent metrics, we also report the expected scores for a random classifier (BASE) with uniform probabilities $p \sim U[0, 1]$.

highest robustness. Finally, we also see that the training approach is crucial. While ATA incorporates not only *toxic* but also *non-toxic* samples in the training procedure, OCA's training routine exposes the model only to toxic samples, leading to underperformance even compared to the random BASE baseline. This is a well-known problem which especially arises when the inliers correlate with outliers in feature space [3, 4]. Interestingly, all methods perform superior on the insult test set, compared to the others. This could be explained by the fact that insults are relatively easy to spot based on certain key words, while threats and identity-related hate are usually context-dependent.

Relevant for the training and deployment of toxicity detection systems, we find that training set size has a significant impact on the generalization performance of such systems. ATA, MIMO and MLP all improve their AUPR scores with bigger training set sizes. Interestingly but not unexpected, this even transfers to toxicity classes not seen during training.

5 Conclusion

Toxicity detection is a challenging task. There are many different types of toxicity and naturally, not all types of toxicity can be observed during training. This is why it is inevitable to have algorithms which are able to learn the abstract toxicity concept and thereby generalize well to unseen toxicity types. Our results show that deep learning methods, which are solely optimized for classification, such as MLPs, lack generalization performance or even tend to fail completely. With ATA and MIMO, we showed that representation learning and ensembling can significantly improve generalization and classification performance. In our future work, we plan to apply the approaches to other datasets, to further test their generalization to toxic comments from other social media sources.

References

- [1] Maeve Duggan. Online Harassment 2017. *Pew Research Center*, 2017.
- [2] Ralf Krestel Betty van Aken, Julian Risch and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *EMNLP*, 2018.
- [3] Max Lübbering, Rajkumar Ramamurthy, Michael Gebauer, Thiago Bell, Rafet Sifa, and Christian Bauckhage. From imbalanced classification to supervised outlier detection problems: Adversarially trained auto encoders. In *Int. Conf. on Artificial Neural Networks*. Springer, 2020.
- [4] Max Lübbering, Michael Gebauer, Rajkumar Ramamurthy, Rafet Sifa, and Christian Bauckhage. Supervised autoencoder variants for end to end anomaly detection. In *Pattern Recognition. ICPR Int. Workshops and Challenges*, 2021.
- [5] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. Outlier Detection using Replicator Neural Networks. In *Proc. of Int. Conf. on Data Warehousing and Knowledge Discovery*, 2002.
- [6] Lukas et al. Ruff. Deep semi-supervised anomaly detection. In *Int. Conf. on Learning Representations*, 2020.
- [7] Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using constrained Adversarial Auto-Encoders. *arXiv preprint arXiv:1806.04972*, 2018.
- [8] Junyi Zou, Jinliang Zhang, and Ping Jiang. Credit Card Fraud Detection Using Autoencoder Neural Network. *arXiv preprint arXiv:1908.11553*, 2019.
- [9] Mohammed Gharib, Bahram Mohammadi, Shadi Hejareh Dastgerdi, and Mohammad Sabokrou. AutoIDS: Auto-encoder based Method for Intrusion Detection System. *arXiv preprint arXiv:1911.03306*, 2019.
- [10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.
- [11] Marton Havasi et al. Training Independent Subnetworks for Robust Prediction. In *Int. Conf. on Learning Representations*, 2021.
- [12] Jigsaw/Conversation AI. Toxic comment classification challenge, 1999.
- [13] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 2006.
- [14] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), 2015.
- [15] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *Proceedings of International Conference on Learning Representations*, 2017.