REM: Efficient Semi-Automated Real-Time Moderation of Online Forums

Jakob Smedegaard Andersen[1,2], **Olaf Zukunft**[1], and **Walid Maalej**[2]

[1]HAW Hamburg, Hamburg, Germany
jakob.andersen@haw-hamburg.de, olaf.zukunft@haw-hamburg.de
[2]University of Hamburg, Hamburg, Germany
maalej@informatik.uni-hamburg.de

Abstract

This paper presents REM, a novel tool for the semi-automated real-time moderation of large scale online forums. The growing demand for online participation and the increasing number of user comments raise challenges in filtering out harmful and undesirable content from public debates in online forums. Since a manual moderation does not scale well and pure automated approaches often lack the required level of accuracy, we suggest a semi-automated moderation approach. Our approach maximizes the efficiency of manual efforts by targeting only those comments for which human intervention is needed, e.g. due to high classification uncertainty. Our tool offers a rich visual interactive environment enabling the exploration of online debates. We conduct a preliminary evaluation experiment to demonstrate the suitability of our approach and publicly release the source code of REM.

1 Introduction

Online forums have become an integral part of many domains to facilitate participation and deliberation; particularly in online journalism (Manosevitch and Walker, 2009). More and more news sites enable users to participate in public debates around their reporting. Users regularly share their feedback, personal stories, and opinions about journalistic content (Häring et al., 2018). While online forums present a valuable space for deliberation and an information source for news organizations (Loosen et al., 2018), news sites are increasingly confronted with inappropriate and toxic content such as hate-speech (Davidson et al., 2017; Kolhatkar and Taboada, 2017) and spam (Chen and Chen, 2015; Martens and Maalej, 2019). Ethical and legal policies put pressure on news organizations to ensure lawful and netiquette compliant participation.

The expanding volume and velocity of user participation makes it increasingly difficult and expensive to rapidly detect and remove undesirable posts (Sood et al., 2012; Gillespie, 2020). Fully automated Machine Learning (ML) approaches for text classifications have shown remarkable improvements over the last years. However, ML models still lack user acceptance and applicability (Brunk et al., 2019; Gillespie, 2020). Fully automated approaches are known to be error-prone (Scharkow, 2013) and rarely reach the level of accuracy required to be applied in real-word settings.

We seek to overcome these limitations of fully automated approaches by letting humans manually correct and confirm artificial predictions. However, looping humans into supervised learning tasks is time consuming and cost intensive and does not scale well with larger workloads. The question arises which instances, i.e. forum posts, should better be assessed by humans. A common way to guide human moderation is to focus on instances where the ML model is unable to provide a reliable prediction (Pavlopoulos et al., 2017).

This paper introduces REM, a new user-centric tool for the semi-automated moderation of online forums, with a particular focus on online journalism. Our tool combines the fields of Human-in-the-Loop (HiL) (Holzinger, 2016) and Visual Analytics (Keim et al., 2008) to enable a more accurate, efficient, applicable, and transparent moderation process. Since the manual moderation of large datasets is tedious and cost intensive, we seek to minimize human efforts by focusing the manual moderation on instances which are most likely classified wrongly. We accomplish an efficient semi-automated moderation by relying on predictive uncertainty (Der Kiureghian and Ditlevsen, 2009).

Uncertainty estimates enable us to deal with instances a classifier can probably not infer correctly (known unknowns). However, classification

Proceedings of the Joint Conference 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, August 1st - August 6th, 2021.

models can also provide misclassifications where a model does not know that its labelling might be wrong (unknown unknowns) (Attenberg et al., 2011). To deal with unknown unknowns, REM provides a rich visual-interactive interface to facilitate the exploratory analysis and labelling of forum discussions. We follow a user-centric moderation process, where moderators can correct arbitrary inferred labels. The uncertainty of predictions is visualized to support and guide moderation decisions. Further, we implement a novel moderation approach to reduce the amount of human effort required to reach a desired accuracy level. In a preliminary ML experiment, we evaluate the suitability and effectiveness of our moderation approach. The goal of REM is to:

- Support an efficient moderation of online forums.
- Facilitate overviewing online debates in news discussions.
- Plan of manual moderation efforts to reach a desired level of accuracy.

The remainder of the paper is structured as follows. Section 2 introduces our novel moderation approach implemented in REM. Section 3 describes the system design. Then, Section 4 presents our user-interface. In Section 5 we shortly describe the results of the preliminary ML experiment to demonstrate the suitability of our moderation approach as the core feature of our tool. Section 6 discusses related work, while Section 7 concludes the paper and outlines further work.

2 Content Moderation with Human-in-the-Loop

Content moderation in online forums is a typical labelling task. It refers to "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann, 2015). Usually, ethical guidelines, moderation policies, or legal constraints are used to guide moderation decisions.

Our tool implements the Human-in-the-Loop (HiL) paradigm in order to achieve a more accurate and accepted moderation compared to fully automatic approaches. HiL describes a computational paradigm that is characterized by humans continually providing feedback, e.g. correcting artificial models in order to obtain a better predictive behaviour (Holzinger, 2016; Zanzotto, 2019). We aim to efficiently involve human moderators by only consulting them when artificial predictions are too unreliable to be trusted. For this, we use the predictive uncertainty (Der Kiureghian and Ditlevsen, 2009; Gal and Ghahramani, 2016) of an ML model to guide human involvement. Recent uncertainty quantification techniques are capable to identify likely-to-be-wrong predictions, which are worth being checked manually (Hendrycks and Gimpel, 2016).

Every moderation strategy is a trade-off between accuracy improvements and manual efforts. Generally, higher accuracy requires larger workloads. Since highly uncertain predictions are over-proportionally wrong (Hendrycks and Gimpel, 2016), the accuracy improvements are expected to saturate and get less rewarding. To the best of our knowledge, REM is the first tool to explicitly use the expected model behaviour evaluated on a representative dataset for providing guidelines about how much manual effort is needed to reach a desired level of accuracy. Section 5 reports on a preliminary evaluation of our moderation approach.

In addition, uncertainty quantification techniques are generally unable to detect all misclassifications, in particular those where the classifier is mistakenly assuming with a high certainty that they are correct (i.e. unknown unknown) (Attenberg et al., 2011). Therefore, our tool additionally relies on the exploratory visualization and analysis of the data (Keim et al., 2008). As in interactive-learning (Höferlin et al., 2012), we support the user-centred moderation of any instance. Using a visual-interactive interface, we assume that humans moderators are able to extract useful information to actively moderate model outcomes. Visual Analytics (Keim et al., 2008) enables moderators to better know and understand their data and thus to become capable to detect outliers, which are potentially misclassified by the model or are prone to a derailment, e.g., toxic users and topics which are prone to rudeness and require special care. Visual Analytics combines the strengths of humans' visual perception and reasoning along with the computational power of machines during the moderation process.

Figure 1 shows the HiL-workflow implemented in REM. New forum comments ① get immediately classified and enriched with uncertainty information ②. Our tool follows a holistic moderation approach, which builds on top of a binary classi-

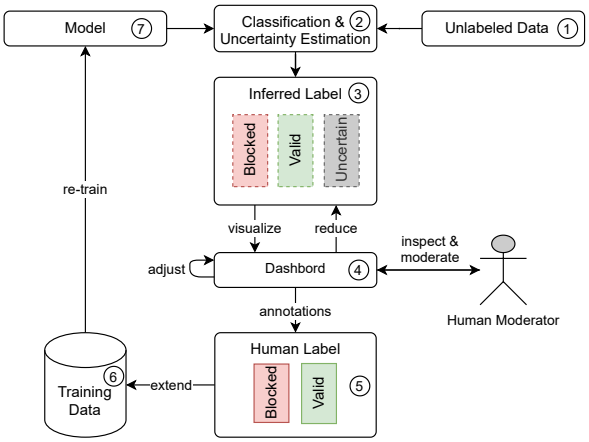

Figure 1: Human-in-the-Loop workflow of REM.

fier. Each comment is either classified as *blocked* or *valid*. Comments can also be marked as *uncertain* ③ if their inferred labelling is too unreliable to reach a desired level of accuracy. Then, human moderators are asked to provide new and more reliable labels for uncertain comments ④. However, we also allow moderators to correct false-positives and false-negatives which are not marked as uncertain. Moreover, our approach integrates an active learning component (Lewis, 1995). Human labelled instances ⑤ are added to the training data ⑥ and used to continually re-train the model ⑦.

Previous studies indicate that such an active learning inspired approach is able to improve the accuracy of existing models (Arnt and Zilberstein, 2003). Since a continuous re-training is inefficient when moderators work in parallel, we implement active learning in batch mode (Hoi et al., 2009). The resulting incremental update of the model weights is particularly important since a model's accuracy is prone to decay over time due to data shifts (Moreno-Torres et al., 2012), i.e. statistical differences in training and operational data.

3 System Design

The components of our tool are depicted in the deployment diagram shown in Figure 2.

The access point of our tool is a web application served by a Node.js[1] server building on top of multiple micro-services. In our prototype, we obtain real-time data by frequently crawling the online forum of a large German news organization. We collect nearly 9,000 user comments daily, distributed across 20 news departments. To ensure a scalable

[1] https://nodejs.org/en/

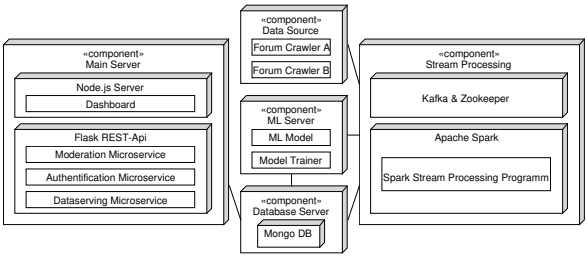

Figure 2: Main components of the semi-automated moderation tool REM.

processing of large volume and volatile real-time data, we implement an ML-pipeline based on the Kappa-Architecture (Kreps, 2014).

We use Apache Kafka (Kreps et al., 2011) as a message broker and the Structural Streaming API of Apache Spark (Zaharia et al., 2010) to implement the data stream processing. In the stream processing pipeline, we first check if comments are already classified with the current version of the model to save computational resources. For non-duplicates, we run text preprocessing steps such as stop word removal and lemmatization. We then apply a neural network based classification model. Then we use Monte Carlo Dropout (Gal and Ghahramani, 2016) to calculate uncertainty estimates. The ML model is built with Tensorflow (Abadi et al., 2016). Model training is performed offline. Finally, the data is persisted and served via MongoDB[2].

4 User Interface

Figure 3 shows the main page of our tool. The user interface consists of three views, which we describe in the following. We share the source code[3] of our prototype together with a video that showcases the tool's main features.[4]

4.1 Context-View

The *Context*-View provides an overview of the comments distribution according to the time-dimension and journalistic entities such as topics, articles, and users (comment writers). The upper bar chart displays the distribution of comments over time. The x-axis represents the time-dimension and the y-axis the total number of comments. Each bar represents a comment label with a three-colour scheme. Blocked comments (e.g. inappropriate, violating

[2] https://www.mongodb.com/
[3] https://github.com/jsandersen/REM
[4] https://youtu.be/cA92Io_xr6Q

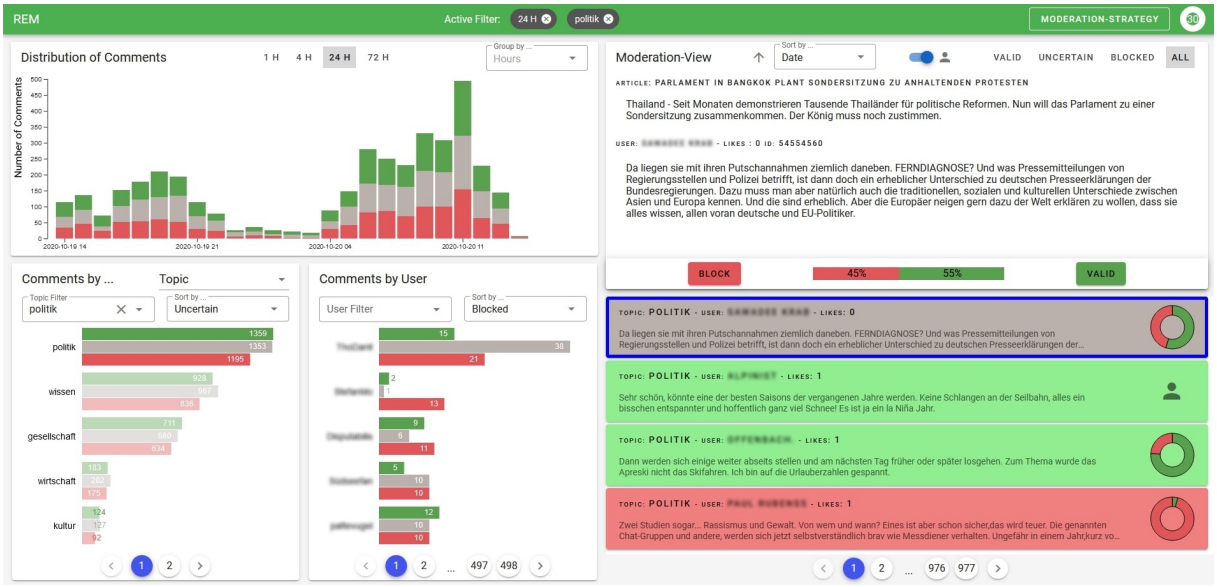

Figure 3: The main page of REM showing the Context-View (left) and the Moderation-View (right).

ect.) are marked as red, valid (none-blocked) comments are green, and uncertain comments are highlighted as grey.

Since the moderation of online forums is a real-time task, the tool focuses on recently added comments. The granularity of the time dimension can be changed through the button group on the top. Possible intervals are minutes, hours, and days. Moderators can select whether to show comments from the last 72 hours to only the last hour.

The lower part of the *Context*-View shows the distribution of comments with regard to journalistic entities, which are topics such as politics or economics and the articles identified by the titles. The second chart depicts the comment behaviour of the users. Each chart can be sorted according to the number of uncertain, blocked, valid, and all comments. All visualizations in the *Context*-View are responsive to filter operations. These can be triggered by clicking on the bars. Specific entities can also be searched over a text-field. Multiple filters can be chained to enable a flexible visual analysis.

4.2 Moderation-View

The *Moderation*-View shown in Figure 3 provides a detailed overview of the selected comments from the *Context*-View. All selected comments are listed here. Each entry on the list consists of the comment's text and additional meta information such as its corresponding topic, the posting user, and the number of recommendations given by other users. Similar to the colour scheme used in the *Context*-View, the colour of each cell represents the current label of the comment. The pie chart visualizes the model's conditional label probability for Blocked and Valid. In highly uncertain predictions, both class outcomes would be nearly equal. If a comment is already labelled by a human, a "human"-icon is shown instead of the pie chart. The list can be filtered to only show uncertain, valid, or blocked comments. Further, the entries can be sorted according to the timestamp or uncertainty. Most uncertain data must be moderated in our approach. Comments that are already manually moderated can also be hidden to enable a faster overview.

The detailed information about a selected comment is shown in the upper part of the view. Additional information about the corresponding article is also provided, followed by the text of the comment. The selected comment is highlighted with a blue box in the comments list. The actual moderation is performed via the buttons down below. An uncertain comment can be blocked or marked as valid. Predictions can also be corrected, e.g. a comment classified as valid can be manually blocked by the moderator. Additionally, the moderator can agree on artificial predictions to provide more training data for the active learning process. Corrections and additional labels are directly synchronized with the database of the training data.

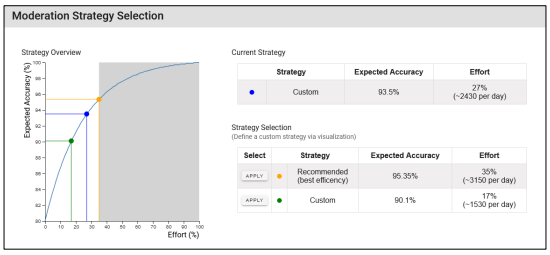

Figure 4: The *Control*-View of REM for managing the moderation strategy.

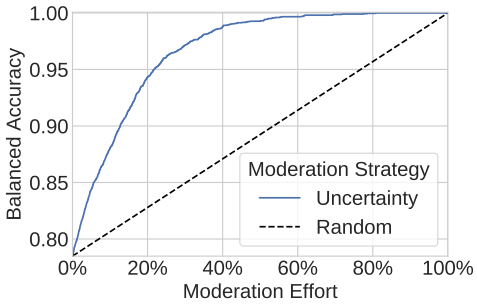

Figure 5: Balanced accuracy in term of moderation effort: uncertainty-based vs. randomly-sampled selection of instances to be moderated.

4.3 Control-View

The *Control*-View is dedicated to steer the moderation process. The view can be activated via the button *"Moderation-Strategy"* on the main page. As described in Section 2, we implement a novel approach to provide guidelines for how much manual effort is needed to efficiently reach a desired level of accuracy. The expected accuracy of the underlying classifier, when a certain amount of the most uncertain predictions are manually validated, is displayed by the line chart shown on Figure 4. On the right a user can select different moderation strategies which are also highlighted in the line chart. For each strategy the expected accuracy and the needed effort is depicted. A user can select a predefined moderation strategy or define a custom strategy by hovering and clicking on a point in the line chart. A moderation strategy affects the number of predictions, which are marked as uncertain. The currently applied strategy is shown above.

Since the efficiency of the moderation is expected to decrease with larger workloads, a point might be reached where further moderation efforts only lead to marginal accuracy improvements. To inform users of such inefficiencies, REM provides a recommended moderation strategy which seeks to optimize human moderation efforts with regard to the accuracy gain. We calculate the recommended moderation effort as the natural point of saturation (Satopaa et al., 2011). Inefficient workload is highlighted by the grey area in the line chart.

Usually, not every moderator should be able to change the moderation strategy and thus the target accuracy of forum moderation. Therefore, the *Control*-View can be secured by assigning specific roles like an administrator.

5 Preliminary ML Experiment

We conduct a preliminary experiment to demonstrate that our semi-automated ML approach is capable of efficiently improving the accuracy of a model during its operational use. For our experiment, we use the dataset provided by Davidson et al. (2017), which consists of 24.782 Twitter comments either labelled as offensive, hate-speech, or neither of them. In our experiment, we classify the comments into blocked (offensive and hate-speech) (83.2% of total) and valid comments (16.8% of total). Since the data is highly imbalanced, we use the balanced accuracy (Brodersen et al., 2010) to measure the performance of the classifier. We split the data into a training and validation set (7868 : 7868) for model training and a test set (9046) to evaluate our approach. The source-code of our experiment is part of our replication package.

We use Sentence-Bert (Reimers and Gurevych, 2019) to compute text encodings. These are used as the input for a feed forward neural network. Further, we apply Monte Carlo Dropout to estimate the uncertainty of the classifications. Our trained classifier reaches a balanced-accuracy of 78.48%. Figure 5 shows the balanced accuracy when a certain percentage of the most uncertain instances of the test data is moderated manually. In our experiment, we simulate manual moderation by selecting the ground truth labels. A workload of 100% corresponds to manually checking 9046 comments, which matches the daily amount of the expected comments in our application scenario. The balanced accuracy of a moderated classifier is computed based on the inferred and manually corrected labels. The results show that an uncertainty based moderation is more efficient than a random moderation strategy, where instances to be labelled are randomly sampled. For instance, moderating 25% of the data based on their uncertainty leads to a balanced accuracy of 96.08%. In comparison, a random moderation strategy requires a moderation

effort of 81.8% to reach the same accuracy and is thus far less efficient.

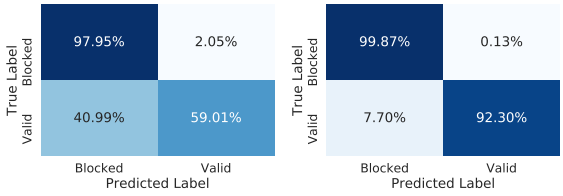

Figure 6: Normalized confusion matrix of the initial classifier (left) and the same classifier when 25% of the most uncertain predictions where moderated (right).

The confusion matrix of the initial and moderated classifier is depicted in Figure 6. The fully automated classifier obviously has difficulties to correctly detect valid comments. Only 59.01% of the valid comments were correctly identified. By moderating only 25% of the data, the detection of valid comments can be increased to 92.30%. As shown in Figure 5, the accuracy of the moderated classifier can be further improved by increasing the amount of human involvement. Thus, our approach is capable of improving the accuracy of a model with a reasonable manual effort.

6 Related Work

There have been previous attempts to efficiently coordinate human involvement to improve the accuracy of ML classifiers. Previous tools mainly focus on the task of interactive model building, also known as active learning (Settles, 2009) and heavily rely on multidimensional projections (Endert et al., 2012). Generally, HiL annotation tools provide a visual-interactive interface to guide human involvement (Höferlin et al., 2012; Bernard et al., 2018). However, tools based on point-visualization are limited in scalability, since data-points will overlap, causing visual clutter. Neves and Ševa (2019) presented a general review of annotation tools for documents.

HiL labelling tools: Seifert and Granitzer (2010) introduce a basic user-centered active learning tool, where humans sequentially select and label instances for the next training iteration. Similar to our approach, the authors utilize the predictive uncertainty to guide human involvement. However, they do not integrate a Visual Analytics component. Heimerl et al. (2012) present a user-centered visual-interactive active learning tool for text documents. Annotators can re-train models in batches and are able to inspect statistics about the model's performance. However, the authors do not consider uncertainty thresholds. The tool provided by Höferlin et al. (2012) enables annotators to manipulate the underlying model directly. This approach requires annotators to be Machine Learning experts, which does not hold for forum moderators e.g. in domains like online journalism. The HiL labelling tool proposed by Choi et al. (2019) facilitates an attention mechanism to explain predictions to annotators. They aim to reduce the time needed to perform annotation decisions and further increase the efficiency of labelling. Our tool might be improved by their findings. Link et al. (2016) introduce a similar semi-automated process for the moderation of social media content. Beside relying on the predictive uncertainty, they also define untrustworthy sources which need additional care. Similar to our approach, human moderation is requested when a prediction does not satisfy a certain confidence level. In contrast, they do not focus on optimizing the moderation in terms of reaching a desired level of accuracy and human efforts needed. Riehle et al. (2020) propose a platform for the semi-automated moderation of online discussions. Similar to our approach, comments are automatically pre-moderated and human moderators can correct or agree on the predicted labels. However, moderators are neither guided to identify comments that require manual attention nor do they assess the effect of the moderation process.

7 Conclusion and Future Work

We introduce a novel tool for the semi-automated moderation of large scale online forums to support content moderators during their daily work. Our tool combines methods from the field of Human-in-the-Loop and Visual Analytics to enable an efficient and more accurate moderation process. We implement a unique approach to reduce and optimize human efforts, building on top of the predictive uncertainty of models. Further, we present a rich uncertainty aware visual-interactive interface to facilitate moderation via exploratory data analysis. Built on top of a big data architecture, our tool is designed to be highly scalable and to enable real-time moderation. A preliminary experiment indicates that our moderation approach is capable of improving the accuracy of a hate and offensive language classifier from 78.48% to 96.08% by only moderating 25% of a test dataset.

REM can be adapted to more generic use-cases, where annotators need to efficiently improve the accuracy of binary classifiers while also making use of active learning. Future work should focus on evaluating our approach regarding its usability, acceptance and usefulness in supporting the moderation of online forums.

Acknowledgments

This work was partly supported by the Hamburg's *ahoi.digital* program within the Forum 4.0 project.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

Andrew Arnt and Shlomo Zilberstein. 2003. Learning to perform moderation in online forums. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 637–641. IEEE.

Josh M Attenberg, Pagagiotis G Ipeirotis, and Foster Provost. 2011. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. 2018. Vial: a unified process for visual interactive labeling. *The Visual Computer*, 34(9):1189–1207.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE.

Jens Brunk, Jana Mattern, and Dennis M Riehle. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 1, pages 429–435. IEEE.

Yu-Ren Chen and Hsin-Hsi Chen. 2015. Opinion spam detection in web forum: a real case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 173–183.

Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 473–482.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2).

James Grimmelmann. 2015. The virtues of moderation. *Yale Journal of Law and Technology*, 17(1):2.

Marlo Häring, Wiebke Loosen, and Walid Maalej. 2018. Who is addressed in this comment? automatically classifying meta-comments in news comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–20.

Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. 2012. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Benjamin Höferlin, Rudolf Netzel, Markus Höferlin, Daniel Weiskopf, and Gunther Heidemann. 2012. Inter-active learning of ad-hoc classifiers for video visual analytics. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 23–32. IEEE.

Steven CH Hoi, Rong Jin, and Michael R Lyu. 2009. Batch mode active learning with applications to text categorization and image retrieval. *IEEE Transactions on knowledge and data engineering*, 21(9):1233–1248.

Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Daniel A Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008. Visual analytics: Scope and challenges. In *Visual data mining*, pages 76–90. Springer.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17.

Jay Kreps. 2014. Questioning the lambda architecture. *O'Reilly Media*. [accessed 2021-03-02].

Jay Kreps, Neha Narkhede, Jun Rao, et al. 2011. Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB*, volume 11, pages 1–7.

David D Lewis. 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *ACM SIGIR Forum*, volume 29, pages 13–19. ACM.

Daniel Link, Bernd Hellingrath, and Jie Ling. 2016. A human-is-the-loop approach for semi-automated content moderation. In *In Proceedings of the ISCRAM 2016 Conference*.

Wiebke Loosen, Marlo Häring, Zijad Kurtanović, Lisa Merten, Julius Reimer, Lies van Roessel, and Walid Maalej. 2018. Making sense of user comments: Identifying journalists' requirements for a comment analysis framework. *SCM Studies in Communication and Media*, 6(4):333–364.

Edith Manosevitch and Dana Walker. 2009. Reader comments to online opinion journalism: A space of public deliberation. In *International Symposium on Online Journalism*, volume 10, pages 1–30.

Daniel Martens and Walid Maalej. 2019. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*, 24(6):3316–3355.

Jose G Moreno-Torres, Troy Raeder, RocíO Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Mariana Neves and Jurica Ševa. 2019. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Dennis M Riehle, Marco Niemann, Jens Brunk, Dennis Assenmacher, Heike Trautmann, and Jörg Becker. 2020. Building an integrated comment moderation system–towards a semi-automatic moderation tool. In *International Conference on Human-Computer Interaction*, pages 71–86. Springer.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Michael Scharkow. 2013. Thematic content analysis using supervised machine learning: An empirical evaluation using german online news. *Quality & Quantity*, 47(2):761–773.

Christin Seifert and Michael Granitzer. 2010. User-based active learning. In *2010 IEEE International Conference on Data Mining Workshops*, pages 418–425. IEEE.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285.

Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, Ion Stoica, et al. 2010. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95.

Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.