Acta Univ. Sapientiae Informatica **12,** 2 (2020) 205–216

DOI: 10.2478/ausi-2020-0012

Machine learning methods for toxic comment classification: a systematic review

Darko ANDROČEC
Faculty of Organization and Informatics, University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
email: dandrocec@foi.unizg.hr

Abstract. Nowadays users leave numerous comments on different social networks, news portals, and forums. Some of the comments are toxic or abusive. Due to numbers of comments, it is unfeasible to manually moderate them, so most of the systems use some kind of automatic discovery of toxicity using machine learning models. In this work, we performed a systematic review of the state-of-the-art in toxic comment classification using machine learning methods. We extracted data from 31 selected primary relevant studies. First, we have investigated when and where the papers were published and their maturity level. In our analysis of every primary study we investigated: data set used, evaluation metric, used machine learning methods, classes of toxicity, and comment language. We finish our work with comprehensive list of gaps in current research and suggestions for future research themes related to online toxic comment classification problem.

1 Introduction

Toxic comments are defined as comments that are rude, disrespectful, or that tend to force users to leave the discussion. If these toxic comment can be auto-

Computing Classification System 1998: I.2.7
Mathematics Subject Classification 2010: 68T50
Key words and phrases: machine learning, toxic comment, deep learning, systematic review

matically identified, we could have safer discussions on various social networks, news portals, or online forums. Manual moderation of comments is costly, ineffective, and sometimes infeasible. Automatic or semi-automatic detection of toxic comment is done by using different machine learning methods, mostly different deep neural networks architectures.

Recently, there is a significant number of research papers on the toxic comment classification problem, but, to date, there has not been a systematic literature review of this research theme, making it difficult to assess the maturity, trends and research gaps. In this work, our main aim was to overcome this by systematically listing, comparing and classifying the existing research on toxic comment classification to find promising research directions. The results of this systematic literature review are beneficial for researchers and natural language processing practitioners.

This work is organized as follows: Section 2 describes in detail the research methodology used for our systematic literature review. The next section lists the results and provides a discussion about obtained results of the systematic review. Our conclusions and future research ideas are provided in the final section.

2 Research methodology

This study has been carried out using the systematic literature review (SLR) methodology described in [16]. First, we have defined the SLR protocol. Then, we performed the study selection and the data extraction process whose outcome is the final list of papers. The main steps of the SLR protocol are listed and elaborated in the next subsections.

2.1 Planning

Planning starts with the identification of the needs for a specific systematic review. We have explained the needs for a systematic review on machine learning methods for toxic comment classification in Introduction section. Next, we have defined the following main research questions:
RQ1: When did the research on toxic comment classification become active in the research community?
RQ2: How is toxic comment classification research reported and what is the maturity level of the research in this field?
RQ3: Which data sets are used to classify toxic comments?

RQ4: Which machine learning methods are used to classify toxic comments?
RQ5: What are main evaluation metrics used to classify toxic comments?

Based on the objectives and research questions of this study, we have defined the review protocol. We have decided to include the following electronic databases: ACM Digital Library, IEEE Xplore, Scopus, and Web of Science Core Collection. These sources are chosen because they represent comprehensive literature in the machine learning field. We also included arXiv, because some highly-cited machine learning papers from researchers of commercial organizations is sometimes published only at this open-access archive service. The search string was simply defined as "toxic comment classification". We defined the following inclusion criteria (IC):

- IC1 - The main objective of the paper must discuss or investigate an issue related to toxic comment classification.
- IC2 - The work must be a research (scientific) paper.
- IC3 - The paper must be written in English.
- IC4 - The study should be published as a conference or a journal paper or a book chapter or an arXiv document.

We excluded papers based on the following exclusion criteria (EC):

- EC1 - Studies that are not related to the research questions.
- EC2 - Studies in which claims are non-justified or studies that had ad hoc statements instead of evidence-based statements.
- EC4 - Papers reported only by abstracts or slides.
- EC5 - Duplicate studies.
- EC6 - Demonstrations, preliminary studies, position papers, technical reports, posters and proof-of-concept papers were excluded.

2.2 Conducting

Conducting is the second step of the systematic review procedure. We have performed the search operation on the mentioned five electronic sources using search string *"toxic comment classification"* on 3th July 2020. We have used a reference management system *Zotero* where we added full texts of the articles to easy our systematic literature review. Our first search resulted in 202 extracted papers. After performing inclusion/exclusion criteria on titles and abstracts, 63 papers remained. After excluding the unrelated and duplicate works, 40 papers remained. The final selection was done by reading the whole text of the papers, and after this phase, we have selected 31 primary studies for our systematic review (Table 1).

ID	Paper title and reference
S1	A supervised multi-class multi-label word embeddings approach for toxic comment classification [5]
S2	Adversarial Text Generation for Googleś Perspective API [13]
S3	Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions [1]
S4	Automation in Social Networking Comments With the Help of Robust fastText and CNN [18]
S5	Avoiding Unintended Bias in Toxicity Classification with Neural Networks [21]
S6	Bangla Toxic Comment Classification (Machine Learning and Deep Learning Approach) [15]
S7	BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection [20]
S8	Challenges for Toxic Comment Classification: An In-Depth Error Analysis [32]
S9	Classification of Abusive Comments in Social Media using Deep Learning [2]
S10	Classification of Online Toxic Comments Using Machine Learning Algorithms [24]
S11	Classification of Online Toxic Comments Using the Logistic Regression and Neural Networks Models [28]
S12	Convolutional Neural Networks for Toxic Comment Classification [9]
S13	Cyberbullying ends here: Towards robust detection of cyberbullying in social media [33]
S14	Detecting Aggression and Toxicity using a Multi Dimension Capsule Network [30]
S15	Detecting Toxicity with Bidirectional Gated Recurrent Unit Networks [17]
S16	Detection of social network toxic comments with usage of syntactic dependencies in the sentences [29]
S17	Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection [31]
S18	Ensemble Deep Learning for Multilabel Binary Classification of User-Generated Content [10]
S19	Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning [12]
S20	Is preprocessing of text really worth your time for toxic comment classification? [19]
S21	LSTM neural networks for transfer learning in online moderation of abuse context [3]
S22	Machine Learning Suites for Online Toxicity Detection [22]
S23	On the Design and Tuning of Machine Learning Models for Language Toxicity Classification in Online Platform [26]
S24	Overlapping Toxic Sentiment Classification Using Deep Neural Architectures [27]
S25	Practical Significance of GA PartCC in Multi-Label Classification [23]
S26	Reading Between the Demographic Lines: Resolving Sources of Bias in Toxicity Classifiers [25]
S27	Stop illegal comments: A multi-task deep learning approach [8]
S28	Tackling Toxic Online Communication with Recurrent Capsule Networks [6]
S29	Towards non-toxic landscapes: Automatic toxic comment detection using DNN [7]
S30	Toxic comments identification in arabic social media [11]
S31	Using Sentiment Information for Preemptive Detection of Toxic Comments in Online Conversations [4]

Table 1: The list of selected primary studies

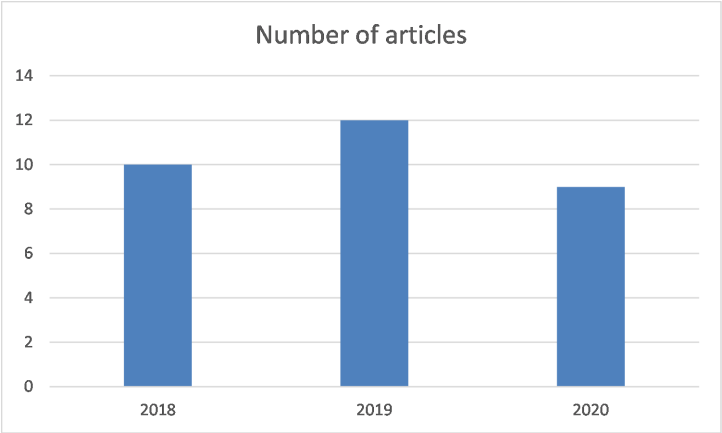

Figure 1: Selected primary studies per year

In the end, we extracted data from the 31 selected primary studies and did a synthesis taking into consideration the stated research questions. The results of our systematic literature review on toxic comment classification are shown in the next sections.

3 Results and discussion

3.1 Temporal overview of studies

The earliest relevant works on toxic comment classification were published in 2018 and their number significantly increased in 2019 (see Fig.1). There is a decrease in the number of studies in 2020, but this is due to the date when we performed the SLR search (3th July 2020, data are actually only for first halve of 2020).

3.2 Types of publications

The number of studies per publication type is demonstrated in Fig. 2. Most of the primary studies are conference papers (twenty), followed by arXiv papers (seven), three journal papers, and one book chapter.

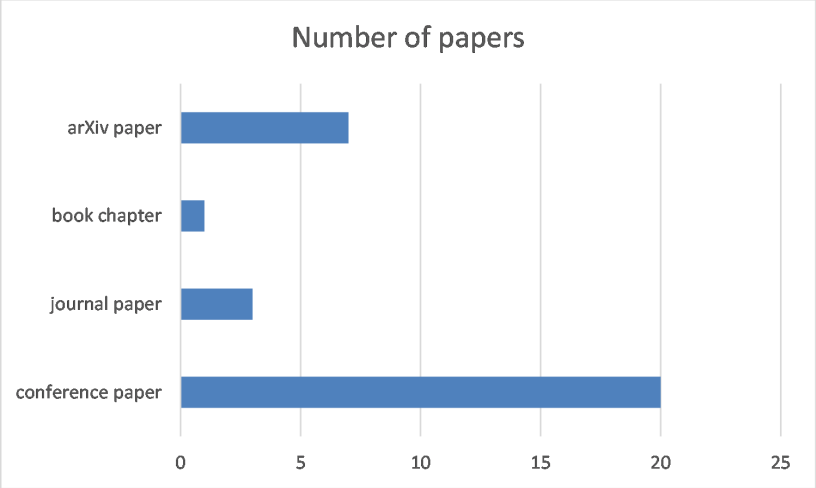

Figure 2: Papers per publication type

3.3 Used data set

Most of the primary studies have used one or two data sets with labelled toxic comments for a supervised machine learning. The most used data set is Jigsaw's data set hosted on Kaggle for Toxic Comment Classification Challenge competition [14]. It is used on twenty-two selected primary studies. This data set was later updated in Jigsaw Unintended Bias in Toxicity Classification Kaggle's competition. The mentioned data set contains a large number of Wikipedia comments which have been labelled by human raters for toxic behaviour. The types of toxicity are: toxic, severe toxic, obscene, threat, insult, and identity hate. Other used data sets are specific to a certain primary studies. Some of these other data sets are created by third parties: Twitter dataset by Davidson, Instagram dataset collected by Hosseinmardi et al., Semeval2018-Task 1 with almost 7000 tweets, The Twitter Hate Speech dataset, dataset of tweets made by members of the U.S. House of Representatives, Wikipedia Detox corpus, and live in-game chat conversations from a video game. The rest of datasets used in primary studies where created by authors of these studies: reviews taken from Udemy, synthetic training data by using Facebook comments that were posted in response to popular news articles, extensive collection of more than 104 million Reddit comments, a dataset taking comments from Facebook pages posts, 9.4K manually labelled entertainment news comments for identifying Korean toxic speech, comments in Hindi and English

both scraped from Facebook and Twitter, and custom prepared data in Arabic language from Facebook, Twitter, Instagram, and WhatsApp.

3.4 Used machine learning methods

Most of the selected primary studies have used more than one machine learning method to classify toxic comments from datasets mentioned in the previous subsection of this work. Table 2 shows in how many primary studies a specific machine learning method was used. The most used and effective methods are different deep neural networks, but often simpler and faster methods such as a logistic regression were used for baseline approaches.

Machine learning method	**Number of papers**
Convolutional neural network (CNN)	12
Logistic regression classifier	9
Bidirectional long short-term memory (BiLSTM)	8
Bidirectional Gated Recurrent Unit Networks (Bidirectional GRU)	6
Long Short Term Memory (LSTM)	5
Support Vector Machine (SVM)	5
Bidirectional Encoder Representations from Transformers (BERT)	4
Naive Bayes	4
Capsule Network	3
Random Forest	2
Decision tree	2
KNN classification	2
Gated Recurrent Unit (GRU)	2
Extreme Gradient Boosting (XGBoost)	2
Recurrent Neural Network (RNN)	2
Bi-GRU-LSTM	1
Gaussian Naive Bayes	1
Genetic Algorithms (GA)	1
Partial Classifier Chains (PartCC)	1

Table 2: Machine learning methods used in primary studies

3.5 Evaluation metrics

To evaluate results of using different machine learning methods to tackle problem of toxic comment classification, authors of primary studies use one or more evaluation metrics. Most used evaluation metrics are F1 score, accuracy, and area under the ROC curve (AUC ROC). All evaluation metrics used in selected primary studies are listed in Table 3.

Evaluation metric	Number of papers
F1 score	15
Accuracy	14
Area Under the ROC Curve (AUC ROC)	9
Custom AUC bias metric	2
Log loss	2
Hamming loss	2
False discovery rate	2
Mean precision	2
Mean recall	2
Pearson correlation coefficients	1
Specificity	1
Mean of the error rates	1
Generalized Mean Bias AUC	1
Subgroup AUC	1
BPSN AUC	1

Table 3: Used evaluation metrics

3.6 Classes of toxicity

Twenty-three of the primary studies have used six classes of toxicity defined in Jigsaw's data set hosted on Kaggle for Toxic Comment Classification Challenge competition [14]: toxic, severe toxic, obscene, threat, insult, and identity hate. Three papers used two classes (toxic or non-toxic). All used classes of toxicity are listed in Table 4.

Classes of toxicity	Number of papers
toxic, severe toxic, obscene, threat, insult, and identity hate	23
toxic or non-toxic	3
hate, offensive, and none	2
overtly aggressive, covertly aggressive, and non-aggressive	1
anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise and trust	1
racist content, sexist, and neutral	1

Table 4: Classes of toxicity identified in primary studies

3.7 Language of toxic comments

Only one paper (S14) analyse the toxic comment for two languages (in this case Hindi and English). Toxic comments in English are used the most (28

primary studies). The rest of studies dealt with Bangla (Bengali), Korean, and Arabic language.

4 Conclusions and future research ideas

Toxic comment classification is a complex research problem tackled by several machine learning methods. That is illustrated by many recent works in literature. After conducting a systematic review literature protocol proposed by Kitchenham and Charters [1], we have selected and analysed 31 primary studies. Our main conclusions are presented as answers to research questions as follows:
RQ1: When did the research on toxic comment classification become active in the research community? – The research on toxic comment classification become active recently, from 2018. It is due to release of Jigsaw's data set [33] that is mostly used in current related papers. From this year the number of paper is growing and this is an indicator that this research topic is actual and trendy.
RQ2: How is toxic comment classification research reported and what is the maturity level of the research in this field? – The most of the work are conference papers, so toxic comment classification is still a novel research topic.
RQ3: Which data sets are used to classify toxic comments? - The most used data set is the Jigsaw's data set hosted on Kaggle for Toxic Comment Classification Challenge competition [33]. Other data sets are mostly created from comments on popular social networks.
RQ4: Which machine learning methods are used to classify toxic comments? - The most used and effective methods are different architectures of deep neural networks, but often simpler and faster methods such as a logistic regression were used for baseline approaches.
RQ5: What are main evaluation metrics used to classify toxic comments? – Main evaluation metrics used to classify toxic comments are: F1 score, accuracy, and area under the ROC curve (AUC ROC).

The toxic comment classification research topic is a very active and challenging theme. Different transformers have recently shown a superior performance in many natural language processing tasks, so we recommend use of transformers for toxic comment classification in future works. Our systematic literature review identified three uses of BERT, but other transformers were not used yet (e.g. Hugging Face Transformers such as GPT-2, RoBERTa, XLM, DistilBert, XLNet... with pre-trained models in Tensor-

Flow 2.0 and PyTorch). Next, multilingual toxic comment classification is yet unsolved problem. In 2020, Jigsaw released multilingual toxic comment classification data set (https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data) that will be a basis for future work on this topic. The dataset is released under CC0, with the underlying comment text being governed by Wikipedia's CC-SA-3.0.

References

[1] H. Almerekhi, H. Kwak, J. Salminen, B. J. Jansen, Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions, *Proceedings of The Web Conference 2020*, Taipei, Taiwan, Apr. 2020, pp. 3033–3040. ⇒208

[2] M. Anand, R. Eswari, Classification of Abusive Comments in Social Media using Deep Learning, *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, Mar. 2019, pp. 974–977. ⇒208

[3] A. Bleiweiss, LSTM neural networks for transfer learning in online moderation of abuse context, *ICAART 2019 - Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, Prague, Czech Republic,2019, pp. 112–122. ⇒208

[4] É. Brassard-Gourdeau, R. Khoury, Using Sentiment Information for Preemptive Detection of Toxic Comments in Online Conversations, ArXiv200610145 Cs, Jun. 2020, Accessed: Jul. 03, 2020. http://arxiv.org/abs/2006.10145. ⇒208

[5] S. Carta, A. Corriga, R. Mulas, D. R. Recupero, R. Saia, A supervised multi-class multi-label word embeddings approach for toxic comment classification, *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Vienna, Austria, 2019, pp. 105–112. ⇒208

[6] A. G. D'Sa, I. Illina, D. Fohr, Towards non-toxic landscapes: Automatic toxic comment detection using DNN, ArXiv191108395 Cs Stat, Nov. 2019, Accessed: Jul. 03, 2020. http://arxiv.org/abs/1911.08395. ⇒208

[7] S. Deshmukh, R. Rade, Tackling Toxic Online Communication with Recurrent Capsule Networks, *2018 Conference on Information and Communication Technology (CICT)*, Jabalpur, India, 2018. ⇒208

[8] A. Elnaggar, B. Waltl, I. Glaser, J. Landthaler, E. Scepankova, F. Matthes, Stop Illegal Comments: A Multi-Task Deep Learning Approach, *ACM International Conference Proceeding Series*, 2018, pp. 41–47. ⇒208

[9] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, V. P. Plagianakos, Convolutional Neural Networks for Toxic Comment Classification, *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, Patras, Greece, Jul. 2018, pp. 1–6. ⇒208

[10] G. Haralabopoulos, I. Anagnostopoulos, D. McAuley, Ensemble Deep Learning for Multilabel Binary Classification of User-Generated Content, *Algorithms*, **13**, 4 (2020). ⇒208

[11] O. Hosam, Toxic comments identification in arabic social media, *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, **11**, (2019) 219–226. ⇒208

[12] M. Ibrahim, M. Torki, N. El-Makky, Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning, *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, Orlando, USA, 2018, pp. 875–878. ⇒208

[13] E. Jain et al., Adversarial Text Generation for Google's Perspective API, *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, USA, Dec. 2018, pp. 1136–1141. ⇒208

[14] Jigsaw, Data for Toxic Comment Classification Challenge. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data`

[15] A. N. M. Jubaer, A. Sayem, Md. A. Rahman, Bangla Toxic Comment Classification (Machine Learning and Deep Learning Approach), *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, Nov. 2019, pp. 62–66. ⇒210, 212

[16] B. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering(2007). ⇒208

[17] V. Kumar, B. K. Tripathy, Detecting Toxicity with Bidirectional Gated Recurrent Unit Networks, *Adv. Intell. Syst. Comput.*, vol. **1034**,(2020) 591–600. ⇒206

[18] S. Mestry, H. Singh, R. Chauhan, V. Bisht, K. Tiwari, Automation in Social Networking Comments With the Help of Robust fastText and CNN, *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, Chennai, India, Apr. 2019, pp. 1–4. ⇒208

[19] F. Mohammad, Is preprocessing of text really worth your time for toxic comment classification?, *CSCE 2018 - Proceedings of the 2018 International Conference on Artificial Intelligence, ICAI 2018*, Las Vegas, USA, 2018, pp. 447–453. ⇒208

[20] J. Moon, W. I. Cho, J. Lee, BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection, ArXiv200512503 Cs, May 2020, Accessed: Jul. 03, 2020. http://arxiv.org/abs/2005.12503. ⇒208

[21] S. Morzhov, Avoiding Unintended Bias in Toxicity Classification with Neural Networks, *2020 26th Conference of Open Innovations Association (FRUCT)*, Yaroslavl, Russia, Apr. 2020, pp. 314–320. ⇒208

[22] D. Noever, Machine Learning Suites for Online Toxicity Detection, ArXiv181001869 Cs Stat, Oct. 2018, Accessed: Jul. 03, 2020. http://arxiv.org/abs/1810.01869. ⇒208

[23] A. P. Patil, A. Mohammed, G. Elachitaya, M. Tiwary, Practical Significance of GA PartCC in Multi-Label Classification, *Proceedings of the 2019 Ieee Region 10 Conference (tencon 2019): Technology, Knowledge, and Society*, Kerala, India, 2019, pp. 2487–2490. ⇒208

[24] Rahul, H. Kajla, J. Hooda, G. Saini, Classification of Online Toxic Comments Using Machine Learning Algorithms, *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, May 2020, pp. 1119–1123. ⇒208

[25] E. Reichert, H. Qiu, J. Bayrooti, Reading Between the Demographic Lines: Resolving Sources of Bias in Toxicity Classifiers, ArXiv200616402 Cs, Jun. 2020, Accessed: Jul. 03, 2020. http://arxiv.org/abs/2006.16402. ⇒208

[26] M. Rybinski, W. Miller, J. Del Ser, M. Nekane Bilbao, J. F. Aldana-Montes, On the Design and Tuning of Machine Learning Models for Language Toxicity Classification in Online Platforms, *Intelligent Distributed Computing Xii*, **798** (2018), pp. 329–343. ⇒208

[27] H. H. Saeed, K. Shahzad, F. Kamiran, Overlapping Toxic Sentiment Classification using Deep Neural Architectures, *2018 18th Ieee International Conference on Data Mining Workshops (icdmw)*, Sentosa, Singapore, 2018, pp. 1361–1366. ⇒208

[28] M. A. Saif, A. N. Medvedev, M. A. Medvedev, T. Atanasova, Classification of Online Toxic Comments Using the Logistic Regression and Neural Networks Models, *Proceedings of the 44th International Conference Applications of Mathematics in Engineering and Economics*, Sozopol, Bulgaria, 2018. ⇒208

[29] S. Shtovba, O. Shtovba, M. Petrychko, Detection of social network toxic comments with usage of syntactic dependencies in the sentences, *CEUR Workshop Proceedings*, Otzenhausen, Germany, 2019, pp. 313–323. ⇒208

[30] S. Srivastava, P. Khurana, Detecting Aggression and Toxicity using a Multi Dimension Capsule Network. Stroudsburg: Assoc Computational Linguistics-Acl, 2019, pp. 157–162. ⇒208
⇒208

[31] A. Vaidya, F. Mai, Y. Ning, Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection, ArXiv190909758 Cs, Mar. 2020, Accessed: Jul. 03, 2020. http://arxiv.org/abs/1909.09758. ⇒208

[32] B. van Aken, J. Risch, R. Krestel, A. Löser, Challenges for Toxic Comment Classification: An In-Depth Error Analysis, ArXiv180907572 Cs, Sep. 2018, Accessed: Jul. 03, 2020. http://arxiv.org/abs/1809.07572. ⇒208

[33] M. Yao, C. Chelmis, D.-S. Zois, Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media, *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, San Francisco, USA, 2019, pp. 3427–3433. ⇒208

Received: July 23, 2020 • Revised: October 5, 2020