Emma Lagren

ARTIFICIAL INTELLIGENCE AS A TOOL IN SOCIAL MEDIA CONTENT MODERATION

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2023

ABSTRACT

Lagren, Emma

Artificial intelligence as a tool in social media content moderation
Jyväskylä: University of Jyväskylä, 2023, 27 p.
Information Systems, Bachelor's Thesis
Supervisor: Clements, Kati

On social media, users engage with each other through consuming and creating user-generated content, the amount of which has increased alongside the growth of the userbase. This presents challenges to social media companies as some uploaded content can be harmful to users and the appeal of the service. Platforms rely on filtering systems and manual labour to remove inadmissible content in the process of moderation. Advancements in the capabilities of artificial intelligence have made it possible to harness the technology for the purposes of moderation. This thesis explores the potential uses of AI in content moderation through features enabled by machine learning. The study was conducted as a literature review. It was found that artificial intelligence can detect instances of toxicity and hate speech as well as harmful images and multimedia content through natural language processing and computer vision technologies. The use of AI enables the scaling of moderation systems, the expedient evaluation of content and expanded moderation capabilities in different languages. Limitations in the implementation of automated moderation systems include inaccuracies, lack of contextual awareness, slow pace of adaptation, and concerns surrounding biases, transparency and freedom of expression.

Keywords: artificial intelligence, content moderation, social media, machine learning, automated moderation

TIIVISTELMÄ

Lagren, Emma
Tekoäly sosiaalisen median moderoinnin työkaluna
Jyväskylä: Jyväskylän yliopisto, 2023, 27 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja: Clements, Kati

Sosiaalisessa mediassa käyttäjät vuorovaikuttavat toistensa kanssa kuluttamalla ja tuottamalla käyttäjien luomaa sisältöä, jonka määrä lisääntynyt käyttäjäkunnan kasvun myötä. Tämä luo haasteita sosiaalisen median yrityksille, sillä osa julkaistusta sisällöstä voi olla haitallista käyttäjille ja palvelun houkuttelevuudelle. Alustat hyödyntävät suodattimia ja manuaalista moderointia poistaakseen sopimatonta sisältöä. Tekoälyominaisuuksien kehittyminen on mahdollistanut teknologian valjastamisen moderointiprosesseihin. Tässä tutkielmassa tarkastellaan tekoälyn käyttömahdollisuuksia sisällön moderoinnissa koneoppimisen mahdollistamien ominaisuuksien avulla. Tutkimus toteutettiin kirjallisuuskatsauksena. Tutkimuksessa havaittiin, että tekoäly pystyy tunnistamaan toksisuutta ja vihapuhetta sekä haitallisia kuvia ja multimediasisältöä luonnollisen kielen käsittelyn ja tietokonenäköteknologioiden avulla. Tekoälyn käyttö mahdollistaa moderointijärjestelmien skaalautumisen, sisällön nopean arvioimisen ja laajentaa eri kielien moderointimahdollisuuksia. Automaattisten moderointijärjestelmien käyttöönoton rajoituksia ovat epätarkkuudet, kontekstin tuntemuksen puute, hidas sopeutumisvauhti sekä huolet liittyen ennakkoasenteisiin, läpinäkyvyyteen ja sananvapauteen.

Asiasanat: tekoäly, moderointi, sosiaalinen media, koneoppiminen, automaattinen moderointi

CONTENTS

1 INTRODUCTION

Online platforms and social media are increasingly becoming integral to the everyday lives of people and provide benefits on a societal scale (Cambridge Consultants, 2019). These platforms act as channels through which users engage with communities and each other, sharing information and ideas in the form of user-generated content (Horne, 2023). The growing popularity of online spaces has led to larger volumes of user-generated content being published, some of which can be illegal or against the guidelines of the platform (Roberts, 2018). Content moderation is the process that monitors uploaded content, and through which unwanted content is identified and removed (Roberts, 2019). The ever-rising amount of user-generated content has created an incentive for online platforms to deploy automated solutions, and corporations are increasingly turning to artificial intelligence to expedite their content moderation processes (Nahmias & Perel, 2021).

The aim of this bachelor's thesis is to examine the possibilities and challenges of harnessing artificial intelligence (AI) as a tool in content moderation. This study can be utilized as an overview of current research and discussion on AI in the context of content moderation. Drawing from previous research, this thesis discusses how artificial intelligence is currently used in content moderation and what challenges its implementation is facing by answering the following research question:

- How can artificial intelligence be utilized in content moderation?

To answer this question, the thesis examines artificial intelligence, machine learning and deep learning capabilities, as well as how these technologies are being utilized in content moderation on different online platforms. In addition, the study discusses the benefits of AI moderation, the limitations facing its implementation as well as the issues it can create.

This thesis was conducted as a literature review. Chosen literature was sourced from searches on Google Scholar and JYKDOK. Keywords included 'artificial intelligence,' 'machine learning,' 'content moderation' and 'automated moderation' used together and separately. Through these keywords a great number of credible sources were found. Examples of moderation systems and deployed technologies were retrieved from the websites of social media entities. Chosen references included textbooks, articles from academic journals, news articles and publications such as transparency reports. Sources were screened for peer review, used references, recency, relevance and publication platform. The Publication Forum (JUFO) assessment for academic journals was checked during the review process. Literature discussing the use of artificial intelligence in content moderation was recent with publishing dates ranging from 2019 to 2023. Older sources were utilized when examining moderation and established technologies, although recent publications were preferred.

This study comprises five chapters: Introduction, Artificial intelligence, Content moderation, AI in content moderation and Conclusion and further research. The first content chapter 'Artificial intelligence' lays out the foundational understanding on what artificial intelligence is and what functionalities are enabled through machine learning. The second content chapter discusses content moderation on social media and the critique surrounding the practice. The third content chapter 'AI in content moderation' explores how machine learning technologies are used in moderation. The last chapter summarizes the findings of this study, discusses its limitations and presents topics for further research.

2 ARTIFICIAL INTELLIGENCE

The aim of this chapter is to provide a baseline understanding of core artificial intelligence technologies that will be utilized in later discussions. As the field of artificial intelligence is vast, the examination will focus on the concepts of AI that are relevant to the topic of the thesis.

2.1 Defining artificial intelligence

Despite significant developments in the capabilities of artificial intelligence, there is no single agreed upon definition of it due to the lack of tools necessary to evaluate, compare and classify AI systems (Bhatnagar et al., 2018). Defining artificial intelligence includes exploring the meaning of intelligence and how it manifests in AI systems (Wang, 2019), which can be difficult to gauge as comparably simple task-oriented systems are being compared to the full breadth of human intellect (Bhatnagar et al., 2018). Due to the complexity of defining intelligence a single definition may not be viable at this stage of AI research (Wang, 2019). However, a clear definition is not a prerequisite for scientific research and discussion, and the boundaries of research fields tend to form over time (Wang, 2019). AI is a broad topic of research, and it can be argued that a universal definition would reduce confusion when discussing across scientific fields, although no working definitions can be outright classified as true or false (Wang, 2019).

As presented by Wang (2019), AI can be seen as a computer system that in some sense emulates the human mind, using its structures, behaviours, capabilities, functions, or fundamental principles as guiding factors. Russell & Norvig (2022) argue that defining AI can be approached from two different dimensions: AI can strive for humanlike performance or try to reach optimal results, and that intelligence manifests itself as either thoughts and reasoning or as observable behaviour. However, it can be argued that an AI is fundamentally an abstraction of human intelligence and is therefore capable of rational decision-making, and that every AI includes a thinking and an acting aspect to be able to execute its functions (Wang, 2019). The standard model of artificial intelligence considers AI to be mainly driven by rationality, and as an intelligent agent it aims to make the best decision in each situation (Russell & Norvig, 2022). For the purposes of this thesis, artificial intelligence will be examined from the viewpoint of the standard model.

2.2 Machine learning

Learning is the process where an agent improves its performance based on observations made about the world, and when a computer engages in the process it's called machine learning (Russell & Norvig, 2022). Through machine learning an AI can make hypotheses about the world based on the data it observes, allowing the computer to adapt to changes in the environment (Russell & Norvig, 2022). The ability to learn enables AI to resolve possible scenarios independently, therefore removing the need for designers to anticipate and provide solutions for every outcome (Alpaydin, 2016). Machine learning creates the capacity for solving complex tasks that would be too demanding for traditional programming languages, such as facial recognition or language processing (Campesato, 2020).

2.2.1 Deep learning

Deep learning is a subset of machine learning that focuses on finding patterns in a set of data (Campesato, 2020). Deep learning networks are organized into at least two layers where each variable interacts with all other variables, meaning that the path from input to output contains multiple steps (Campesato, 2020; Russell & Norvig, 2022). Due to the architecture of deep learning networks consisting of multiple layers of nodes, deep learning relies heavily on large amounts of data and computing power in its operation, although once these requirements are met there is little need for manual interference (Alpaydin, 2016; Russell & Norvig, 2022). Deep learning is widely used in multidimensional abstract processes such as speech recognition and visual object recognition, exceeding human performance in such tasks (Russell & Norvig, 2022).

Networks trained using deep learning methods are referred to as neural networks due to early researchers aiming to emulate the neural structure of the human brain with computational circuits (Russell & Norvig, 2022). Convolutional neural networks (CNNs) are deep neural networks consisting of one or more convolutional layers that map the features of an image, feeding the output to a pooling layer that then classifies the image into a category (Campesato, 2020; Russell & Norvig, 2022). CNNs are especially successful in classification tasks, making them well-suited for processing images, audio, and natural language (Campesato, 2020; Russell & Norvig, 2022). Recurrent neural networks

(RNNs) operate in cycles, allowing them to use their own outputs as inputs in the next cycle, creating an internal memory (Russell & Norvig, 2022). RNNs are effective in tasks that process data sequentially, such as language modelling and machine translation (Russell & Norvig, 2022). Neural network learning algorithms can be trained in batches, where training is done all at once on a complete dataset that can be incrementally updated with minibatches (Alpaydin, 2016). With larger datasets online learning is preferred, where small updates are applied to connections between variables, allowing seamless adaptation without the need to stop for data collection and retraining (Alpaydin, 2016).

As highlighted by Campesato (2020), despite the impressive results deep learning has produced, challenges remain, such as biases and lack of explainability, among others. Machine learning agents can unintentionally pick up biases from algorithms or used training datasets, possibly perpetuating social and societal issues (Russell & Norvig, 2022). Lack of transparency and explainability can lead to users not trusting the systems they use, slowing down the adoption of AI (Russell & Norvig, 2022). Some initiatives have been launched to tackle these issues, such as EU regulations enforcing the explainability of decisions and industry frameworks defining best practices for ethical AI design (Russell & Norvig, 2022).

2.3 Natural language processing

Language is a complex form of presenting information, containing grammatical and semantic rules alongside subtle mechanics such as humour to be interpreted together in a general context to form meaning (Alpaydin, 2016). To engage with humans, learn, and advance scientific understanding of languages, computers use natural language processing (NLP) to interface with this aspect of communication (Russell & Norvig, 2022). NLP is currently of particular interest to the machine learning community, being utilized in chatbots, search engines, speech recognition and machine translation (Campesato, 2020).

In order to represent text, NLP uses a language model to determine the likelihood of a set words appearing next, thereby being able to suggest how a sequence might end (Russell & Norvig, 2022). The bag-of-words language model draws from a large, predefined vocabulary split into various categories, and in text generation words from a specific category are selected until an end-of-sentence indicator is reached (Alpaydin, 2016; Russell & Norvig, 2022). Categories are chosen based on the task, for instance spam filters consider the words “opportunity” and “offer” to be discriminatory (Alpaydin, 2016). However, the bag-of-words model assumes every word to be independent of others and cannot generate coherent sentences, therefore being primarily used in classification and text analysis (Alpaydin, 2016; Russell & Norvig, 2022). N-gram language models are probabilistic, examining interactions between words and determining the next word based on all previous words in a sentence (Russell & Norvig, 2022). These language models can perform diverse tasks such as language identification, spelling correction and sentiment analysis (Russell & Norvig, 2022).

Deep learning can produce impressive language models that are not as adversely affected by linguistic complexities compared to classical rule-based language models, yielding higher accuracy but slower output (Campesato, 2020; Russell & Norvig, 2022). NLP language models built on deep learning methods utilize CNNs and more commonly RNNs (Campesato, 2020). Sequence-to-sequence language models comprise of two RNNs, one for the source and another for the target, and are commonly used for machine translation (Russell & Norvig, 2022). Unlike CNN or RNN-based models that process inputs one word at a time, transformer models utilize a self-attention mechanism to process input sequences in their entirety and can therefore capture long-term and local context (Russell & Norvig, 2022). Similarly to other deep learning methods, NLP models require a large quantity of training data, access to which is provided by the internet and repository projects like Common Crawl (Russell & Norvig, 2022).

2.4 Computer vision

As presented by Russell & Norvig (2022), computer vision enables a computer to sense visual stimuli and act based on the information it observes. Computer vision utilizes extracted features, obtained by applying simple computations to an image, to interpret information from an image. Extracted features generally include edges, textures, optical flow, and segmentation. Alternatively, computer vision can be approached through object models that provide statements about the general properties of objects, and rendering models that describe the processes that produce the stimulus. Computer vision can be utilized in observing human behaviour, building tagging systems, creating, and modifying images and controlling movement with vision. (Russell & Norvig, 2022).

Modern systems approach image classification using appearance, which can be impacted by variables including lighting, occlusion, and deformation (Russell & Norvig, 2022). CNNs are especially effective at image classification, but still require large datasets to train for the aforementioned changes in object appearance (Russell & Norvig, 2022; Voulodimos et al., 2018). The ImageNet dataset contains over 14 million labelled and categorised datapoints which has played a significant role in the development of image classification systems, alongside the MNIST dataset with 70,000 images generally used as a warmup dataset (Russell & Norvig, 2022). Image classifiers predict what object is in an image, whereas object detectors find multiple objects in an image, reporting their class and location (Russell & Norvig, 2022). Being unaffected by transformations such as translation, rotation, and scale, CNNs are effective at object detection as well as other computer vision tasks including facial recognition, action and activity recognition and human pose estimation (Voulodimos et al., 2018).

3 CONTENT MODERATION

This chapter explores content moderation as a concept, with a particular emphasis on the practices of large social media platforms and commercial content moderation, although other models are also presented. Common critique surrounding content moderation is examined, highlighting the social, societal, and ethical problems it can create.

3.1 Defining content moderation

Content moderation is the organised practice of monitoring, evaluating, approving and removing user-generated content (UGC) uploaded to websites, social media platforms and other online outlets (De Gregorio, 2020; Roberts, 2019). Screening UGC can take place before the content is distributed or after it has been uploaded, triggering reactively as a result of receiving complaints from users, administrators or external parties (De Gregorio, 2020; Roberts, 2019). UGC acts as the currency which engages users to a platform, allowing them to create and consume content in the form of text, images, video, or a combination of multiple types of media (Roberts, 2018). Through social media people are put into direct contact with one another, enabling them to share information and form relationships and communities, but despite the positive use cases UGC can also contain harmful and dangerous material including false information, violent or adult content and hate speech among others (Gillespie, 2018; Horne, 2023). As social media has in some sense become an extension of the real world, evidence suggests that it can contribute to the spread of 'offline harms' (Blackburn & Zannettou, 2022; Horne, 2023). The aim of moderation is to enforce positive communications online while minimizing aggression and anti-social behaviour (De Gregorio, 2020).

As corporate entities operating on the internet, online platforms are required to engage in the practice of content moderation to protect users or groups from one another, remove unwanted content, and to present the service in an attractive light to shareholders and the public (Gillespie, 2018). Through content moderation platforms also ensure they are compliant with local legislation that governs their operations by not allowing illegal material to be hosted on their services (Roberts, 2019). However, despite presiding over the legal framework social media platforms operate in, governments have been dethroned as the primary regulators of speech by the private companies that own the largest social networks in the world (Wilson & Land, 2021). Platforms enforce local law alongside their own policies in the form of community guidelines, which respond to contemporary fears and concerns around sexual content, obscenity, and violence, often prohibiting otherwise lawful speech when drawing from the operator's beliefs and norms (Gillespie, 2018; Keller & Leerssen, 2020). Conversely, community guidelines can also be used to serve business processes by for example only allowing pictures of properties on a real estate listing site (Keller & Leerssen, 2020). The importance of enforcing a platform's own rules increases as policymakers and the public demand the removal of offensive, but not always illegal, content from online outlets (Keller & Leerssen, 2020). Therefore, platforms do not shape public discourse themselves, but as hosts of the discourse, they shape the shape of it through moderation (Gillespie, 2018).

3.2 Content moderation models

3.2.1 Commercial content moderation

Commercial content moderation, as coined by Roberts (2019), is content moderation practiced by commercial websites, social media properties, and media properties that rely on UGC in their production cycle. Large volumes of UGC ensure that users return to a continually updated feed and regularly engage with the platform, yet high traffic poses complex challenges for moderation (Roberts, 2018; Roberts, 2019). Pre-moderating UGC happens before it is published, providing a high level of control at the cost of a substantial reduction to the amount of UGC uploaded (Veglis, 2014). Pre-moderation by human actors removes the satisfaction experienced by the uploader as they are left waiting for their submission to be cleared and is disrupting in spaces where interaction happens in real time (Veglis, 2014). Automation has significantly expedited pre-moderation by reducing the need for human intervention using filters, which apply pre-defined rules to determine if a submission is admissible, and techniques like hash matching where the digital fingerprints of files are compared (Llansó, 2020; Veglis, 2014). In automated pre-moderation the identifiers of unwanted content must be applied to the filters in advance, for example the list of blocked keywords and hashes (Llansó, 2020).

Nearly all platforms opt for a post-moderation approach, where contributions are immediately public and inappropriate content is removed after the fact (Gillespie, 2018; Roberts, 2019). While algorithmic systems have been developed and deployed to take on the enforcement of platform rules and community guidelines, a large share of the work is done manually by human content moderators (Roberts, 2018). This process is set off when a user flags UGC as offensive or otherwise inappropriate (Crawford & Gillespie, 2016). Flagging is the mechanism that enables platforms to moderate large amounts of content manually, as detecting harmful content can be outsourced to the entire userbase instead of dedicated moderators (Crawford & Gillespie, 2016). Flags prompt a review by a professional moderator, who applies internal rules to determine if the contribution will be removed or kept on the site (Roberts, 2018; Roberts, 2019). The advantage of post-moderation is that uploading UGC and other user activities can happen instantaneously, although the lack of screening can lead to users being exposed to harmful material which can remain unnoticed for an indefinite period of time (Gillespie, 2018; Veglis, 2014).

3.2.2 Community-driven moderation

Community moderation is a form of content moderation where the process is primarily handled by the users of the platform instead of by the company that owns the service, used mainly on community-centric platforms including Wikipedia, Twitch and Reddit (Seering, 2020). Volunteer community moderators act as community leaders and can create growth in user numbers as well as contribution quality through thoughtful moderation (Seering, 2020). Some platforms include flagging features which engage users in the process of moderation and can bring issues into moderators' attention, although the system can also be abused through false reports or users reporting an admissible contribution en masse in an attempt to have it removed (Seering, 2020). In user moderation each user can rate an individual contribution up or down by one point, and posts meeting or exceeding a threshold may be hidden (Veglis, 2014). Despite relying on users to manage day-to-day moderation on a smaller scale and rarely interfering in how communities are run, large platforms have been known to exercise their authority and remove entire communities if their presence is deemed harmful, regardless of user protest (Rozenshtein, 2023). It is worth noting that some platforms, including imageboards 4chan and 8chan as well as alternative social media sites like Gab, choose to engage in little to no moderation of the UGC present on their sites (Zeng & Schäfer, 2021). When no official moderation scheme exists, users engage in spontaneous moderation where comments are posted about other comments (Veglis, 2014).

3.2.3 Moderating decentralised social media

Decentralised or distributed social media, commonly called the Fediverse, is a network of independent servers called instances that interact with each other through federation, where instances connect in a peer-to-peer fashion (Anaobi et al., 2023; Rozenshtein, 2023). Each of these instances is comparable to a social media application, connected with a shared protocol, allowing users to interact with their peers regardless of the instance they are signed up to (Anaobi et al., 2023; Rozenshtein, 2023). Therefore, the Fediverse is physically decentralised but logically interconnected (Anaobi et al., 2023).

Due to the shared protocol being decentralised, each instance can choose what content is allowed to flow through the network and set their own content moderation standards, for example blocking formats like image or video, or other instances (Rozenshtein, 2023). Instances are policed by independent administrators who determine what content moderation policies are effective on the instance (Anaobi et al., 2023). However, only a small percentage of instances share the load across multiple moderators, and evidence suggests it can lead to overwhelming workloads and the employment of less sophisticated policy strategies (Anaobi et al., 2023). According to Rozenshtein (2023), since no instance can control another instance and there is no central authority, a user or piece of content cannot be entirely banned from the network as long as it exists on an instance. However, it is noted that decentralised social media can self-police by refusing to interact with a shunned instance, cutting it off from most of the network and leaving it isolated. Further, the architecture of the Fediverse allows users to choose the instance that suits them, enabling users to have greater power in the event of dissatisfaction compared to centralised platforms. Even with the lack of a central authority, the Fediverse exists on servers located in nations that can interfere if illegal content is being hosted on an instance, imposing some level of content moderation (Rozenshtein, 2023).

3.3 Critique of content moderation

3.3.1 Scale, complexity, and the worker

User-generated content is being uploaded in unfathomable quantities (Roberts, 2018). In the first half of 2022 social media service X, formerly known as Twitter, required users to remove over six million pieces of content and took enforcement action on over five million accounts for violating community guidelines (X, 2023). From April to September of 2023, Facebook and Instagram removed over 90 million and 150 million pieces of content respectively in EEA countries alone (Meta, 2023a; Meta, 2023b). Out of Instagram’s 150 million inadmissible submissions, 76 million were removed by a team of 15,000 human content reviewers (Meta, 2023b). Manually reviewing content is expensive and difficult to scale as upload volumes increase (Cambridge Consultants, 2019). Due to the amount of UGC uploaded, the time employees are allocated to decide the admissibility of a contribution is typically measured in seconds, resulting in thousands of decisions per day (Gillespie, 2018; Roberts, 2018).

Evaluation of UGC happens at the intersection of the nature of the content, its intent, its meaning and its unintended consequences (Roberts, 2019). Decisions to intervene must balance offence and importance, follow ethical obligations and respect the complexities of political, cultural and social discussion in an environment where value systems compete (Gillespie, 2018). Content is evaluated against the guidelines of the platform and the larger environmental context to determine its admissibility (Roberts, 2019). When screening UGC, moderators of Facebook follow a 10,000-word document of moderation guidelines divided into categories of harmful behaviour, sensitive content and legal violations (Wilson & Land, 2021). As UGC is complex and therefore difficult to assess, human moderators must possess knowledge about a platform's audience and culture to make decisions accurately (Roberts, 2019).

Given the large amount of multifaceted UGC moderators evaluate in a short duration, errors and disparities in accuracy are known to happen (Wilson & Land, 2021). Gillespie (2018) and Roberts (2018) highlight the case of a Vietnam War-era photo being removed from Facebook for containing child nudity, disregarding its importance as a piece of anti-war media. Tobin et al. (2017) found that human moderators apply content evaluation guidelines inconsistently. The researchers asked Facebook to explain its moderating decisions regarding 49 contested contributions and the company stated its moderators made mistakes in 22 of them. They also noted that offensive language may survive the moderation process if it is not violent or derogatory enough to meet the platform's definition of hate speech.

Moderating UGC is often outsourced and undertaken in low-wage and low-status environments by temporary contract workers (Roberts, 2019). Acknowledgements of moderation practices and the human workforce behind them often omit details of location, working conditions and the background of employees (Roberts, 2018). Content moderators often work covered by non-disclosure agreements, maintaining the secrecy of company policies (Roberts, 2018; Roberts, 2019). Human moderators are exposed to material ranging from benign to disturbing, which can lead to the development of mental disorders including post-traumatic stress disorder (Gillespie, 2018). The wages of moderators can vary from one to four dollars per hour to being compensated per submission reviewed (Gillespie, 2018), incentivising moderators to prioritize speed over accuracy.

3.3.2 Inequalities in global moderation

Developing Asian and African countries are seen by social media companies as time and resource-intensive yet unprofitable markets (De Gregorio & Stremlau, 2023). Facebook devotes 87% of its global budget on classifying misinformation to the United States despite making up 10% of the platform's daily users (Frenkel & Alba, 2021). A critical factor in moderation inequality is language diversity, as insufficient moderation resources are allocated to minor languages and hate speech detection systems struggle with the range of dialects and vernacular languages (De Gregorio & Stremlau, 2023). Evidence suggests that hateful expressions in non-Western languages and English content in non-Western contexts are more likely to be overlooked by filters and moderation processes (Udupa et al., 2023). Due to the lack of viable moderation tools, African countries have relied on blocking content, direct censorship and internet shutdowns to combat online speech and offline harms (De Gregorio & Stremlau, 2023).

In the most extreme cases insufficient moderation can enable atrocities. Wilson & Land (2021) highlight the ethnic cleansing in Myanmar as "the most harmful use of social media by a government thus far" (p. 1034). They claim that posts on Facebook containing narratives designed to stoke fear about the Rohingya minority led to a wave of violence where at least 10,000 Rohingya civilians were killed. According to the researchers these posts were uploaded by high-ranking members of Myanmar's military government and Buddhist nationalists. Presented numbers claim that 780,000 Rohingya have been forcibly displaced to neighbouring countries as of September of 2018. The spread of hate speech has been attributed to inadequate moderation tools and lack of moderators that speak the local language (De Gregorio & Stremlau, 2023). Facebook has since accepted its central role in the genocide and employed Burmese-speaking moderators (De Gregorio & Stremlau, 2023; Wilson & Land, 2021).

3.3.3 Transparency and censorship

Platforms have little incentive to share their moderation practices and outcomes with the public, as constantly evolving systems are burdensome to document and admitting errors can be used to the detriment of the platform (Keller & Leerssen, 2020). The rules of social media platforms are publicly available in the form of community guidelines, but the way those policies are translated into action is not explained or open to scrutiny (Wilson & Land, 2021). This presents issues for the freedom of expression, as the flow of information on social media platforms is organised to serve business processes and maximise profit instead of ensuring integrity (De Gregorio, 2020). Without transparency and explainability of decision-making processes, debates about platforms become speculation and drafting legislation regarding their operation is rendered ineffective (De Gregorio, 2020; Keller & Leerssen, 2020).

Langvardt (2017) presents that previously the law of free expression was subject to law itself as states were the primary entities engaged in censorship. It is stated that currently the mantle is held by a small number of technology oligarchs that exercise state-like power while being politically unaccountable. In 2017 the CEO of Cloudflare withdrew security support from a white nationalist site, expressing concern over having the power to make the decision (Gillespie et al., 2020; Keller & Leerssen, 2020). Langvardt (2017) highlights a resemblance between the Golden Shield censorship system used in China and the moderation practices of America's private sector, noting that their guiding principles differ merely in the cultural dimension. It is presented that entrusting online speech rights to corporations alone is a dangerous outcome and meaningful legislation is needed to guarantee the right of freedom of expression (Langvardt, 2017)

In the United States content moderation is regulated under CDA 230 which exempts companies from legal liability for their users' actions while permitting platforms to police the service (Gillespie et al., 2020). Requirements for transparency across platforms are currently not implemented and companies self-regulate, though some insight is provided through transparency reports (De Gregorio & Stremlau, 2023; Gillespie et al., 2020). The European Union has developed frameworks and regulations to combat disinformation and increase platform transparency in the form of the EU Code of Practice on Disinformation and more recently the EU Digital Services Act (Galantino, 2023). However, Galantino (2023) argues that the Digital Services Act provides insufficient protection of European freedom of expression due to the uncertainty of regulating private entities whose moderation practices affect the flow of information. Social media processes are shaped predominantly in the US and EU, which presents challenges to the global south in how to ensure justice and equal treatment for their citizens (Gillespie et al., 2020). Many initiatives concerning the issue of harmful online content in Africa have been seen as proxies for censoring speech, which can be justified in some cases (De Gregorio & Stremlau, 2023).

4 AI IN CONTENT MODERATION

In this chapter the potential of using artificial intelligence as a content moderation tool is explored through current technologies and deployed moderation systems. The benefits and limitations of their functionalities are examined alongside concerns surrounding the use of AI moderation. Lastly the prospects of AI content moderation are presented.

4.1 Overview of current technologies

Artificial intelligence used in content moderation is also called algorithmic or automated moderation (Gorwa et al., 2020). AI moderation is commonly deployed as an adjacent system to other moderation methods, for instance assisting human moderators in flagging content (Gongane et al., 2022). The shift to achieve fully automated moderation is not a purely technical transition, as AI offers solutions to scale a moderation system to manage the growing amounts of UGC (Elkin-Koren, 2020). It is worth noting that moderation can be offered as a service to online platforms, meaning it is not always necessary to develop in-house moderation systems (Cambridge Consultants, 2019).

As discussed previously, training an AI model requires a large, labelled dataset. In the context of social media content moderation these datasets usually consist of extracted and annotated UGC, although other datasets such as ImageNet can be utilized (Gongane et al., 2022). X is the preferred platform for collecting data for the creation of fake news and hate speech datasets (Gongane et al., 2022). It should be noted that social media entities can utilize uploaded UGC and existing infrastructure to gather their own datasets for training internal systems. Transparency reports from Facebook and Instagram (Meta, 2023a; Meta, 2023b) outline a feedback loop between human review and AI technology, where decisions made by the moderation team are labelled and used as training data for automated systems. However, some machine learning technologies require little to no training data, such as the few-shot learning method used in moderating Facebook (Meta, 2021). The Few-Shot Learner aims to act on new types of harmful content within weeks, and it can be deployed to reinforce existing AI moderation systems (Meta, 2021).

Natural language processing methods can be used to detect detrimental text-based UGC on social media platforms (Gongane et al., 2022). Text classifiers can analyse the features in text to classify it into a category, for example hate speech or toxicity (Andročec, 2020; Duarte et al., 2017). The primary machine learning methods used in toxic comment classification are CNNs and other neural network architectures including transformers which have recently shown superior performance in NLP-related tasks (Andročec, 2020). Sentiment analysis can be performed by utilizing the bag-of-words language model, which classifies words and their variants with a score to determine if the text is positive or negative (Cambridge Consultants, 2019). N-grams used together with other AI techniques have proven especially proficient in NLP tasks as they can be trained to flag misspelt or alphanumeric words (Cambridge Consultants, 2019). Zhu et al. (2021) developed a framework of unsupervised algorithms that analyse words in their sentence-level context and can detect euphemisms. It was discovered that the model can identify euphemisms with greater accuracy than the current state-of-the-art as well as discover previously unknown hidden meanings. The researchers believe that the framework can be utilized in moderation as an efficient alternative to manual evaluation.

Social media platforms host multimedia content including images, GIFs and videos, and the detection of harmful contributions in these formats often combine multiple machine learning models (Gongane et al., 2022). CNNs enable image analysis through object detection and scene understanding which can detect objects of interest from an image (Cambridge Consultants, 2019). Karabulut (2020) presents a CNN-based automatic moderation system that analyses and classifies uploaded images and automatically obfuscates restricted content. It was found that the classification network correctly labelled 90,3% of entries and the obfuscation algorithm on average covered 68% of areas classified as explicit (Karabulut, 2020). CNNs paired with NLP technologies that detect and categorise associated text allow AI to identify harmful memes, instances of hate speech and fake news articles (Cambridge Consultants, 2019; Gongane et al., 2022). Similar techniques are used by Meta in the Rosetta machine learning system deployed on Facebook and Instagram (Sivakumar et al., 2018). Rosetta uses a CNN to recognize and transcribe text detected in a specific region, and paired with an image classification model the system can understand the context of the text and the image together (Sivakumar et al., 2018). RNNs can be utilized to analyse content consisting of multiple images or frames that are relative to each other, such as videos and GIFs (Cambridge Consultants, 2019).

4.2 Benefits, limitations and issues

Utilizing AI moderation solutions allows social media platforms to expand further even as the amount of UGC rises since automated systems can be scaled to manage the volume of uploaded material (Gillespie, 2020). Evidence suggests that AI-based tools can detect detrimental content on social media with high accuracy and speed (Gongane et al., 2022). Facebook has stated that 99% of removed terrorist content is flagged by AI systems instead of users, reducing the amount of exposure to extremist content (Nahmias & Perel, 2021). NLP techniques can improve moderation capabilities by providing moderators with translation tools for assessing UGC in different languages and expediating the evaluation process through the inclusion of context (Cambridge Consultants, 2019). Utilizing automated systems can be seen as an ethical alternative to the manual labour of human moderators, as reducing the amount of manually reviewed content limits employees' exposure to disturbing material (Gillespie, 2020; Gongane et al., 2022). AI-powered techniques such as nudging, chatbots or automatic keyword extraction can be used to promote socially positive online engagement by discouraging users from posting harmful content (Cambridge Consultants, 2019).

Despite the advancements in technology, AI faces limitations in its moderation capabilities. Unclear and inconsistent definitions of prohibited content restrict the ability of automated systems to target specific material (Duarte et al., 2017). UGC is constantly evolving and as most AI approaches rely on training the system before deployment, AI must be retrained to combat new types of content (Cambridge Consultants, 2019). NLP tools have been observed to have limited utility outside the context they were trained in, meaning they cannot perform reliably when deployed for other uses (Duarte et al., 2017). Effective content moderation requires contextual awareness which many AI systems are incapable of, although improvements in this sector can be observed (Cambridge Consultants, 2019; Udupa et al., 2021). Automated systems can address issues related to 'super spreaders,' bot activities and trending devices such as hashtags, but are incapable of moderating extreme speech circulating in closed communities that are built on existing social ties, such as WhatsApp groups (Udupa et al., 2023).

With the introduction of AI to content moderation, issues regarding the use of automated systems have been raised. Algorithmic tools and training datasets are developed by humans and can therefore contain biases, mirroring social and societal inequalities (De Gregorio & Stremlau, 2023). Included biases can lead to moderation processes disproportionately targeting minority groups or those with minority views (Duarte et al., 2017). Machine learning tools are evaluated based on their accuracy, yet accuracy often refers to how closely the machine's interpretation of content matches a human's interpretation of the same content (Llansó, 2020). The 2020 Ranking Digital Rights assessment found that no evaluated company fully disclosed how the content users consume is curated, ranked or recommended (Udupa et al., 2021). Machine learning-based systems can be difficult to scrutinize if they are not designed with transparency and explainability as guiding principles (Llansó, 2020). Deciphering takedowns and flagging decisions of AI moderation systems is challenging and the criteria on which those decisions were based on remain unknown (Gorwa et al., 2020). Full automation would eliminate dissenting actors from within the content moderation system as machines cannot provide insight into its operation by supplying information to outside parties (Roberts, 2018). Concerns have been raised around the use of AI technologies as instruments of state surveillance and manipulating online discourse (Udupa et al., 2023).

4.3 Future of AI content moderation

Cambridge Consultants (2019) argue that effective automated moderation is not possible and in the foreseeable future human moderators continue to be essential for reviewing contextual and nuanced content. However, social media platforms have reached the scale where automating moderation processes is the only viable option (Gillespie, 2020). Areas of interest for researching automated moderation solutions as identified by Gongane et al. (2022) include context-aware moderation, explainable AI (XAI) and moderating multimedia content. Alongside the development of new technologies, moderation system design could focus on creating tools to support human moderators (Gillespie, 2020).

Frameworks have been proposed to address the issues related to the use of AI in content moderation and to explore the benefits of human-AI collaboration. Elkin-Koren (2020) suggests introducing an adversarial AI that reflects public interest to dispute the decisions made by automated moderation systems. The system aims to combat the authority platforms hold over public discourse by ensuring a level of freedom is present in online social spaces (Elkin-Koren, 2020). Lai et al. (2022) developed a collaborative system between humans and AI called conditional delegation where rules are created to evaluate which regions of an AI model are trustworthy, and the deployed AI model only acts in instances that belong to the trustworthy regions. It was found that the AI model paired with conditional delegation is more effective than the model working alone, and the practice is promising when applied to the context of content moderation. The researchers claim that conditional delegation allows moderators to decide when to trust or distrust AI. However, it is emphasised that further research is required to realise the impact of conditional delegation in moderation.

5 CONCLUSION AND FURTHER RESEARCH

In this thesis the use cases and limitations surrounding the utilization of artificial intelligence in content moderation were examined, as AI has been identified as a solution to online platforms' challenges of moderating growing amounts of user-generated content. Artificial intelligence was defined as a computer system that aims for the best outcome in each situation. Relevant subfields of AI were explored, and it was found that through machine learning AI can make predictions based on previous observations, and deep learning enables AI to process natural language and detect visual stimuli. Moderation was found to be the practice of evaluating user-generated content and applying guidelines to determine if it is appropriate for the platform. Content moderation systems use filters to pre-moderate uploads and rely heavily on human labour to manage post-moderation. Concerns surrounding moderation include the lack of transparency, misuse of moderation to silence speech, global inequalities in moderation capacity and the working conditions of human moderators.

The final chapter covering the use of AI in content moderation synthesized previous discussion to answer the research question:

- How can artificial intelligence be utilized in content moderation?

It was found that natural language processing tools can detect toxicity, hate speech and harmful euphemisms. Computer vision capabilities enable AI systems to analyse images, GIFs and videos to find unwanted pieces of content. Moderating multimedia combines multiple machine learning technologies to detect instances of fake news or harmful memes. The benefits of using AI moderation include the scalability of systems, accurate and expedient content evaluation and expanded moderation capabilities in different languages. Identified limitations relate to inaccuracies in targeting harmful material, the lack of contextual awareness, the inability to quickly adapt to evolving content and the narrow use cases of systems targeting specific content. The issues surrounding the use of AI in moderation in part mirror those of moderation itself, as concerns include the lack of transparency and limited explainability of decisions as well as the fear of censorship. Alongside these issues, concerns have been raised about biases in algorithms and datasets that can disproportionately affect minority groups.

Throughout the course of this study, limitations in its scope were identified. Most of the automated content moderation systems in use today are owned by private corporations that lack incentive to publish the performance records of their moderation systems. Therefore, the information needed to make observations is not always available. Some knowledge could be gleaned from transparency reports, but the machine learning technologies used in these systems, the harmful content they target and the accuracy of their decisions was not publicly disclosed. Studies of experimental technologies tested in the context of content moderation was able to parse this information gap to some degree. However, how these technologies would be implemented into a content moderation system and to what extent they would retain their performance capabilities remains unclear.

As the use of artificial intelligence in social media content moderation is expanding, further research into the field is required. Despite being popular forms of media, tools for moderating video material and livestreams are rudimentary and rely on detecting known content through matching technologies. The moderation capabilities of different machine learning technologies such as hybrid models are worth investigating, as well as refining existing natural language processing and computer vision tools. The development of context-aware moderation systems could reduce the number of false positives.

REFERENCES

Alpaydin, E. (2016). *Machine Learning: The New AI*. The MIT Press.

Anaobi, I. H., Raman, A., Castro, I., Bin Zia, H., Ibosiola, D., Tyson, G. (2023). Will Admins Cope? Decentralized Moderation in the Fediverse. *WWW'23: Proceedings of the ACM Web Conference 2023*, 3109-3120. https://doi.org/10.1145/3543507.3583487

Andročec, D. (2020). Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica, 12*(2), 205-216. https://doi.org/10.2478/ausi-2020-0012

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., Halina, M., Loe, B. S., hÉigeartaigh, S. Ó, Martínez-Plumed, F., Price, H., Shevlin, H., Weller, A., Winfield, A., Hernández-Orallo, J. (2018). Mapping Intelligence: Requirements and Possibilities. In: *Philosophy and Theory of Artificial Intelligence 2017*. Springer, Cham. https://doi.org/10.1007/978-3-319-96448-5_13

Blackburn, J. & Zannettou, S. (2022). Effect of Social Media Networking on Real-World Events. *IEEE Internet Computing, 26*(2), 5-6.

Cambridge Consultants. (2019). *Use of AI in online content moderation*. Ofcom. https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation

Campesato, O. (2020). *Artificial Intelligence, Machine Learning, and Deep Learning*. Mercury Learning & Information.

Crawford, K. & Gillespie, T. (2016). What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society, 18*(3). https://doi.org/10.1177/1461444814543163

De Gregorio, G. & Stremlau, N. (2023). Inequalities and content moderation. *Global Policy, 00*, 1-10. https://doi-org.ezproxy.jyu.fi/10.1111/1758-5899.13243

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review, 36*. https://doi.org/10.1016/j.clsr.2019.105374

Duarte, N., Llansó, E. & Loup, A. (2017). Mixed messages? The limits of automated social media content analysis. Center for Democracy & Technology.

Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720932296

Frenkel, S. & Alba, D. (2021, October 23). In India, Facebook Grapples With an Amplified Version of Its Problems. *The New York Times*. https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html

Galantino, S. (2023). How Will the EU Digital Services Act Affect the Regulation of Disinformation? *SCRIPTed, 20*(1). DOI: 10.2966/scrip.200123.89

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720943234

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A. & Myers West, S. (2020). Expanding the debate about content moderation : scholarly research agendas for the coming policy debates. *Internet Policy Review, 9*(4). DOI: 10.14763/2020.4.1512

Gongane, V. U., Munot, M. V. & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining, 12*(1). https://doi.org/10.1007/s13278-022-00951-3

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1). https://doi.org/10.1177/2053951719897945

Horne, B. D. (2023). Is Automated Content Moderation Going to Solve Our Misinformation Problems? *Information Matters, 3*(1). https://dx.doi.org/10.2139/ssrn.4359981

Karabulut, D. (2020). *Neural Networks Based Automatic Content Moderation on Social Media* [Master's Thesis, Tartu University]. Tartu Ülikool. http://hdl.handle.net/10062/72115

Keller, D. & Leerssen, P. (2020). Facts and Where to Find Them: Empirical Research on Internet Platforms and Content moderation. In: *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press. https://doi.org/10.1017/9781108890960

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y. & Tan, C. (2022). Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In: *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 54,* 1-18. https://doi.org/10.1145/3491102.3501999

Langvardt, K. (2017). Regulating Online Content Moderation. *Georgetown Law Journal, 106*(5). https://dx.doi.org/10.2139/ssrn.3024739

Llansó, E. J. (2020). No amount of «AI» in content moderation will solve filtering's prior-restraint problem. *Big Data & Society, 7*(1). https://doi.org/10.1177/2053951720920686

Meta. (2021, December 8). *Harmful content can evolve quickly. Our new AI system adapts to tackle it.* https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/

Meta. (2023a, October 27). *Transparency Report for Facebook.* https://transparency.fb.com/sr/dsa-transparency-report-oct2023-facebook/

Meta. (2023b, October 27). *Transparency Report for Instagram.* https://transparency.fb.com/sr/dsa-transparency-report-oct2023-instagram/

Nahmias, Y. & Perel, M. (2021). The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. *Harvard Journal on Legislation, 58*(1).

Roberts, S. T. (2018). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday, 23*(3). https://doi.org/10.5210/fm.v23i3.8283

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Rozenshtein, A. Z. (2023). Moderating the Fediverse: Content Moderation on Distributed Social Media. *Journal of Free Speech Law, 3*(1), 217-235.

Russell, S. J. & Norvig, P. (2022). *Artificial intelligence: A Modern Approach* (Fourth edition). Pearson.

Seering, J. (2020). Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1-28. https://doi.org/10.1145/3415178

Sivakumar, V., Gordo, A. & Paluri, M. (2018, September 11). Rosetta: Understanding text in images and videos with machine learning. *Engineering at Meta*. https://engineering.fb.com/2018/09/11/ai-research/rosetta-understanding-text-in-images-and-videos-with-machine-learning/

Tobin, A., Varner, M. & Angwin, J. (2017, December 28). Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up. *ProPublica*. https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes

Udupa, S., Hickok, E., Maronikolakis, A., Schuetze, H., Csuka, L., Wisiorek, A. & Nann, L. (2021). Artificial Intelligence, Extreme Speech and the Challenges of Online Content Moderation. AI4Dignity Project. https://doi.org/10.5282/ubm/epub.76087

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society, 10*(1). https://doi.org/10.1177/20539517231172424

Veglis, A. (2014). Moderation Techniques for Social Media Content. In: *Social Computing and Social Media. SCSM 2014. Lecture Notes in Computer Science, vol 8531*. Springer, Cham. https://doi.org/10.1007/978-3-319-07632-4_13

Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience, 2018*. https://doi.org/10.1155/2018/7068349

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence, 10*(2), 1-37. DOI: 10.2478/jagi-2019-0002

Wilson, R. A. & Land, M. (2021). Hate Speech on Social Media: Content Moderation in Context. *Connecticut Law Review, 52*(3), 1029-1076.

X. (2023, April 25). *An update on Twitter Transparency Reporting*. https://blog.twitter.com/en_us/topics/company/2023/an-update-on-twitter-transparency-reporting

Zeng, J. & Schäfer, M. S. (2021). Conceptualizing «Dark Platforms». Covid-19-Related Conspiracy Theories on 8kun and Gab. *Digital Journalism, 9*(9), 1321-1343. https://doi.org/10.1080/21670811.2021.1938165

Zhu, W., Gong, H., Bansal, R., Weinberg, Z., Christin, N., Fanti, G. & Bhat, S. (2021). Self-Supervised Euphemism Detection and Identification for Content Moderation. *2021 IEEE Symposium on Security and Privacy,* 229-246. https://doi.org/10.48550/arXiv.2103.16808