IRIE
International Review of Information Ethics Vol. 29 (03/2021)

Diogo Cortiz, Arkaitz Zubiaga

Ethical and technical challenges of AI in tackling hate speech

Abstract:
In this paper, we discuss some of the ethical and technical challenges of using Artificial Intelligence for online content moderation. As a case study, we used an AI model developed to detect hate speech on social networks, a concept for which varying definitions are given in the scientific literature and consensus is lacking. We argue that while AI can play a central role in dealing with information overload on social media, it could cause risks of violating freedom of expression (if the project is not well conducted). We present some ethical and technical challenges involved in the entire pipeline of an AI project - from data collection to model evaluation - that hinder the large-scale use of hate speech detection algorithms. Finally, we argue that AI can assist with the detection of hate speech in social media, provided that the final judgment about the content has to be made through a process with human involvement.

Keywords:
Artificial Intelligence, Ethics, Online Harms, Hate Speech, Bias

Outline:

Author(s):

Prof. Dr. Diogo Cortiz
- Pontifícia Universidade Católica de São Paulo (PUC-SP)
- Brazilian Network Information Center (NIC.br)
- dcortiz@pucsp.br

Prof. Dr. Arkaitz Zubiaga
- Queen Mary University of London (QMUL)
- a.zubiaga@qmul.ac.uk

ISSN 1614-1687

1. Introduction

One of the most prominent challenges we have in the digital society is dealing with a contemporary phenomenon that has emerged from the intensive use of social networks: Online Harms. In the scope of this research, we can understand by Online Harms any illegal or unacceptable activity that can put a person, a group of people, and even democratic institutions at risk, often leading to consequences for the offline world. Some examples of the most common activities related to this movement are the spread of disinformation and fake news, hate speech, cyberbullying and foreign interference in internal affairs, which could impact not only a group of people but sometimes a nation as a whole.

In April 2019, the British Government released the document "Online Harms White Paper" aiming to propose a regulatory framework as a response to a possible 'free circulation' of harmful content on the Internet. In their words: "Given the prevalence of illegal and harmful content online, and the level of public concern about online harms, not just in the UK but worldwide, we believe that the digital economy urgently needs a new regulatory framework to improve our citizens' safety online".

To justify the initiative, the document lists a series of studies and evidence of threats in the digital universe. For example, a survey published by the NHS showing that one in five children (between 11-19 years) reported having experienced cyberbullying; a study produced by the University of Oxford's Computational Propaganda Project which discuss evidences of organized and structured manipulation on social media campaigns in 48 countries; a study by Reuters Institute study showing that 61% of people want the government to do more to separate what is real and what is fake; and an international survey showing that two thirds (64%) of female journalists had experienced online abuse - death or rape threats, sexist comments, cyberstalking, account impersonation, and obscene messages and 47% did not report the abuse they had received, and 38% admitted to self-censorship in the face of those abuses.

At first, this seems to be a valuable initiative to deal with a problem that still haunts a global society, but as the document also invited individuals and organizations to respond to questions, several academics and entities sent their contributions listing a series of criticisms of the proposal. Barker and Jurasz (2019) argue that the "Online Harms White Paper" is not fully adequate because the government wants to introduce legislation without understanding the full scope of the problem. The authors question that the proposal to create a government regulatory agency can be problematic and undermine the justice system that has attributions on the subject. They also argue that protecting freedom of expression and protecting the right to participation should be the criteria that guide social media regulation to ensure online equality. Harbinja *et al.* (2019), as a response from The British Irish Law Education and Technology Association (BILETA) also suggests that if a new regulator is needed (which the authors are skeptical), it must be independent to avoid being influenced by politics and industry. In the opinion of the authors is potentially undemocratic and does not support the standards of the rule of law of a democratic society.

The British government's initiative, although well-intended, is subject to criticism because it is a complex problem involving political, social, economic, and technological issues. And this challenge is not an exclusive concern of the United Kingdom but is present in other nations around the world. In Brazil, the National Congress is debating several bills to deal with "Online Harms" issues, especially the bill "PL 2630/2020", which became known as the "Fake News bill". This bill has been subject to several criticisms for proposing greater control over the data transfer and communication. For example, in the bill there is an article that requires all messaging services, such as WhatsApp, to store messages for three months as a mechanism to track possible harmful content. The bill has not yet been approved and is still under discussion, but it shows how the debate on these issues is a high priority.

The scope of this paper is not to address public policies, laws, or regulations discussions. We present an initial discussion on these topics to help understand that the problem of "Online Harms", especially Fake News and Hate Speech, is a complex and multifaceted challenge present in different sectors. The purpose of this paper is to discuss the ethical and technical challenges of using Artificial Intelligence (AI) to combat certain online threats. We agree that the debate on regulation is important, necessary, and must continue. However, we argue that there are still open fundamental questions about data processing. How to deal with an increasing volume of sensitive data that often depends on context and interpretation? What are the possible processes, techniques, and tools for identifying and tracking harmful content? Can AI assist in this process with

responsibility and effectiveness? Even though there is a regulation in force, the absence of answers to these technical questions will make it more difficult to comply with the established rules.

In a study for UNESCO, called "Countering online hate speech", Iginio *et al.* (2015) point out that social media platforms act reactively and take action only when users report harmful content, but that they could do much more in a proactive way. In their words:

> *"Social networking platforms could, however, take a more proactive approach. They have access to a tremendous amount of data that can be correlated, analysed, and combined with real life events that would allow more nuanced understanding of the dynamics characterizing hate speech online. Vast amount of data is already collected and correlated for marketing purposes" (Iginio, 2015).*

The academic and technical community has been studying computational approaches to assist in this process, especially on how to identify harmful content. We must consider that the quantity of data in digital media is bulky and the speed of the dynamics (i.e. posts and shares) of the content is extremely high. In this sense, we will discuss whether AI can help detect harmful content (in the midst of so much content) before it harms a person, a group of people, or an entire nation as well as the possible ethical implications of using AI in this subject.

As a case study we will present and discuss a project that we conducted for detecting hate speech using state of the art techniques in the area of AI (Machine Learning and Natural Language Processing). We will present the entire pipeline of an AI model training and for each stage, we will discuss the ethical and technical challenges of implementation. Although the case under discussion is about a model for detecting specifically hate speech, we argue that the process and its challenges are very similar to other projects that use machine learning to combat different kinds of online harms, such as Misinformation and Fake News.

2. Considerations about Hate speech

Social media platforms provide a new way to interact and produce content without prior supervision of an editorial process. Ellison and Boyd (2013) argued that the "implicit role of communication and information sharing has become the driving motivator for participation". If we analyze the first interactions that usually happen on new social media platforms, we usually find a conversation between friends, users sharing unpretentious information and producing creative content. But as the platform becomes popular as a free space, people start using it to discuss other topics, focusing on political and moral discussions for example, which can lead a small portion of users to use it as an environment for radicalization.

Suler (2004) presented the idea of "The Online Disinhibition Effect" as a proposal to understand why some people act in a more self-disclose and intense way when they are online than they would if they were in person. The author raised six main factors: dissociative anonymity, invisibility, asynchronicity, solipsistic introjection, dissociative imagination, and minimization of authority. "The Online Disinhibition Effect" can normalize extremist thinking, facilitate group polarization (Beadle, 2017), and stimulate antisocial behaviors.

It is a growing concern that is not reduced to just a technical concern but involves actors from different sectors. One difficulty is the lack of a precise definition of what hate speech is. The line between freedom of expression and hate speech is blurred. It is not the objective of this research to define with legal precision what hate speech is, so we chose to look for the concept of hate speech that the platforms are considering.

Facebook says that they do not allow hate speech because it creates an environment of intimidation and exclusion, which can, in some cases, lead to violence in the real world. Below is the definition of hate speech by Facebook.

> *"We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status.*

> *We define attack as violent or dehumanising speech, harmful stereotypes, statements of inferiority or calls for exclusion or segregation" (Facebook, 2020).*

Twitter does not provide a specific definition of hate speech, but they describe their stance against this kind of content.

> *"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories" (Twitter, 2020)*

YouTube also has a hate speech policy:

> *"Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender, Identity and Expression, Nationality, Race Immigration, Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin Veteran Status" (Youtube, 2020)*

As we can see in the above citations, although there is no strict legal definition of hate speech, there is indeed an effort by platforms to seek an understanding of hate speech to somehow combat them expressly through terms of use and code of conduct. In 2019, during a discourse at Georgetown University, Mark Zuckerberg argued that Facebook's mission is to support freedom of speech while in some way combat hate speech that can promote violence or dehumanize people or groups of people. However, he makes it clear that there is no consensus and this is one of the most difficult areas to deal with at the moment:

> *"I believe people should be able to use our services to discuss issues they feel strongly about — from religion and immigration to foreign policy and crime. You should even be able to be critical of groups without dehumanizing them. But even this isn't always straightforward to judge at scale, and it often leads to enforcement mistakes. Is someone re-posting a video of a racist attack because they're condemning it, or glorifying and encouraging people to copy it? Are they using normal slang, or using an innocent word in a new way to incite violence? Now multiply those linguistic challenges by more than 100 languages around the world."*

One point highlighted by Zuckerberg is the speed and volume of content posted on social networks (as we discussed earlier), which makes large-scale action for combat difficult, mainly if it is performed only by a group of human moderators. Perhaps this is one of the reasons why platforms prefer to act more reactively, which was criticized by Iginio *et al.* (2017). In this research, we argue that AI can be a support tool in the process of detecting potential hate speech. In the next section, we will describe our project to develop an AI system to help to identify hate speech and discuss the technical and ethical challenges.

3. Ethical and technical challenges

In a research initiative to assess the advantages and challenges of using AI to assist in the hate speech identification process, we have developed a project to develop and train an AI model to detect hate speech in a specific language (Portuguese of Brazil PT-BR). We used state of the art approach techniques in the area of Natural Language Processing to detect hate speech in text sentences.

The Natural Language Processing (NLP) is a sub-area of AI that works with the processing of a vast volume of data in natural language. Among various approaches and techniques, when working with text, it is possible to perform a series of tasks: sentiment analysis, named entity recognition, text generation, questions and answers, classification, among others. In our case, we developed and trained a model to classify sentences as ordinary text or hate speech text.

There are different techniques and models of text classification. We used the state-of-the-art approach in NLP. It is important to note that the area underwent a revolution in late 2018 when Google researchers presented a new Language Model called BERT - Bidirectional Encoder Representations from Transformers (Devlin *et al*, 2018). This model was made available by Google under an open-source license, which spurred the academic community and other companies to implement and evolve it. Different models based on the BERT architecture were implemented and showed great results in different fields of applications. For example, a group of researchers from Facebook AI and the University of Washington developed RoBERTa, a more robust model of the original BERT that achieved state-of-the-art results on different leaderboards with SQuAD and GLUE (Liu *et al.*, 2019). Since then, new language models have been developed inspired by the Transformers architecture of neural networks and BERT itself.

It is important to note that BERT is a language model, and a language model is just a type of probability distribution over sequences of words. Simply put, BERT has a representation of words in a given language (trained from Wikipedia and Books Corpus, in this case). For BERT to be useful in a context, we need to adapt it for a specific task, which is called fine-tuning. In our case, the project aimed to apply BERT to classify hate speech, so we needed to adapt (fine-tuning) the language model for classification with a set of labeled examples of hate speech.

To fine-tune a model, there must be a dataset with labeled examples. The collection of examples can be collected from the Web and then annotated (whether there is hate in a given sentence) by a group of experts. Once we have the labeled dataset organized and with a sufficient number of examples, we can adapt (fine-tune) the language model to learn how to classify hate speech (learning from the examples). The step after training is testing and validation. It is the most common process for training an AI model. In Figure 1, we present the pipeline of our project. In every stage, we listed the ethical and technical challenges that will be discussed later.

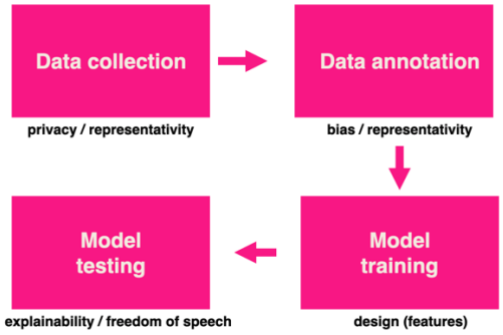

Figure 1 - Pipeline of AI Project

As we can see in Figure 1, the pipeline of our project is composed of four macro stages: Data collection, Data annotation, Model training, and Model testing. We listed the ethical and technical challenges (**in bold**) to be discussed in this paper. Although this pipeline is the one adopted in our project, it also represents a common method when developing AI and Machine Learning. In this perspective, we argue that those challenges are not restricted to our project and are present in the development of AI in different areas and sectors.

3.1. Dataset preparation: data collection and data annotation

Our project aims to train a model to identify hate speech in a sentence. In this case, we need to use an approach called "supervised learning", which requires an organized and labeled data set. Suppose you want to create a system that will classify an image as an apple or a banana. Initially, the system does not know any of the fruits. You could 'teach' the model what apples and bananas are by example, using a dataset in which every image has a label saying whether it is an apple or a banana. This is called a supervised dataset because there was a prior process to annotate the data. The same approach is useful for training a model to detect hate speech. We need a dataset with annotated sentences, which involves two tasks (as described in Figure 1): Data collection and Data annotation.

Data collection is the process of gathering a sample of sentences from social media platforms. This is usually done using APIs provided by the platforms themselves or using data scraping techniques. In this step, a sample of user-generated content (for example, posts, tweets, and others) is collected to be annotated later. In Figure 1 we listed some ethical challenges at this stage. The first one is representativeness. We must pay attention to the filters we use in the collection to ensure a more diverse and representative dataset. If we use a specific set of keywords, we can end up with data that represents the reality of only a group of users. The collection can also have a representativeness problem if the process involves the specific part of a network: suppose the researchers choose to collect data from a particular forum (some Chan, for example) or a precise subnet of a social network (for example, collecting data only from a specific group of users in Twitter), they will end up with a biased and unrepresentative data set. There is ongoing research looking into investigating appropriate methods to collect and measure dataset representativeness, but it is still an open question.

Privacy is another relevant challenge. The collection process involves using content posted by real users who do not necessarily want to be identified. One of the datasets that we used in our project was created by Fortuna *et al*. (2019). The authors were concerned with removing the user ID as an anonymization process. This is an important action, but it does not fully solve the problem because it is possible to re-identify users using different techniques. One can argue that the data is public and available on a social network, so it is not a privacy problem. However, a counterpoint argument is that the user posted something in social media only to express their thoughts, opinions, emotions, and did not agree to have their content used to train an intelligent agent. Privacy is a central theme on the global agenda with the number of regulations and legislation being discussed and created around the world.

Another activity of data preparation is the annotation process. Once the data has been collected, the next stage is to annotate all the sentences. In this case, experts on the topic are recruited to label a sentence as hate speech or ordinary speech. The same sentence is usually annotated by at least three distinct annotators as a strategy to bring greater plurality to the annotation process. Then some measure of agreement is applied between the annotators, such as Fleiss's Kappa (Fleiss, 1971). In the dataset released by Fortuna *et al*. (2019), for example, there was a low concordance value (K = 0.17) among annotators. The authors argued the result may be a consequence of having invited non-specialist annotators (Information Science student volunteers).

Sap *et al*. (2019) investigated how the insensitivity of the annotators to differences in dialects can lead to racial prejudice in models of automatic detection of hate speech, which can cause even greater damage to minority populations. This is a delicate subject that we would like to highlight in this work. The annotation process is fundamental in any AI project, especially when using the supervised learning approach. At a certain level, annotators influence the future behavior of an AI system. We should be concerned that the group of annotators must be plural to avoid problems of representativeness and bias. It is also possible to use priming techniques – as used by Sap *et al*. (2019) – to reduce bias in the annotation process. Still, this is an open question that the technical community, together with cognitive and social scientists, psychologists, linguists, among others, are trying to address.

3.2 Model Training: design

Once the dataset is ready, with sentences collected and annotated, the next step is training. Different models can be used for different tasks. The choice of the model and its design (choice of variables that will be used in training) can influence the learning and future behavior of the system. Obermeyer *et al.* (2019) described a health system that had a racial bias. At a given risk score to receive special treatment, black patients were considerably sicker than white patients. According to the authors, this bias appeared because the system

predicts health care costs rather than illness. The dataset used in training presented some type of bias, because unequal access to healthcare means that the health system usually spends less money caring for Black patients than for white patients.

But in addition to the bias in the data set, the 'design' of the model helped to amplify the harmful consequences towards black people. If the objective is to develop a fair system for allocating health services, then the system is expected to use data on the health condition of patients rather than their spending. Obermeyer *et al.* (2019) did an experiment to attempt to solve this problem. They maintained the same basis - sample, model and training process - but changed the labels: rather than future cost (as the original design), they created an index variable that combined health prediction with cost prediction. This new "design" of the project reduced the model's bias by 84%.

The aforementioned study is not directly related to the hate speech detection task, but it shows that the choices made when developing the model could influence its behavior. This situation is recurred in several areas, including the detection of hate speech detection. As seen previously, our project uses a language model (BERT) as a starting point. We then train (fine-tuning) it for the specific hate speech detection task using a particular data set. We saw in the last section that the data set could have bias, but it is important to mention that a language model may also present bias, which can contaminate the model's final training. For example, the GPT-3 language model shows traces of gender, race and religion bias (Brown *et al.*, 2020). In this sense, some studies are also emerging as an alternative to study and possibly mitigate bias not only in the data, but also in the algorithms (Mozafari *et al.*, 2020).

3.3 Model testing: explainability

After training, the next phase is testing and evaluating the model. Can a model with 96% accuracy be considered a good model? Technically we should check other metrics (such as F1-score) to evaluate the model in a more broadly manner, which takes into account the rate of false positive, false negative among others. Even if a model performs 96% on all performance metrics, does that mean it is a good model?

For a long time, computer scientists, programmers and engineers used performance metrics to assess whether a model was suitable for use in production. Perhaps this is one of the reasons that in the past few years some cases have emerged to show that AI systems could have unintended consequences. One of the most famous cases was the study published by the investigative journalism agency ProPublica that showed that a system used by some states in the USA to calculate criminal recidivism had a racial bias against blacks (Angwin, 2016).

We argue that performance metrics are not the only ones that should be considered to evaluate a model. We need to define new assessment metrics for AI models that take into account not only technical requirements, but also social aspects. However, it is not a simple thing, since the technical complexity itself imposes difficulties.

One of the biggest technical and ethical challenges is for AI to explain its decisions. This is such a sensitive issue that today there is an investigation field named "Explainable AI" to research and develop techniques that can bring some kind of interpretation to the decisions of a model. State-of-the-art architectures in the field of AI are so complex that it is difficult to understand why a system has made a particular decision.

When we started testing our model, we received the following inquiry from a policymaker who was interested in the topic: "What does the system consider to be hate speech?". Despite the performance results of our model (BERT-based model trained with the data set published by Fortuna *et al.* (2019)) surpass the benchmarks presented by Pari (2019), our challenge was to also be able to interpret the model. We discussed previously that there is not an established definition for hate speech, so our concern was to investigate what our model understands as hate speech.

We have implemented state-of-the-art techniques for interpreting AI models, such as CAPTUM (Kokhlikyan, 2020), as shown in the image on the next page:

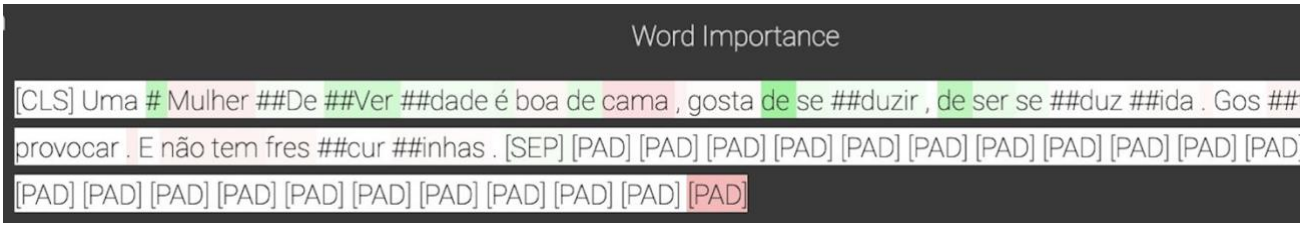

Figure 2 - Interpreting the Model

As can be seen in Figure 2, the techniques for interpreting AI models are still rudimentary. The technique presents in green the tokens (words or pieces of words) that most influence the model to assign a sentence to a class (for example, hate speech). In red are the tokens that most exclude the sentence of this class. The technique presents some clues, but it is not possible to have a deep understanding of the model's behavior. If we cannot explain the decisions, can we use the model in sensitive areas?

In the regulation of AI algorithms, a right to explanation is a right the citizen has to be given an explanation for the output of the system. In our project, we advocate that the automated hate speech detection process is just a tool to help the moderation team deal with the large volume of content posted daily. The system would only be responsible for placing a "mark" on the content so that it could be analyzed by a human moderator. In this case, the lack of a plausible explanation by the AI is less sensitive, because the final decision on whether or not to remove content is up to the moderator. But imagine a situation in which you have a bank credit denied by a fully automated process or someone goes to jail because the system decided it. If the AI fails to explain its decision, does it make sense for it to remain in operation? This is also an open question with technical and ethical consequences.

4. Conclusion

In this paper, we listed some ethical and technical challenges of Artificial Intelligence. Even though we discussed specifically a system to detect hate speech, the challenges are not limited to this scenario and are present in many sensitive areas – healthcare, bank credit approval, etc. We explored the complexity of an AI and how ethical discussions are present in different steps of the project – from data collection to model evaluation.

Based on that, we investigated whether an AI could be a technical tool to assist in the social challenge of hate speech. At first, according to all ethical and technical challenges present, we do not advocate that AI should be adopted in isolation as a strategy to solve the problem of hate speech, because it may jeopardize freedom of expression and human rights. We suggest that, alternately, the tool could cooperate with humans, by monitoring content and flagging likely hateful content which will subsequently be validated by human moderators. They will decide (and justify) whether content should ultimately be removed or not, but the AI can alleviate the task by sifting through the large volumes of content in social media by flagging up content to be checked.

The scope of this paper also served to provide an overview of the main challenges of AI for content moderation. We presented different questions that impact the use of AI. These questions mix ethical and technical issues, which requires greater integration between different areas of knowledge and cooperation across disciplines in order to be answered and tackled effectively.

5. References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. Machine Bias. ProPublica, 2016. https://doi.org/http://dx.doi.org/10.1108/17506200710779521

Barker, K. and Jurasz, O. Online Harms White Paper Consultation Response. Striling Law School & The Open University Law School, 2019.

Beadle, S. How does the Internet facilitate radicalization? London, England: War Studies Department, King's College, 2017.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. Language Models are Few-Shot Learners, 2020. http://arxiv.org/abs/2005.14165

Cortiz, D. O Design pode ajudar na construção de Inteligência Artificial humanística?, p. 14-22 . In: 17º Congresso Internacional de Ergonomia e Usabilidade de Interfaces Humano-Tecnologia e o 17 º Congresso Internacional de Ergonomia e Usabilidade de Interfaces e Interação Humano-Computador. São Paulo: Blucher, 2019.

ISSN 2318-6968, DOI 10.5151/ergodesign2019-1.02

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. http://arxiv.org/abs/1810.04805

Ellison, N. B., & Boyd, D. M. Sociality Through Social Network Sites (W. H. Dutton (ed.); Vol. 1). Oxford University Press, 2013. https://doi.org/10.1093/oxfordhb/9780199589074.013.0008

Facebook (2020). Community Standards;. Available from: https://www.facebook.com/communitystandards/objectionable content.

Fleiss, J. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378, 1971.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. A Hierarchically-Labeled Portuguese Hate Speech Dataset. Proceedings of the Third Workshop on Abusive Language Online, 94–104, 2019.https://doi.org/10.18653/v1/W19-3510

Harbinja, E., et al. "Online Harms White Paper: Consultation Response, BILETA Response to the UK Government Consultation'Online Harms White Paper', 2019.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. Countering online Hate Speech. UNESCO, 2015.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. Captum: A unified and generic model interpretability library for PyTorch, 2020. http://arxiv.org/abs/2009.07896

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. http://arxiv.org/abs/1907.11692

Mozafari, M., Farahbakhsh, R., & Crespi, N. Hate speech detection and racial bias mitigation in social media based on BERT model. PLOS ONE, 15(8), e0237861, 2020. https://doi.org/10.1371/journal.pone.0237861

MIT Technology Review. 10 Breakthrough Technologies, 2020. Available from: https://www.technologyreview.com/10-breakthrough-technologies/2020/

Nash, V. Revise and resubmit? Reviewing the 2019 Online Harms White Paper. Journal of Media Law, 11(1), 18–27, 2019. https://doi.org/10.1080/17577632.2019.1666475

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science, Vol 366 (6464), 447–453, 2019. https://doi.org/10.1126/science.aax2342

Pari, C; Nunes, G; Gomes, J. Avaliação de técnicas de word embedding na tarefa de detecção de discurso de ódio. In: Encontro Nacional De Inteligência Artificial E Computacional (ENIAC), 16 ,2019, Salvador. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, p. 1020-103, 2019. DOI: https://doi.org/10.5753/eniac.2019.9354.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678, 2019. https://doi.org/10.18653/v1/P19-1163

Suler, J. The Online Disinhibition Effect. CyberPsychology & Behavior, 7(3), 321–326, 2004. https://doi.org/10.1089/1094931041291295

Sun C., Qiu X., Xu Y., Huang X. How to Fine-Tune BERT for Text Classification?. In: Sun M., Huang X., Ji H., Liu Z., Liu Y. (eds) Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science, vol 11856. Springer, 2019. Cham. https://doi.org/10.1007/978-3-030-32381-3_16

YouTube. Hate speech policy. 2020. Available from: https://support.google.com/youtube/answer/2801939?hl=en

Twitter. Hateful conduct policy, 2020. Available from: https://help.twitter.com/en/rules-and-policies/hateful-conductpolicy

Zuckerberg, Mark. Mark Zuckerberg Stands for Voice and Free Expression, 2019. Available from: https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/