

# An Overview of Toxic Content Datasets for Artificial Intelligence Applications to Educate Students Towards a Better Use of Social Media

Sara Havzi<sup>1,2</sup> and Davide Taibi<sup>1</sup>

<sup>1</sup>*Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, 90146 Palermo, Italy*

<sup>2</sup>*Faculty of Technical Sciences, University of Novi Sad, Serbia*

**Keywords:** Online Toxic Content, Datasets, Artificial Intelligence, Education.

**Abstract:** The Internet has become an integral part of life, providing numerous benefits to its users. However, due to freedom of speech and lack of control, the Internet is becoming a breeding ground for spreading harmful/toxic content. Since young people are the most active Internet users, protecting them from harmful online content is extremely important. One of the directions within which this could be conducted is educating young people about the consequences of using online toxic language and building powerful artificial intelligence-based tools such as Virtual Learning Companions that could educate youth in recognising online toxic content and upgrading their social media and self-protection competencies. To be able to build such tools, quality online datasets are needed. This paper is a brief overview of 9 selected English language online toxic content datasets published between 2020 and 2022 among 70 we found in the literature that could help educate young people on this topic.

## 1 INTRODUCTION


Online toxic content is becoming more frequent and widespread on social networks and in the online world. The hatred directed towards members of certain ethnic groups, races or religions can be conveyed through any form of expression. Represented on different platforms through comments, pictures, memes, cartoons and movies, gestures and videos can be spread offline or online. Recent studies have shown that adolescents are particularly susceptible to the influence of toxic content, fake news, and online hate speech (Boer et al., 2020, Kansok-Dusche et al. 2022). This is motivated by the fact that adolescents constitute the highest percentage of social network users, especially TikTok (Zheluk et al., 2022). For example, a report by the Pew Research Center (Anderson et al., 2022) found that adolescents who use social media are more likely to be exposed to false information and are less likely to be able to distinguish between fact and fiction. One of the most common effects of toxic


content exposure on adolescents is the impact on their well-being with direct consequences on their engagement, participation, and performances in school contexts, as well as a more severe possibility of developing mental health problems.

(Nixon, 2014) demonstrated that cyberbullying has a significant impact on adolescents' health. The author highlights cyberbullying as an emerging international public health concern.

Studying and preventing online toxicity is especially important as exposure to its various forms can negatively affect mental health (Baier et al., 2019, Nixon, 2014), increase the risk of some serious mental health concerns, anxiety, stress, depression, and suicidal thinking (Martínez-Monteagudo et al., 2020). In most extreme cases, online abuse can lead to suicide attempts (Hinduja & Patchin, 2019). Moreover, exposure to these issues can affect the students' well-being in the school context thus critically influencing the learning performance of students (Al-Rahmi et al., 2022).

In light of these concerns, it is crucial to develop tools and resources that can support students in

<sup>a</sup> <https://orcid.org/0000-0001-7077-8780>

<sup>b</sup> <https://orcid.org/0000-0002-0785-6771>

counteracting social media threats by increasing their awareness of social media threats and supporting the implementation of learning activities aimed to educate adolescents about the dangers of online toxic content, fake news, and hate speech. This can include educational tools, resources, and apps that help adolescents identify and avoid harmful content and teach them how to evaluate the information they find online critically. To this aim, new trends and approaches in the Artificial Intelligence research domain could play a key role.

Additionally, it is essential to involve educators and parents in educating adolescents about these issues. This can include providing training and resources for educators in their classrooms and partnering with parents to develop strategies for talking to their children about these issues.

Since a large part of the population is relying on media as their primary information source (Lazer et al., 2018), it is of high priority to build tools to detect toxic content on social media, furthermore to educate society, especially the young population to be able to detect such content.

In recent years, researchers are developing a range of artificial intelligence (AI) and machine learning (ML) algorithms to detect online harmful content of different kinds, as well as tools and companions to teach the public about different kinds of online toxic content and toxic content spreading and understand how social media toxic content is affecting the population (Baru et al., 2019), (Roberto Sanchez Reina et al., 2022).

To have these algorithms and tools to effectively fight harmful content on social media, data quality is a prerequisite for the quality of entire systems. Authors are highlighting how crucial it is to have high-quality datasets for developing machine learning models (Gudivada et al., 2017), (Jain et al., 2020), (Paullada et al., 2021) Authors (Gudivada et al., 2017) indicate that outliers in training datasets can cause instability or non-convergence in ensemble learning, as well as that “incomplete, inconsistent, and missing data can lead to drastic degradation in prediction” (Gudivada et al., 2017).

There is no standard definition of online toxic content, hate speech or fake news. Moreover, there is no standard classification of terms related to online toxic content. Therefore, we consider toxic content as any online content that is harmful, abusive, offensive and has or might harm an individual, group of individuals, society or organisation.

It is very important to build online toxic content datasets to be able to develop algorithms for the detection of such phenomena, as well as to educate

the population, especially young people about this harmful content.

These datasets are important in education for several reasons the availability of labelled datasets concerning toxic content can help:

- educators and researchers to identify and classify these forms of content. These datasets are a fundamental part to develop tools and resources for identifying and avoiding these types of content, and to train machine learning models to automatically detect and classify potentially harmful content,
- researchers to study the prevalence and effects of these forms of content. Datasets can be used to inform the development of educational software and resources, that support the understanding of the impact of these forms of content with a particular focus on adolescents, and
- educators should develop and update traditional curricula by introducing social media literacy activities to educate students on how to identify and cope with these forms of content and critically evaluate the information they find online.

In this paper, we present a study in which the literature between 2020 and 2022 has been analysed to detect the most relevant datasets focused on toxic content, that can be used to support the development of Artificial Intelligence applications aimed at increasing the student’s awareness of social media threats. In the inclusion criteria for selecting these datasets, particular emphasis has been oriented on the studies that take reproducible features into account. More details will be provided in section 3, but in principle, these features assure that the dataset made available in these studies respects quality criteria such as the availability of documentation concerning their use. For our study, only datasets in the English language have been considered.

## 2 RELATED WORKS

Many of the open datasets found on popular Internet platforms such as Kaggle or Google Dataset Search have a great possibility of being unreliable. Very often, there is no information about the data collection method or annotation. However, since datasets of good quality are prerequisites in developing quality online toxic content detection systems based on machine learning, many researchers are creating datasets in the field.

Moreover, most researchers observe only one category of online toxic content, specifically hate speech, which is the most frequent. As far as we know, a systematic review of the literature concerning dataset for online toxic content detection compares the different datasets and highlight their use in the educational context. However, we found some papers summarising the existing datasets and some systematic literature reviews in the field that addressed the most common datasets.

Zhang and Ghorbani (Zhang & Ghorbani, 2020) conducted an overview of online fake news, with a short overview of the existing online fake news datasets. According to the authors, there is a lack of labelled fake news datasets which is a prerequisite for building an effective detection system for online misleading information. The addressed datasets were published from 2009 to 2019. Authors conclude that most of the datasets are small, that truthfulness is predominantly scaled as true or false and that most datasets collect news spontaneously published by human creators on mainstream media.

D'Ulizia et al. (D'Ulizia et al., 2021) conducted a survey of evaluation online fake news datasets. They systematically reviewed twenty-seven popular datasets for fake news detection by doing a comparative analysis. This research is a rare survey that addresses the detailed analysis of datasets as a primary objective of the research.

Poletto et al. (Poletto et al., 2021) did a thorough systematic review of resources and benchmark corpora for hate speech detection. Authors systematically analyse the resources made by the community available in different languages. Authors were looking for papers published until 2020. In this paper, lexica, corpora or both were addressed, and 11 benchmark datasets were identified among those. Most of the corpora were from Twitter (24), next are news websites (6), then Reddit with 5 corpora. The annotation strategy is mostly binary schemed.

Interestingly, many corpora didn't provide guidelines for the annotators. Most of the datasets that authors found provided links to the datasets, with some requiring user registration to be available. The authors highlighted the risk of creating too many biased systems or overfitting due to the problem of the classification of hate speech. The authors mentioned the possibility of bias in annotated data as well.

Alkomah and Ma (Alkomah & Ma, 2022) did a literature review of textual hate speech detection methods and datasets. The authors pointed out that many datasets are of poor quality since they are not regularly updated, and many datasets have overlaps

in data. However, the authors were only looking at Twitter datasets in English, which significantly reduces the comprehensiveness of the research. In addition, Alkomah and Ma highlight the potential bias problem in building large datasets due to the manually performed annotation process.

Many short or thorough research papers address the datasets in the field of online toxic content. However, authors mostly focus on one type of online toxic content such as online fake news or online hate speech, choosing different synonyms for these types.

Online toxic content is increasing interest among the research community and more literature reviews addressing the datasets appeared in the last two years (Madukwe et al., 2020; Raponi et al., 2022).

### 3 AI APPLICATIONS FOR EDUCATIONAL PURPOSES

Artificial intelligence (AI) has great power in changing the way we think and learn, with an increasing application in various fields.

AI has proven to be increasingly applicable in various ways within education and is more and more often used for educational purposes in the last few years. In 2020, (Chen et al., 2020) conducted research of comprehensive review of Artificial Intelligence in Education (AIEd) and highlighted that there is an increasing interest in and impact of AIEd research, as well as those traditional AI technologies were commonly adopted in the educational context, but that advanced techniques were rarely adopted. In the past years, AI-built-in virtual agents for simulation-based learning is becoming more popular (Dai & Ke, 2022) AI enables personalised learning for students (Chen et al., 2020; Della Ventura, 2017).

The usage of AI in education, as in other fields has numerous perspectives. For example, when it comes to social media, AI can help to provide content to social media users, detecting harmful content on social media and supporting users by preventing them from seeing harmful content on social media (Taibi, et al, 2021). However, due to different social media policies and free speech, these tools don't always detect toxic online content. Therefore, one key role of AI solutions in detecting toxic and harmful content on social media is creating powerful educational tools that could support and educate users, especially regarding building their social literacy competencies and self-protection skills.

Since there is a rapid growth of concerns regarding the bad influence of online toxic content on

social media, researchers are developing more tools and experiments to educate young people and increase their social media literacy.

## 4 METHODOLOGY

We conducted a mapping review to see which terms are used in the scientific community referring to online toxic content.

There is no standard definition of online toxic content, hate speech or fake news. Moreover, there is no standard classification of terms related to online toxic content. Equally important, hate speech is usually treated the same as cyberbullying, where only online hate speech could be considered a synonym for cyberbullying since offline hate speech is not the same as hate speech in the online world (Fulantelli et al., 2022). Therefore, we use toxic online content as the umbrella term for all the terms related to online fake news and online hate speech.

Next, we conducted an umbrella review to detect existing systematic reviews in the field. Finally, we searched for novel machine learning methods in online toxic content detection in Scopus and Web of Science, trying to identify datasets in the field. This paper presents the part of the wider study focused on identifying online toxic content datasets.

We identified more than 70 datasets in our research but chose to present the ones that:

- are publicly available datasets
- dataset papers published between 2020. and 2022.
- for this paper, we chose only datasets written in English, and,
- we only chose the datasets presented in the studies where authors explained the annotation guidelines (in hate speech datasets) or the data method of deciding truthfulness (in the fake news datasets).

The results were nine datasets divided into two main online toxic content categories – “(online) hate speech” and “(online) fake news”.

## 5 ONLINE TOXIC CONTENT DATASETS

Dataset names and links are provided in Table 1.

Table 1: Dataset names and links.

Dataset Name	Link
FAKENEWS NET	<a href="https://github.com/KaiDMML/FakeNewsNet">https://github.com/KaiDMML/FakeNewsNet</a>
NELA-GT-20	<a href="https://github.com/MELALab/nela-gt">https://github.com/MELALab/nela-gt</a>
DGHSD	<a href="https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset">https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset</a>
HATEXPLAIN	<a href="https://github.com/hate-alert/HateXplain">https://github.com/hate-alert/HateXplain</a>
NJH	<a href="https://bit.ly/dataset-NJH">https://bit.ly/dataset-NJH</a>
ETHOS	<a href="https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset">https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset</a>
CONVABUSE	<a href="https://github.com/amandacurry/convabuse">https://github.com/amandacurry/convabuse</a>
UCBERKLEY-MHS	<a href="https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech">https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech</a>
FAKEEDIT	<a href="https://github.com/entitize/fakeddit">https://github.com/entitize/fakeddit</a>

FakeNewsNet - (Shu et al., 2020) presented a multiple dimension fake news data repository containing two datasets. The authors used Politifact and GossipCop to collect truth labels for the data. FakeNewsNet repository has one of the largest and most comprehensive online fake news datasets since it has featured in news content, social context and spatiotemporal information. This means that these datasets can be used in a wider spectrum, especially in understanding fake news propagation and fake news intervention.

NELA-GT-20 - (Gruppi et al., 2021) is an update of the NELA-GT-19 dataset (Gruppi et al., 2020). It is a fake news dataset with more than 1.8 million news articles from 519 sources collected in 2020. Different from the previous version of datasets, NELA-GT-20 has tweets embedded. NELA-GT-20 has almost double more sources than NELA-GT-19.

DGHSD - (Vidgen et al., 2021) published an online hate speech dataset, which we named DGHSD for this paper. This dataset consists of nearly 40 000 entries generated and labelled by trained annotators. Hateful entries take up about 54% of DGHSD. The authors provided detailed guidelines for the 20 recruited annotators, which two expert annotators oversaw in online hate speech..

HATEXPLAIN - (Mathew et al., 2020) presented HateXplain, the benchmark online hate speech dataset. As sources, authors used Twitter and Gab<sup>3</sup>. Authors did not consider reposts or posts containing links, pictures, or videos, but emojis were included. Annotation is done at three levels – classification of

<sup>3</sup> <https://gab.com/>

text as hate speech or offensive, annotating the target groups of the hate speech, and annotating parts of the text which are the reason for the given annotation. Annotators had brief instructions for annotating. The pilot task had 621 annotators, and 253 were selected for the main task. Three annotators annotated each post. The dataset consists of more than 20 000 entries.

NJH – (Bianchi et al., 2022) is an online hate speech dataset with more than 40.000 tweets about immigration. Authors downloaded over 150M tweets between 2020-21. The authors made annotation guidelines for 10 undergraduate researchers from the University of Liverpool (UK) and Syracuse University (US). Annotators were trained until they achieved satisfactory reliability. NJH is a well-labelled dataset with 7 types of labels. Due to deleting or suspending tweets, or private accounts, approximately 25% of tweets in this dataset are no longer available.

ETHOS - (Mollas et al., 2020) published an online hate speech dataset based on YouTube and Reddit comments. ETHOS is a textual dataset with two variants – binary and multi-label, with a thorough protocol. Binary ETHOS is consisted of 998 entries, while a multi-label ETHOS is considered 433 hate speech messages. Targets of this dataset are wide – from religion to gender, disability, race etc.

CONVABUSE - (Curry et al., 2021) presented the first English study on online hate speech towards three AI systems. We decided to include this dataset due to its originality of source since most datasets are focused on social media. The authors made data from two systems publicly available since the third one wasn't released due to privacy reasons. Two systems were text-based chatbots, while the third system was voice-based. Annotators were gender studies students. Annotators were deciding on abusive, or non-abusive content, abuse severity, and types such as ableism, homophobia, intellectual, racism, sexism, sexual harassment, transphobia and general.

UCBERKLEY-MHS - (Kennedy et al., 2020) presented a general method for measuring complex variables on a continuous interval spectrum by combining supervised deep learning with the Constructing Measures approach to faceted Rasch item response theory (IRT) (Furr & Bacharach, 2007) In their work, they used this method on 50000 entries online hate speech dataset. For the annotators, they used Amazon Mechanical Turk and 10000 annotators. Authors made scale items for labeling, proposing guidelines for the annotators through questionnaire, mostly using Likert-scale. Sources that

authors used for the dataset were YouTube, Twitter and Reddit.

FAKEEDIT - (Nakamura et al., 2020) is one of the largest online fake news datasets. Not only that – this dataset has over 1 million samples in forms of text, comments data, images and metadata. The dataset was multi-labelled and enables fine-grained online fake news classification. FakeEdit provides implicit fact-checking, and as a source, authors used 22 different subreddits from Reddit. Authors used 2-way, 3-way and 6-way labelling. The first layer of labelling was classifying sample as true or false. The 3-way labelling with having “middle” label as “the sample is fake and contains text that is true”. The final labelling was categorising the fake news type: true, satire/parody, misleading content, imposter content, false connection and manipulated content.

In Table 2, publications and the year of publication for the chosen datasets are shown.

Table 2: Datasets and publications.

Dataset Name	Publication	Year
FAKENEWSNET	(Shu et al., 2020)	2020
NELA-GT-20	(Gruppi et al., 2021)	2020
DGHSD	(Vidgen et al., 2021)	2021
HATEXPAIN	(Mathew et al., 2020)	2022
NJH	(Bianchi et al., 2022)	2022
ETHOS	(Mollas et al., 2020)	2021
CONVABUSE	(Curry et al., 2021)	2021
UCBERKLEY-MHS	(Kennedy et al., 2020)	2020
FAKEEDIT	(Nakamura et al., 2020)	2020

Our review identified more than 70 datasets, but most were before 2019. and were either mentioned in the related work or widely used in the research community. Many datasets we found reported in the literature were not publicly available anymore. Most datasets were published in 2016, and an even higher number of covid-19 related datasets from 2019. Some authors are connecting this with the US Elections 2016 (Varma et al., 2021).

Most online toxic content datasets have Twitter as a source since it is the easiest source to gather data. However, Twitter also has some limitations, some as privacy concerns, and some are connected to the decay of tweets over time due to deleted, suspended tweets or private accounts. This is a very common issue with Twitter datasets (Tromble et al., 2017) Researchers are more aware of the importance of building datasets with different sources, for example, (Curry et al., 2021) showed the importance of

Table 3: Datasets, platforms and sizes.

Dataset Name	Platform	Size
FAKENEWSNET	Politifact, GossipCop	300 000+
NELA-GT-20	519 media sources, Twitter	1.8 million
DGHSD	Synthetically generated	40000+
HATEXPLAIN	Twitter, Gab	20000+
NJH	Twitter	40000+
ETHOS	YouTube, Reddit	998
CONVABUSE	Facebook Messenger Chatbot and E.L.I.Z.A. chatbots	4500+
UCBERKLEY- MHS	Twitter, Reddit, YouTube	41000
FAKEEDIT	Reddit	185445

exploring AI systems and chatbots as sources. The datasets we on our list show that there are more datasets created with sources different from Twitter.

Annotation is one of the most important tasks of creating a dataset. Given that this is often a manual human work, there is a high probability of bias and the annotators will be biased. That's why guidelines for annotators are so important, as well as the number of annotators. In annotating, there can be several problems such as not understanding the difference between terms. For example, (Davidson et al., 2017) proposed an error-analysis on annotations as well as classifier performance with arguments that in the field of hate speech (toxic content as well) there are unclear definitions, and homophobic and racist terms are identified more frequently than sexist. (Sap et al., n.d.) found strong evidence that extra attention should be paid to dialect as one possible factor of racial biases. Kocon et al. (Kocoń et al., 2021) claim that separating annotator groups significantly impacts the performance of hate detection systems. Authors also provide an annotator ID with each entry in their dataset, to enable further research since online hate speech is sensitive and complex topic (Ognibene & Taibi, 2022).

### 5.1 Datasets and AI in the Purpose of Education

Datasets of good quality are very important in building powerful tools that could help educate society in recognising and dealing with the online toxic content, since the correct data is a prerequisite to good quality tools. Datasets could have numerous and different applications in the course of education.

For example, building powerful guidebots or virtual agents that could help young people learn to distinct toxic content from non-toxic content on social media, could have a great educational impact in educating students. In order to empower young people to build social media competences and self-protection skills, AI based solutions that are using quality datasets are of particular importance.

Virtual learning companions (VLC) could be a good way for empowering adolescents regarding the threats of social media. Researchers recognised the need for such companions, and one example is the COURAGE project VLC. Even though social media platforms use their AI algorithms to detect some types of online toxic content, their priority is still focused on highly clicked content and viral content, as well as to generate as much traffic as possible. COURAGE VLC, based on AI, comprises adaptive detectors of content and network threats, user models to support personalised interventions as well as content and educational activity recommendations (Ognibene & Taibi, 2022).

## 6 CONCLUSIONS

Online toxic content is gaining momentum and scientists are making efforts to counteract or at least reduce the occurrence of toxic content on the Internet. The development of machine learning methods for these purposes is becoming more and more popular, and the quality of the data is a prerequisite for the progress and development of these methods.

It is not only important to develop tools and algorithms to detect online toxic content, but it is essential to educate the students on these concerns and to develop tools and resources that could be used for educational purposes, in order for the users to be able to identify and avoid harmful content.

Therefore, it is critical to address and build the online toxic content datasets, as well as to conduct the quality criteria for the data.

It is of high priority to building the datasets from different sources, and since in our literature research, we haven't found any TikTok datasets, which would play an important role in educating students and young people since TikTok is the most downloaded application in 2021, as well as it is medium primarily used by young people under 30 years old (Zheluk et al., 2022) In order to be able to educate and build social media competencies, datasets from crucial sources should be built and publicly available for the researchers to build educational tools.

An important role in the quality of datasets is played by annotators, the source of the data and the data collection method. Hence, detailed annotators guidelines and training are needed in order to avoid bias. It is important to document the demographic characteristics of the annotators, so that possible bias can be verified and identified, as well as to give the annotators precise definitions of the terms they are manually classifying. Here we come to another challenge for the research community dealing with online toxic content and that is the formal definition of the terms. Authors use different definitions and names for types of online toxic content and this leads to confusion, therefore datasets may be invalid.

Last, but not least, it is extremely important to promote reproducibility within research in this area, to ensure that algorithms really detect online toxic content, that datasets are valid, and to create an opportunity to thereby educate people on best practices and the most effective examples.

In this paper, we present nine datasets that we identified amongst 70 datasets we found, and are publicly available, have proposed their annotation guidelines and are related to online toxic content.

The future work in the educational context should be building educational tools using proposed datasets.

## ACKNOWLEDGEMENTS

This work was developed in the framework of the project COURAGE—A social media companion safeguarding and educating students (Nos. 95567) and funded by the Volkswagen Foundation in the topic Artificial Intelligence and the Society of the Future.

## REFERENCES

- Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. In *Information (Switzerland)* (Vol. 13, Issue 6). MDPI. <https://doi.org/10.3390/info13060273>
- Anderson, M., Vogels, E. A., Perrin, A., & Rainie, L. (2022). *Connection, Creativity and Drama : Teen Life on Social Media in 2022. November.*
- Baier, D., Hong, J. S., Kliem, S., & Bergmann, M. C. (2019). Consequences of Bullying on Adolescents' Mental Health in Germany: Comparing Face-to-Face Bullying and Cyberbullying. *Journal of Child and Family Studies*, 28(9), 2347–2357. <https://doi.org/10.1007/s10826-018-1181-6>
- Baru, C., Institute of Electrical and Electronics Engineers, & IEEE Computer Society. (2019). *2019 IEEE International Conference on Big Data : proceedings : Dec 9 - Dec 12, 2019, Los Angeles, CA, USA.*
- Bianchi, F., Hills, S. A., Rossini, P., Hovy, D., Tromble, R., & Tintarev, N. (2022). "It's Not Just Hate": A Multi-Dimensional Perspective on Detecting Harmful Speech Online." <http://arxiv.org/abs/2210.15870>
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. In *Computers and Education: Artificial Intelligence* (Vol. 1). Elsevier B.V. <https://doi.org/10.1016/j.caeai.2020.100002>
- Curry, A. C., Abercrombie, G., & Rieser, V. (2021). *ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI.* <https://github.com/amandacurry/convabuse>.
- Dai, C. P., & Ke, F. (2022). Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. In *Computers and Education: Artificial Intelligence* (Vol. 3). Elsevier B.V. <https://doi.org/10.1016/j.caeai.2022.100087>
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language.* [www.aai.org](http://www.aai.org)
- Della Ventura, M. (2017). Creating Inspiring Learning Environments by means of Digital Technologies: A Case Study of the Effectiveness of WhatsApp in Music Education. *EAI Endorsed Transactions on E-Learning*, 4(14), 152906. <https://doi.org/10.4108/eai.26-7-2017.152906>
- D'Ulizia, A., Caschera, M. C., Ferri, F., & Grifoni, P. (2021). Fake news detection: A survey of evaluation datasets. *PeerJ Computer Science*, 7, 1–34. <https://doi.org/10.7717/PEERJ-CS.518>
- Furr, R. M., & Bacharach, V. R. (2007). Item Response Theory and Rasch Models. *Psychometrics: An Introduction*, 314–334.
- Gruppi, M., Horne, B. D., & Adali, S. (2020). *NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles.* <http://arxiv.org/abs/2003.08444>
- Gruppi, M., Horne, B. D., & Adali, S. (2021). *NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles.* <http://arxiv.org/abs/2102.04567>
- Gudivada, V. N., Ding, J., & Apon, A. (2017). *Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations Big Data Management View project Research Topic "Data Quality for Big Data and Machine Learning" in Frontiers in Big Data View project Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations.* <https://www.researchgate.net/publication/318432363>
- Hinduja, S., & Patchin, J. W. (2019). Connecting Adolescent Suicide to the Severity of Bullying and Cyberbullying. *Journal of School Violence*, 18(3), 333–346. <https://doi.org/10.1080/15388220.2018.1492417>
- Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2020). Overview and Importance of

- Data Quality for Machine Learning Tasks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3561–3562. <https://doi.org/10.1145/3394486.3406477>
- Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). *Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application*. <http://arxiv.org/abs/2009.10277>
- Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing and Management*, 58(5). <https://doi.org/10.1016/j.ipm.2021.102643>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Madukwe, K., Gao, X., & Xue, B. (2020). *In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets*. 150–161. <https://doi.org/10.18653/v1/2020.alw-1.18>
- Martínez-Monteaquedo, M. C., Delgado, B., Díaz-Herrero, Á., & García-Fernández, J. M. (2020). Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying. *Psychiatry Research*, 286. <https://doi.org/10.1016/j.psychres.2020.112856>
- Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). *HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection*. <http://arxiv.org/abs/2012.10289>
- Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2020). *ETHOS: an Online Hate Speech Detection Dataset*. <https://doi.org/10.1007/s40747-021-00608-2>
- Nakamura, K., Levy, S., & Wang, W. Y. (2020). *r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection*. <https://www.journalism.org/2019/06/05/many-americans->
- Nixon, C. (2014). Current perspectives: the impact of cyberbullying on adolescent health. *Adolescent Health, Medicine and Therapeutics*, 143. <https://doi.org/10.2147/ahmt.s36456>
- Ognibene, D., & Taibi, D. (2022). *Designing Educational Interventions to Increase Students' Social Media Awareness-Experience From the COURAGE Project* <https://www.researchgate.net/publication/366595820>
- Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *Patterns* (Vol. 2, Issue 11). Cell Press. <https://doi.org/10.1016/j.patter.2021.100336>
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. In *Language Resources and Evaluation* (Vol. 55, Issue 2, pp. 477–523). Springer Science and Business Media B.V. <https://doi.org/10.1007/s10579-020-09502-8>
- Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). *A Benchmark Dataset for Learning to Intervene in Online Hate Speech*. <http://arxiv.org/abs/1909.04251>
- Raponi, S., Khalifa, Z., Oligeri, G., & di Pietro, R. (2022). Fake News Propagation: A Review of Epidemic Models, Datasets, and Insights. *ACM Trans. Web*, 16(3). <https://doi.org/10.1145/3522756>
- Roberto Sanchez Reina, J., Scifo, L., & Lomonaco, F. (2022). *Empirically Investigating Virtual Learning Companions to Enhance Social Media Literacy*. <https://www.researchgate.net/publication/365683043>
- Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. A., & Allen, P. G. (n.d.). *The Risk of Racial Bias in Hate Speech Detection*. Association for Computational Linguistics. [www.figure-eight.com](http://www.figure-eight.com)
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). *FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media*. <http://arxiv.org/abs/1809.01286>
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3), 171–188. <https://doi.org/10.1089/big.2020.0062>
- Taibi, D., Fulantelli, G., Monteleone, V., Schicchi, D., & Scifo, L. (2021). An Innovative Platform to Promote Social Media Literacy in School Contexts. In *Proceeding of ECEL 2021 20th European Conference on e-Learning*, Berlin, Germany.
- Tromble, R., Storz, A., & Stockmann, D. (2017). *We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference*. <https://ssrn.com/abstract=3079927> [Electronic copy available at: https://ssrn.com/abstract=3079927](https://ssrn.com/abstract=3079927)
- Varma, R., Verma, Y., Vijayvargiya, P., & Churi, P. P. (2021). A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic. In *International Journal of Intelligent Computing and Cybernetics* (Vol. 14, Issue 4, pp. 617–646). Emerald Group Holdings Ltd. <https://doi.org/10.1108/IJCC-04-2021-0069>
- Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2021). *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection*. <https://github.com/bvidgen/>
- Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterisation, detection, and discussion. *Information Processing and Management*, 57(2). <https://doi.org/10.1016/j.ipm.2019.03.004>
- Zheluk, A. A., Anderson, J., & Dineen-griffin, S. (2022). *Adolescent Anxiety and TikTok: An Exploratory Study*. 14(December 2021). <https://doi.org/10.7759/cureus.32530>