Denoising Labeled Data for Comment Moderation Using Active Learning

Andraž Pelicon[◇], Vanja Mladen Karan[*], Ravi Shekhar[†], Matthew Purver[*◇], Senja Pollak[◇]

[◇]Jožef Stefan Institute, [*]Queen Mary University of London, [†]University of Essex
andraz.pelicon@ijs.si, m.karan@qmul.ac.uk, r.shekhar@essex.ac.uk,
m.purver@qmul.ac.uk, senja.pollak@ijs.si

Abstract

Noisily labeled textual data is abundant on internet platforms that allow user-created content. Training models, for tasks such as comment moderation, on such data may prove difficult as the noise in the labels prevents the model from converging. In this work, we propose using active learning methods for *denoising* data for model training. The goal is to use active learning to sample the most informative, noisily-labeled examples and send them to the oracle for reannotation, thus increasing data quality while reducing the overall cost of reannotation. In this setting, we tested three existing active learning methods, namely DBAL, Variance of Gradients (VoG), and BADGE, applying them to data denoising for the problem of comment moderation. We show that active learning can be effectively used for data denoising. However, care should be taken when choosing the algorithm for this purpose.

Keywords: Data denoising, Active Learning, Comment moderation

1. Introduction

Comment moderation has recently garnered much attention in both scientific circles and industrial applications (Vidgen and Derczynski, 2020). Internet platforms that allow user-created content (e.g., internet forums, social media platforms, news platforms with comment sections) aim to remove overly toxic, hateful, or profane content through moderation to keep comment threads civil, and in some cases have a legal responsibility for their content. Increasing data volumes, however, mean that moderation has become infeasible to perform manually, especially if required in near real time (Andersen et al., 2021). Consequently, manual moderation is often supplemented with automated tools (Shekhar et al., 2020).

In recent years, due to their performance and adaptability to diverse tasks, large contextualized language models (LLMs) based on transformer architectures (Devlin et al., 2019; Peters et al., 2018) have become ubiquitous in Natural Language Processing. The main training paradigm for LLMs is transfer learning, where the model is first pretrained on large amounts of unlabeled text and subsequently fine-tuned on labeled data of specific downstream tasks, including for offensive speech detection (Pelicon et al., 2021b). However, the fine-tuning stage still relies on having sufficient amounts of high-quality labeled data, which can be challenging to obtain. In the case of comment moderation, sufficient amounts of labeled data are often available if human moderator decisions can be treated as labels (see e.g. Pavlopoulos et al., 2017b). However, these labels can be relatively noisy, as popular platforms garner a lot of traffic, making moderation mistakes common (Miok et al., 2022). To train production-quality models on such data, a relabeling campaign is often warranted to improve annotation quality. However, such a large-scale re-annotation is often infeasible due to excessive costs and time commitment.

Active learning (AL) is a paradigm generally used to optimize annotation efforts when the annotation budget is limited (Fang et al., 2017). AL methods aim to identify samples that are most informative and will benefit the training of a prediction model. This is achieved by sampling unlabeled examples based on an *acquisition function* and sending it for annotation. Annotation is then performed by an *oracle*, an outside information source that usually consists of a human expert. In this way, training a well-performing model while keeping data annotation costs low is possible. This paper explores the idea of leveraging AL to guide the re-annotation efforts that address the noisy label issues. We implement three AL methods to identify promising examples for relabeling. These are manually re-labeled and used to train a new, better comment-moderation model. This can be repeated for several iterations.

The contributions of this paper are the following:

- We propose an AL-based approach for guiding the re-annotation efforts of examples with noisy labels.
- We extensively evaluate the proposed framework in a realistic scenario using data from a real-world news provider.
- We make our implementation of the framework and three AL methods, along with all data and models, publicly available.[1]

[1] Data And Code

LREC-COLING 2024, pages 4626–4633
20-25 May, 2024.

While we focus on comment moderation in this work, we believe the proposed approach can be readily extended to other text classification tasks.

2. Related work

Active learning The main two categories of AL methods are *Uncertainty sampling* and *Diversity sampling*. The former relies on confidence scores of prediction models to select new examples, while the latter aims to find examples dissimilar to already labeled ones, bringing new information. The earliest AL methods fall into the *uncertainty sampling* category. Among the early and established methods are the least confidence score (Settles and Craven, 2008), margin score (MacKay, 1992), and confidence entropy (Hwa, 2004). The Query-by-Committee (Seung et al., 1992) approach trains a committee of prediction models and measures its uncertainty, most often via voter entropy (Dagan and Engelson, 1995) or Kullback-Leibler divergence (McCallum et al., 1998). These approaches are model agnostic and computationally inexpensive and have been used in a variety of tasks (Weber and Plank, 2023).

Several uncertainty sampling methods have been developed, particularly for neural network models. Examples of these are Deep Bayesian Active Learning (Gal et al., 2017), Variance of Gradients (Agarwal et al., 2022), and Variational Adversarial Active Learning (Sinha et al., 2019). Among *Diversity based* methods, the typical example is the core-set approach (Sener and Savarese, 2017) that transforms AL into a set covering problem. Given an annotation budget and an area around an example (a given radius), it selects the set of examples that cover the whole dataset.

Methods that combine criteria from both categories are particularly interesting, aiming to balance uncertainty and diversity. One such method is BADGE - the Batch Active Learning by Diverse Gradient Embeddings (Ash et al., 2019). Most neural approaches tend to perform better for neural network models than model agnostic approaches; however, they were mostly benchmarked on image data.

While more prominent in the domain of image processing, active learning is also becoming widely researched in the domain of natural language processing (Zhang et al., 2022). Margatina et al. (2021) have devised an acquisition function that leverages contrastive examples in order to gauge the informativeness of unlabeled examples. Yuan et al. (2020) propose an acquisition function that leverages transformer models' masked language modeling loss. This way, the acquisition function does not depend on the classification loss of a fine-tuned model, which is usually poorly calibrated in neural models. All these approaches are tested in a standard active learning setting with a pool of unlabeled examples.

Noisy labels Several recent studies have experimented with AL in the presence of label noise. Gupta et al. (2019) assume that the oracle is imperfect and outputs a noisy label. To counterattack this, they introduced a denoising layer in the neural network model that robustifies AL in the presence of noisy oracles. However, this study mostly considers training in the presence of imperfect oracles as opposed to a noisily labeled data pool. A study that resembles ours is the one that presents QActor (Younesian et al., 2021), a framework for using AL in the presence of noisily labeled data, but it was benchmarked only on image data.

Comment moderation The task of comment moderation is predominantly modeled using neural models (Pavlopoulos et al., 2017a), especially the Transformer-based large language models (Badjatiya et al., 2017; Zampieri et al., 2019; Zia et al., 2022). Recent studies tackle comment moderation in multi- or cross-lingual settings, leveraging multilingual LLMs (Pelicon et al., 2021a,b; Pamungkas et al., 2021; Shekhar et al., 2020; Bigoulaeva et al., 2023; Jiang and Zubiaga, 2024). To improve the detection of offensive and hateful language on social media as well as on news platforms, approaches have lately been proposed that enrich modeling with user- and platform-related features, e.g., Mosca et al. (2021); Risch and Krestel (2018); Haber et al. (2023).

3. Experimental setting

Our setting is an example of pool-based active learning. In this setting, we have a small initial training set, T, on which we train our initial or seed model. Additionally, we have a large pool of unlabeled instances M. In our case, M contains data with noisy labels. From M, we sample instances that will produce higher-value information for model training. We use an acquisition function A to gauge the informativeness of the sampled instances and we send n instances to an oracle O for reannotation subject to the available reannotation budget B.

3.1. Methods

We tested three AL methods for denoising training data: Deep Bayesian Active Learning (DBAL), Variance of Gradients (VoG), and Batch Active Learning by Diverse Gradient Embeddings (BADGE) method. All three were specifically designed for use with gradient-based models and were tested on image data in their initial implementation. While methods benchmarked on NLP tasks do exist, the chosen methods are widely used and have been shown to achieve robust performance on a number of different tasks. The first two methods, DBAL and Variance of Gradients fall into the uncertainty sampling AL category, while the latter, BADGE, belongs to the category of hybrid approaches.

The **DBAL** method proposes training a deep Bayesian neural network as our classifier. The training of a Bayesian classifier was approximated by using Monte Carlo dropout. We ran the inference model 10 times with the dropout and selected the 50 examples for the reannotation where the variance was highest. The motive was that if there is a large change in the model's confidence with a small change in the network, then the model is not robust for that sample.

The second method in the uncertainty sampling category was the VoG approach of Agarwal et al. (2022). It computes for each input example a set of gradient matrices w.r.t. that input example $\{S_1, ..., S_K\}$ one for each model checkpoint $k \in [1, K]$. These are used to estimate the variance across checkpoints for each element of S. The average of these variances is the score of the example. The intuition behind this method is that atypical, out-of-distribution examples will have less consistent gradients with higher variance across checkpoints. The original paper of Agarwal et al. (2022) uses this method on images, so S_k are gradients to the pixels of the input image, and thus the same size for all examples. We adapted this to text data and calculated S_k as gradients to the outputs of the embedding layer. Consequently, an input example of M tokens using an embedding of length N would yield S_k of size $M \times N$. The rest of the calculation proceeds exactly as described by Agarwal et al. (2022). When running this method, we used $K = 20$ checkpoints.

The last chosen method, **BADGE**, computes the prediction y and the gradient *gx* of the loss L(x, y) w.r.t. last layer's parameters. This gradient is then used as an embedding, a representation of the input example. Given *gx* of the examples in the noisily labeled pool, we then run k-MEANS++ initialization and take the initial cluster centers as examples for reannotation. Though this method was initially tested on image data, no special adaptation was needed for use with text.

3.2. Dataset

The majority of AL settings assume an initial clean dataset. However, our goal is to leverage the existing noisy dataset. We use the publicly available 24sata newspaper comment dataset (Shekhar et al., 2020), in the Croatian language.[2] The dataset contains comments moderated by 24sata's moderators based on the newspaper's policy: rules include the removal of hate speech, abusive statements, threats, obscenity, deception & trolling, vulgarity, and comments that are not in Croatian. We used comments from 2019 and sampled the comments with the help of dataset maps (Swayamdipta et al., 2020); 75% of examples were sampled from the easy-to-learn region, and 25% were sampled from the ambiguous region. This was to ensure a dataset representing the range of difficulty present. We engaged and trained three annotators to relabel the sampled dataset to obtain an initial clean dataset according to our annotation guidelines. The dataset is split into a training set (1700 examples), a validation set (250 examples), and a test set (250 examples). We trained a seed model for the experiment using the initial clean dataset.

3.3. Evaluation

We test the active learning algorithms for denoising the task of comment mdoeration. We pose the comment moderation task as a classification task where we map instances *x* into *y* = ***C****{1, .., C}* discrete classes. We model this task with Transformer neural networks based on the multilingual BERT model (Devlin et al., 2019).

The active learning pipeline is run for 10 rounds. In each round, we use the model trained in the previous round and sample 50 instances from pool *M*, starting with the seed model for the first round. These examples are sent to an oracle for reannotation. The oracle consists of one annotator from our initial reannotation campaign and labels the sampled instances using the same annotation guidelines the annotators used in the initial training data reannotation. The annotated instances are then added to the initial training set *T*, and a new model is trained on this expanded training set. This pipeline is run for each of the three active learning approaches. The model's performance in each round is evaluated using a macro F1 score.

We compare the models trained using active learning approaches with two baselines:

- **No-reannotation** - in each round, we sample 50 examples with an AL algorithm *A* from pool *M*, but we do not send them for reannotation.
- **Random addition, no active learning** - in each round, we select 50 examples from the pool *M* randomly, add them to the training set with their original moderator-generated labels, and retrain the model. No reannotation is performed on the selected examples. This baseline represents the standard fine-tuning approach one would employ in the presence of a noisily labeled dataset. To minimize the effect of randomness, we present the average of random sampling done thrice.

[2] Available at `https://clarin.si/repository/xmlui/handle/11356/1399` (Pollak et al., 2021)

4. Results

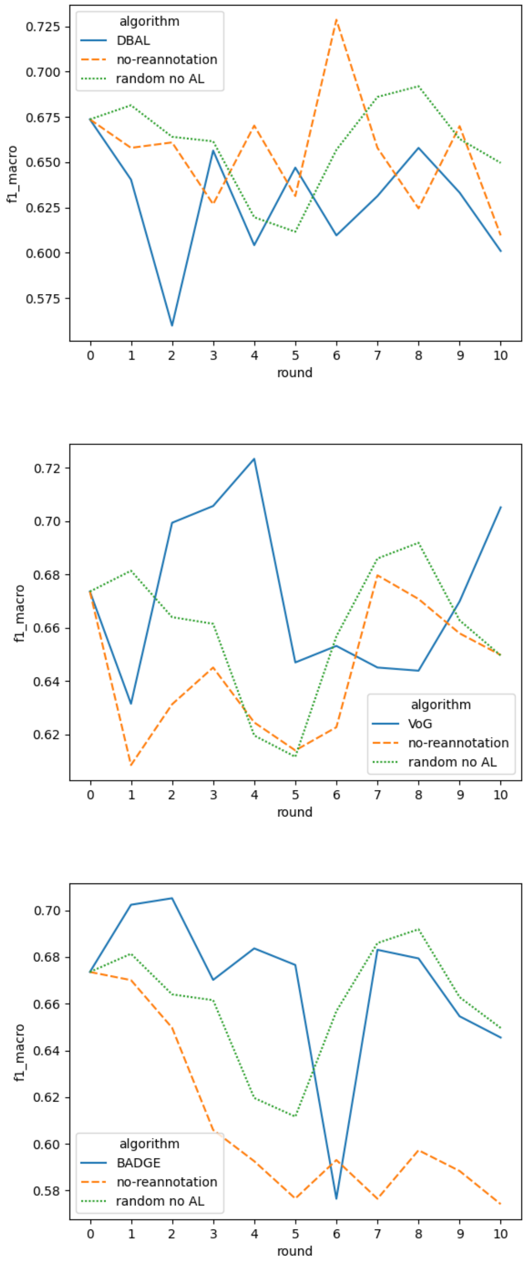

Figure 1: Comparison of three AL methods (DBAL, VoG, BADGE), with two baselines. In Baseline 1, 'no-reannotation', the same AL method is used to select new samples to add to the training set, but they are not sent for reannotation (the original noisy moderator-generated labels are trusted). In Baseline 2, 'random no AL', we randomly select new samples, again keeping the original labels; this represents the standard fine-tuning approach without AL.

Our experiment shows that using active learning is beneficial for training comment moderation models in the presence of large proportions of noisy data. Using these methods allows us to forgo relabeling of the whole dataset and thus reduces the time and cost of relabeling campaigns. The results show (see Figure 1) that two of our algorithms outperform the 'random addition, no active learning' baseline. The VoG algorithm achieves the highest macro F1 score of all three algorithms. While performing a bit worse, the BADGE algorithm seems to have the most consistent performance across the first five rounds of all the tested algorithms with the least variance in the results. While we observe a slower start and a considerable dip in performance from the variance of the gradients algorithm, the performance of the BADGE algorithm stays within the 0.02 point difference in terms of macro F1 score across the rounds. However, the performance drops off significantly, starting with round 6. The Monte Carlo dropout algorithm performs worst in our setting as it does not outperform the baselines. We do not see this algorithm as useful while training models in the presence of noise in the dataset.

Further, the results show that generally, the reannotation is a crucial step in the proposed approach and setting. Without reannotating the sampled instances, the performance of the BADGE and VoG algorithms significantly drops to the point where there are no clear benefits from the random baseline. However, this effect is not observed with the MC algorithm, where reannotation does not seem to impact the model's performance that much.

According to the results, the active learning framework can give us performance benefits up to a certain point. After that point, the model sufficiently converges that reannotation and careful sampling have diminishing returns. According to the the results from the BADGE and VoG algorithms, this cut-off point is algorithm dependent. We observe the BADGE algorithm being useful in our setting up until round 5, while the VoG algorithm performs well even after 10 reannotation rounds. This result also potentially speaks about the quality of the instances sampled by each algorithm. This cut-off point is generally hard to predict, so we suggest employing an early stopping strategy when training in an AL framework in the presence of noise.

Finally, we analyze the types of instances the three algorithms sample in the current setting with noisy datasets. Regarding class labels, we observe that all three algorithms sample instances mainly from the not-offensive class. This is not surprising due to the construction of the noisily labeled pool in the current dataset, where most instances are initially marked as not offensive. The reason for this is that moderators often can't keep up with the high data volumes at peak times, and many examples never even get seen during moderation. Such examples are, therefore, by default, labeled

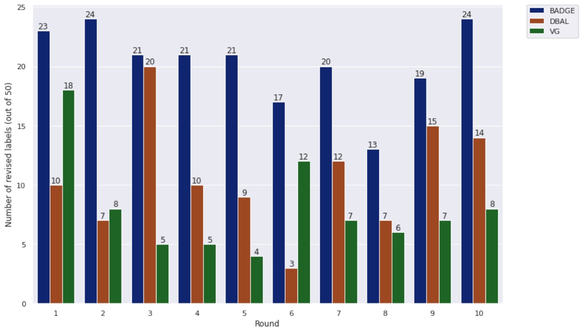

Figure 2: Number of examples with originally noisy labels denoised during reannotation out of a total of 50 sampled examples per round.

as not offensive. Due to this fact, we expect more noise to occur in the not-offensive class.

Further, from Figure 2 we see that all algorithms sampled instances from the noisy pool M that were initially incorrectly labeled. In each round, each algorithm sampled at least 10% of instances, which were eventually relabeled by the oracle with the opposing class label. The BADGE algorithm seems to sample the highest proportion of noisily labeled examples - in each round, approximately 50% of sampled instances had to be relabeled by the oracle. We also noticed that, depending on the round and algorithm used, reannotation can have a strong impact even with relatively low amounts of data reannotated. We would like to explore this phenomenon in the future by developing methods to identify which sampled examples do not need reannotation, thus further decreasing the budget needed for annotation.

5. Conclusion and Future work

In this study, we presented a framework for using active learning methods in the presence of label noise. We tested three active learning algorithms on the task of comment moderation with the goal of denoising the training data and training a performant model while reducing the annotation budget needed for denoising the training data. Our results show that active learning methods can effectively be used in this setting and on the chosen data domain, however the particular active learning algorithm has to be carefully chosen; in our case, for example, the DBAL algorithm did not yield desirable results.

Even though reannotation is a crucial step in the proposed approach, we have observed that, depending on the active learning algorithm and round of the experiment, a considerable proportion of the examples did not change their labels during reannotation. For future work, we would like to explore approaches to reduce the annotation effort even further by eliminating the need to relabel instances with already correct labels while maintaining the same performance of trained models.

6. Limitations and Ethical Considerations

Limitations A limitation of this work is that we focus on the domain of comment moderation and thus cannot claim to what extent our results will generalize to different domains or tasks without further experimentation. Moreover, we use alternatives to the BERT-based classification model, such as the more recent GPT or LLama-2 (Touvron et al., 2023) models, which should also be explored in future work. Finally, the language of our dataset is Croatian; while it should be representative of a typical low-resource indo-european language, our approach may behave differently for larger, more popular languages (e.g., English or Chinese) or typologically different languages from other families (e.g., Japanese or Finnish).

Ethical considerations Our research is grounded in utilizing existing publicly available datasets and established methods. Our primary objective is to provide an in-depth analysis of the robustness of various active learning algorithms. Our re-annotated dataset will contribute to online comment moderation research, particularly in the context of low-resource languages. It is worth acknowledging that the dataset and models we present theoretically could be used to train generative offensive speech models. However, it is essential to note that such models can already be developed using much larger datasets that are readily accessible. An alternative application of our dataset and models is to enhance our understanding of the limitations of current offensive speech detection tools. This understanding can bolster these tools and make them more resilient.

Acknowledgements

We acknowledge financial support from the Slovenian Research Agency (ARRS) core research programs Knowledge Technologies (P2-0103) and the projects CANDAS (J6-2581), SOVRAG (J5-3102) and EMMA (L2-50070); as well as the UK EPSRC via the projects Sodestream (EP/S033564/1) and ARCIDUCA (EP/W001632/1); and the project RobaCOFI (Robust and adaptable comment filtering), which indirectly received funding from the European Union's Horizon 2020 research and innovation action programme via the AI4Media Open Call #1 issued and executed under the AI4Media project (Grant Agreement no. 951911).

7. Bibliographical References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. 2022. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378.

Jakob Smedegaard Andersen, Olaf Zukunft, and Walid Maalej. 2021. Rem: efficient semi-automated real-time moderation of online forums. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 142–149.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2023. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation*, 57(4):1515–1546.

Ido Dagan and Sean P Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. *arXiv preprint arXiv:1708.02383*.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.

Gaurav Gupta, Anit Kumar Sahu, and Wan-Yi Lin. 2019. Noisy batch active learning with deterministic annealing. *arXiv preprint arXiv:1909.12473*.

Janosch Haber, Bertie Vidgen, Matthew Chapman, Vibhor Agarwal, Roy Ka-Wei Lee, Yong Keong Yap, and Paul Röttger. 2023. Improving the detection of multilingual online attacks with rich social media data from Singapore. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12705–12721, Toronto, Canada. Association for Computational Linguistics.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Computational linguistics*, 30(3):253–276.

Aiqi Jiang and Arkaitz Zubiaga. 2024. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *arXiv preprint arXiv:2401.09244*.

David JC MacKay. 1992. The evidence framework applied to classification networks. *Neural computation*, 4(5):720–736.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*.

Andrew McCallum, Kamal Nigam, et al. 1998. Employing em and pool-based active learning for text classification. In *ICML*, volume 98, pages 350–358. Citeseer.

Kristian Miok, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Šikonja. 2022. To ban or not to ban: Bayesian attention networks for reliable hate speech detection. *Cognitive Computation*, pages 1–19.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021a. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online. Association for Computational Linguistics.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021b. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Senja Pollak, Marko Robnik-Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose G. Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 99–109, Online. Association for Computational Linguistics.

Julian Risch and Ralf Krestel. 2018. Delete or not delete? semi-automatic comment moderation for the newsroom. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 166–176.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in croatian and estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Leon Weber and Barbara Plank. 2023. ActiveAED: A human in the loop improves annotation error detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8834–8845, Toronto, Canada. Association for Computational Linguistics.

Taraneh Younesian, Zilong Zhao, Amirmasoud Ghiassi, Robert Birke, and Lydia Y Chen. 2021. Qactor: Active learning on noisy labels. In *Asian Conference on Machine Learning*, pages 548–563. PMLR.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. *arXiv preprint arXiv:2010.09535*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. *arXiv preprint arXiv:2210.10109*.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1435–1439.