AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "CONTENT MODERATION"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Georgi Kostadinov** (Imagga) **Chris Georgiev** (Imagga) **Pavel Andreev** (Imagga) **Ralitsa Golemanova** (Imagga)

This report is part of the deliverable D2.3 - *"AI technologies and applications in media: State of Play, Foresight, and Research Directions"* of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Content moderation

Current status

During the last few years, ***content moderation*** powered by Artificial Intelligence has grown exponentially. It has developed in ways that were unimaginable just a decade ago, breaking concepts and widening the horizons of what's possible.

Automated content moderation has been fuelled by ever-evolving machine-learning algorithms that constantly improve in accuracy and speed. Just 10 years ago, image recognition was only able to classify and detect basic objects and shapes. Now, thanks to the advancements of deep learning, image recognition algorithms for instant detection of all types of inappropriate visual content are a reality.

Automated (also referred to as semi-automated) content moderation thus offers important new capabilities for businesses of different venues that need effective screening of digital content. The AI moderation platforms address a number of key challenges that online platforms and companies face, including:

- Huge amounts of user-generated content need to go online immediately, but still have to be monitored for appropriateness, safety, and legality. This can make it difficult for online platforms to grow and scale internationally if they don't have an effective way to ***screen all postings*** — textual, visual, and even live streaming. Without moderation, these businesses risk great reputational harm, along with other negative consequences.
- Content moderation has to happen ***in real-time***, which is especially difficult for live streaming and video that are becoming the most popular content formats. The complexity of screening visuals, texts and moving images at the same time is tremendous.
- ***User safety*** and especially the ***protection of vulnerable groups*** is becoming a priority in legislation that covers digital platforms. This means that in many places across the globe, online businesses are required by law to have solid Trust and Safety programs and protection mechanisms based on content moderation. This is necessary not only to ensure the upholding of their internal principles and guidelines, but to safeguard consumers.
- The ***stress and harmful effects on human moderators*** from exposure to shocking, violent and disturbing content is significant. Digital businesses aim to minimise these negative consequences and to protect their moderating teams from the worst content.
- Digital platforms have to be able not only to scale in terms of countries and amounts of content that goes live, but also to adapt to ***quickly changing circumstances and norms for content appropriateness***.
- Public ***manipulation, political propaganda, disinformation*** through fake news, and the rise of deepfakes are disturbing yet prevailing new phenomena online. Both official authorities and online platforms need an effective way to fight them, and machine learning algorithms are the key to that.

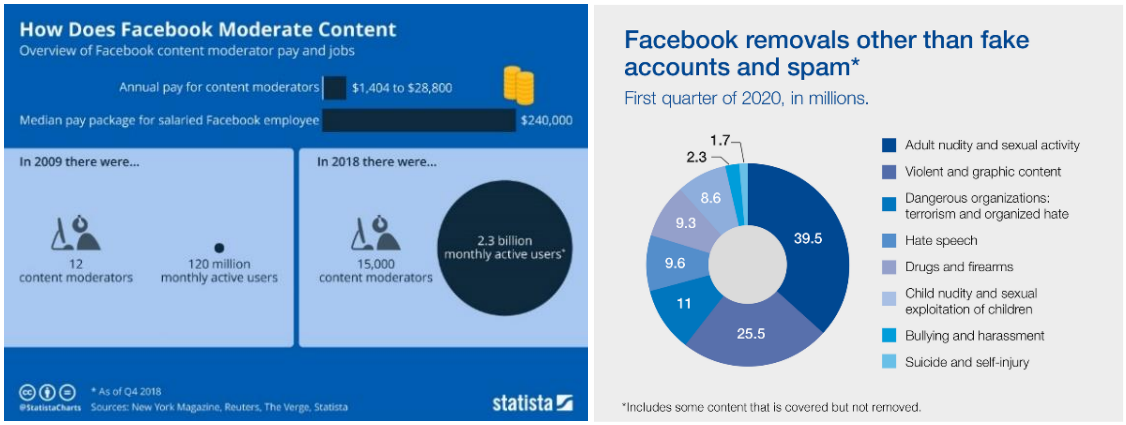

Figure 1: Some interesting moderation statistics: a) Facebook content moderation between 2009 and 2018[1], and b) Facebook content removals per category in 2020[2].

Research challenges

While holding great potential and already showing impressive results, there are challenges that AI-powered content moderation is facing.

One of the major issues with which automated content moderation is struggling ***is recognising context***. Machine learning algorithms can find it difficult to differentiate between subtle cultural and social trends and phenomena. For example, if the algorithm is set to remove all nudity, this is what it would do - even if the nudity is related to art or important news pieces. A prominent example was the case from 2016 when Facebook removed the photo of the iconic Vietnamese 'napalm girl' who is naked[3].

A growing body of ***national and international regulations*** are already affecting online platforms that allow the sharing of user-generated content and communication between users. The regulations are steadily pushing for effective content moderation in order to protect people and ensure a fair and safe environment for all.

The European Union is gradually moving forward with its Digital Services Act[4], and so are individual countries like the United Kingdom, France, Germany, and the United States with their national legislation. These new regulations include provisions to provide adequate protection of users against inappropriate and harmful content. The rationale of these legal requirements

[1] Image source: Statista, How Does Facebook Moderate Content (2019) https://www.statista.com/chart/17302/facebook-content-moderator/

[2] Image source: P. Barrett, Who Moderates the Social Media Giants? (2020): https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf (p. 10)

[3] S. Levin, J. Carrie Wong and L. Harding, "Facebook backs down from 'napalm girl' censorship and reinstates photo", The Guardian (2016): https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo

[4] European Commission, Proposal for a Regulation on the European Parliament and of the Council on a Single Market for Digital Services (DSA Act) and amending Directive 2000/31/EC, 15 December 2020, COM(2020) 825 final. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN

seems understandable, especially considering the infamous cases of crimes and murders that have been live streamed on major platforms like Facebook, YouTube, and Periscope[5].

On the other end, this might create more fears of ***moderation bias and censorship***, resulting in limiting the freedom of expression, and the right of the user to know why their content was removed. What is clear is that we need to better understand the context, the political situations in different regions of the world as well as cultural and religious particularities and properly transfer that knowledge to any AI algorithm that will deal with automated content moderation.

Another important challenge that AI platforms need to overcome is ***multilingual and multicultural moderation***. While they are getting better at it and are surely improving how content moderation in different languages is performed, there are still obstacles in the way. The process is not only about acknowledging the direct meaning of words and phrases, but their ***social and cultural connotations*** that may make them offensive and inappropriate. In this respect, the more feedback machine learning algorithms receive, the better they can become at spotting the nuances in content — which is definitely not a mission impossible, but simply a gradual process that takes processing large amounts of data.

Live streaming and live video are another interesting challenge for AI-based content moderation platforms. They generate such a substantial amount of data per second that manual moderation is simply an impossible task. Moreover, applying AI on each frame of the live broadcast generates high platform costs. A fast and accurate AI needs to be developed to overcome these hurdles and reach efficient and cost-effective moderation of live streaming and videos.

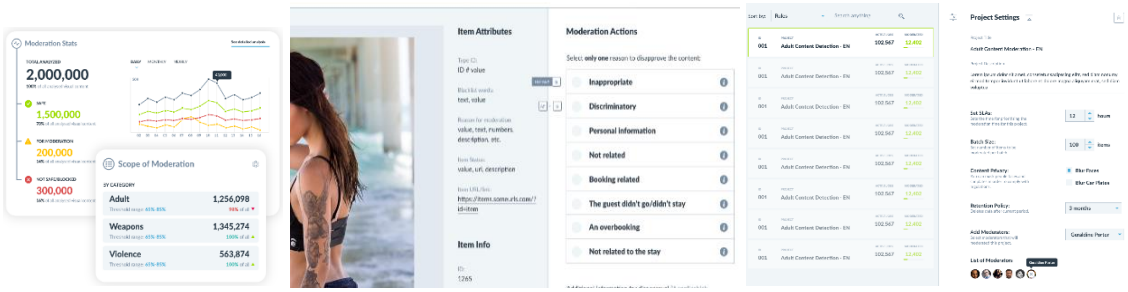

Figure 2: A content moderation tool by Imagga[6].

Societal and media industry drivers

Vignette 1: Automatic moderation of content harmful for vulnerable communities

Merry is a manager in an online newspaper with a solid and long-standing reputation for monitoring content that might be harmful for vulnerable communities. Her job is to ensure every piece of information that gets published through the media's channels meets company standards for content that might be shown to vulnerable groups. She needs to check not only

[5] O. Solon, "Why a rising number of criminals are using Facebook Live to film their acts", The Guardian (2017): https://www.theguardian.com/technology/2017/jan/27/rising-numbers-of-criminals-are-using-facebook-to-document-their-crimes

[6] Image source: https://imagga.com/content-moderation-platform (provided by Imagga)

facts and textual references, but also visuals. In a way, Merry is a modern-day fighter against discrimination – one of the plagues of today's digital world.

But doing all of this on her own – even with her fact-checking and moderation teams – is a monstrous task. The process simply takes enormous amounts of time and effort. That's why Merry needs a viable and scalable solution for checking and preventing the spread of inappropriate visuals.

This is especially important in some parts of Europe and other countries of the world, where protection of basic human rights is not fully enforced and problematic content can easily slip in. Ordinary citizens, and sometimes politicians share problematic content for their own gain – and it's very difficult to sift through what's acceptable and what's not. Merry has to figure out how true the lead is and whether to publish it.

With the help of an AI-powered content moderation platform, Merry can screen various materials around the topic for authenticity. She can catch textual references, as well as photos that have already been posted, for example.

Vignette 2: Real-time moderation of live-streaming content

Daniel is a content editing manager at an online news outlet. He's in charge of guaranteeing that all published content complies with the standards of the media and the legal framework. His most challenging task is to ensure the compliance of live streams. Catching inaccurate and harmful video content in real time is a tough nut to crack. Without technical support, it is a burdensome task to monitor live streaming content as it occurs.

This is where a content moderation platform based on machine learning algorithms can kick in. It processes all live video streams, checking them for inappropriate verbal and visual elements. If there are such, the platform can immediately signal this to the editors.

Daniel can test the capabilities of the AI platform in practice during an important live streaming with a local politician at a public rally. The situation is uncontrollable, as it's a place full of people where anyone can appear and take over the stage. With the help of the moderation platform, Daniel can have the live stream screened for problematic content throughout the whole event.

Future trends for the media sector

The role of content moderation in the media sector cannot be overestimated — in fact, it's crucial for its wholesome development on a couple of levels:

- Content moderation algorithms are the key to fighting the *online disinformation*. They can spot inaccuracies in textual data and fake or old visuals and videos. These capabilities can be a game changer for the media sector that direly needs adequate fact-checking in the ocean of information that gets published daily.
- Protecting vulnerable groups, under-represented communities and women as well as fighting terrorism will be of utter importance in the future and media companies need to find ways to better understand such content and adopt or develop tools to address it.

- On the basis of the massive volumes of content that machine learning algorithms process, they can also make *predictions about the types of content that needs to be moderated*. This can be of huge importance for combating harmful tendencies in user-generated content sharing.
- Content moderation, if done in-depth, can help in detecting the intent of disinformation in order to differentiate between positive (for example, to keep state secrets) and negative (to influence opinions and harm society).
- With the fast adoption of Metaverse, a new set of moderation requirements are emerging as this new medium merges reality with virtual worlds. Fake and authentic will have a totally new meaning and content moderation platforms of tomorrow need to adapt to this new trend of virtualisation of almost everything.

Goals for next 10 or 20 years

The long-term vision for AI-powered content moderation is a truly ambitious one.

First and foremost, content moderation would need to leave the semi-automatic status it currently has. To be fully useful, scalable, and powerful, it should be ***more autonomous***. The vision is that AI-powered content moderation platforms would become a monitor that is always on and oversees any and all content that goes online. It would screen for all types of abnormalities to ensure protection of users from harmful and illegal content, and a safe online environment without fake news — thus taking care of everything from violence and nudity to propaganda, radicalisation and disinformation, and all that's in between.

The second big goal for content moderation's evolution is ***self-learning*** — which is already in motion, but can reach new heights. With the data that is being fed in the moderation platform in real time, the machine learning algorithm becomes better and better. It expands its knowledge base with practical examples and input from moderators. With time, this is how the AI can become more independent from humans in terms of feedback loop. In the foreseeable future, this can reach a point where the moderation platform becomes an autonomous machine that identifies and filters content accurately and effectively with no human input.

A third long-term goal for content moderation is the creation of ***instant and efficient on-device moderation***. Nowadays, moderation is executed on the server end, only after a piece of content has already gone live. This means that harmful content can sneak in for a moment and be accidentally shown to users. In the near future, moderation would be possible on the user device itself. This would happen before the content has gone live. This advancement would enable the prevention of illegal and disturbing content appearing on the device level, thus ensuring full protection for end users.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu