Proceedings of the 2nd International Conference on Social Psychology and Humanity Studies
DOI: 10.54254/2753-7048/36/20240443

Harnessing the Power of Artificial Intelligence to Combat Abuse, Bias, and Discrimination in Social Media Algorithms

Xuwen Lin[1,a,*]

[1]*Master of Science in Communication, Northwestern University, Evanston, IL, USA*
a. xuwenlin2021@u.northwestern.edu
**corresponding author*

Abstract: In every corner of social media, abuse, bias, and discrimination are prevalent phenomena. Misinformation spreading, cyberbullying, and even the swaying of public opinion and user viewpoints are examples of ABD (Abuse, Bias, and Discrimination). It is also important to remember that Artificial Intelligence (AI) systems have the potential to function biasedly, which could result in unfair user interactions and information distribution. This essay aims to highlight the potential hazards associated with social media use in the modern world. As an essential instrument for spreading information, it should be transparent, secure, equitable, and inclusive. This paper can increase the diversity of information flow, lessen the effects of abuse, bias, and discrimination, improve the social media environment, and increase the value of social media by skillfully utilizing AI technology. This paper's primary focus is on AI's numerous approaches and tactics to identify and lessen bias, abuse, and discrimination on social media. It emphasizes the significance of data diversity and quality and how to refine algorithms to make them fairer and more transparent. It also delves into stricter and more precise regulations and user education for social media, ensuring, with the help of AI algorithms, it becomes a safer, more inclusive, and fairer space.

Keywords: Social Media, Artificial Intelligence, Abuse, Bias, Discriminatio

1. Introduction

Social media is a vast data source for training and fine-tuning artificial intelligence models. These AI models, such as recommendation algorithms and content filters, significantly shape the content users encounter on these platforms. Unfortunately, the data they rely on can be tainted by the very issues of ABD (Abuse, Bias, and Discrimination) that social media platforms face. The deeply ingrained stereotypes in our society find their way into the data, propagating biases and discriminatory content through Artificial Intelligence (AI) systems [1]. In this context, the intersection of social media and artificial intelligence models magnifies the existing problems of ABD, making it crucial to address these issues at both the data and AI model levels. This integration highlights the interconnectedness of social media, artificial intelligence models, and the persistent challenges of abuse, bias, and discrimination in the digital realm.

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/36/20240443

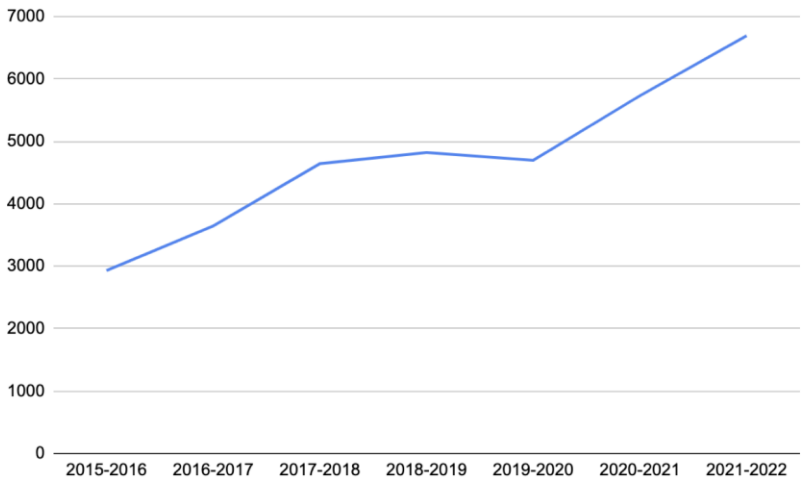

Figure 1: The number of articles, books, and journals that contained the terms 'Artificial Intelligence' and 'Social media' in their titles from 2015 to 2022.

Figure 1 shows the increasing trend in articles, books, and journals that featured the terms 'Artificial Intelligence' and 'Social Media' in their titles from 2015 to 2022. This data suggests a growing interest among people in the relationship between AI and social media. As the digital age ushers in, people's lives are intricately intertwined with social media, influencing billions of thoughts, behaviors, and decisions. Beyond personal connections, social media has become pivotal for businesses, governments, and NGOs [1]. However, with this influence comes a dark side: the challenges of Abuse, Bias, and Discrimination (ABD) in online platforms.

Large AI models heavily depend on their training data, which can be prone to malicious manipulation, including toxic inputs, biases, ideological attacks, opinion manipulation, misinformation, and privacy breaches. AI algorithms on social media platforms are often criticized for their inherent biases and discrimination. Such algorithms can exacerbate societal inequalities, restrict information flow diversity, diminish information dissemination quality, and even compromise societal fairness, intensifying societal divisions [1]. For instance, as shown in Figure 2 and Figure 3, a YouTube video titled "Joe Biden falls on stage at US Air Force Academy ceremony," a total of 2,334 comments were extracted using Netlytic. Of these comments, 87 contained negative emotional words such as "bad," "creepy," "evil," "embarrassed," "scary," "ashamed," and "awful." Additionally, 61 comments included swear words or derogatory terms, with examples being "embarrassing," "disgrace," "fool," "loser," "coward," "fk," and "selfish." The presence of negative emotional words and derogatory terms suggests an unfavorable bias towards the event or Joe Biden himself.

DOI: 10.54254/2753-7048/36/20240443

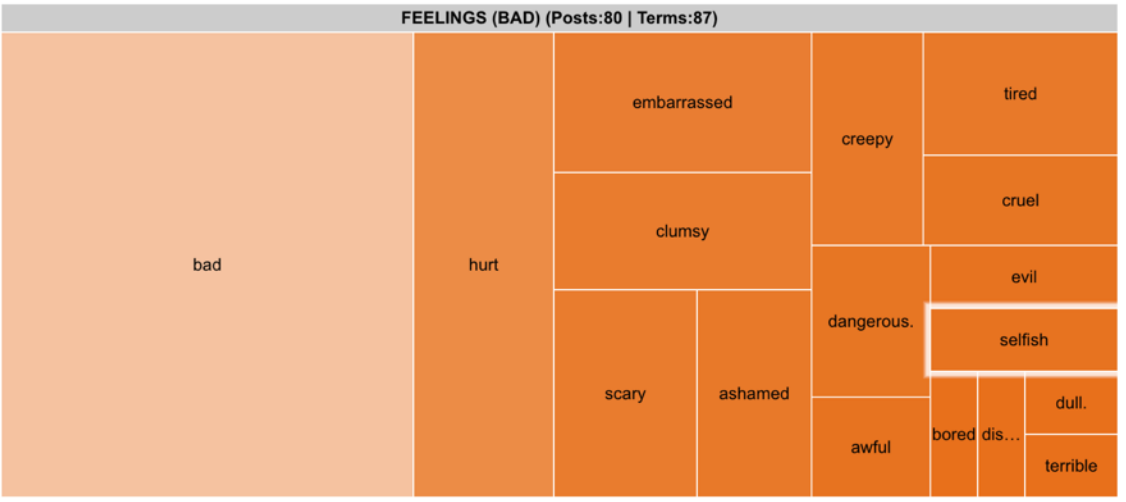

Figure 2: Netlytic Text Analysis of YouTube video "Joe Biden falls on stage at US air force academy ceremony." – Feelings (bad).

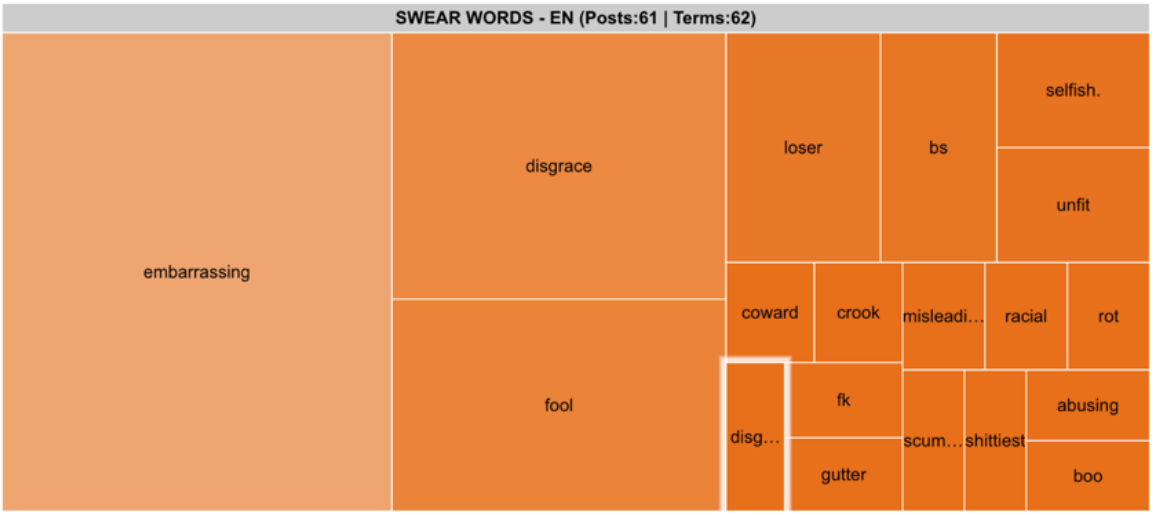

Figure 3: Netlytic Text Analysis of YouTube video "Joe Biden falls on stage at US air force academy ceremony." – Swear Words.

There is a wide array of such circumstances, misinformation is one of them. The New York Times received over 4,000 instances of election misinformation from readers and analyzed various types of deceptive content shortly before the midterm elections. These included "hoax floods" of false claims surrounding major news events, poorly labeled campaign ads on social media platforms, Russian manipulation campaigns on Reddit, voter suppression attempts through misleading information, deceptive claims about candidates including manipulated images and videos, misleading text messages from political campaigns, and the use of "attack pages" by both Democrats and Republicans to mock and criticize opponents on social media [2].

Furthermore, cyberbullying, especially among young adolescents. A study led by NIH-funded researchers has found that cyberbullying, especially when experienced by young adolescents, is closely linked to a higher chance of suicidal thoughts and attempts. The research, based on data from over 10,000 adolescents, showed that the likelihood of reporting suicidal thoughts or attempts was more than four times higher in people who were the targets of cyberbullying. This connection remained significant even after adjusting for other factors like family conflict and racial

DOI: 10.54254/2753-7048/36/20240443

discrimination. Importantly, cyberbullying's impact was independent of in-person bullying. The study highlights the need for routine screening for cyberbullying as a risk factor for suicide and underscores the negative consequences of virtual space bullying, urging parents, educators, and healthcare providers to be vigilant regarding this issue [3].

Another concern is manipulation. According to the Oxford Internet Institute's 2020 media manipulation survey, coordinated social media manipulation tactics are now a major threat to democracy worldwide. In the 81 surveyed countries, evidence of such campaigns was found in all of them, marking a 15% increase from 2019. Governments, political parties, and public relations firms are now generating misinformation on an industrial scale, with over 93% of countries using disinformation as part of political communication. The report highlights a surge in social media manipulation, with state actors employing private sector "cyber troops" to amplify their messages, often using citizen influencers to spread manipulated content. Social media platforms like Facebook and Twitter have spent significant sums combating these efforts. In many nations, computational propaganda is commonplace and has an impact on free speech and public opinion. It uses data-driven targeting, abusive tactics, and phony accounts. [4].

Additionally, algorithmic biases, according to a UC Davis study, users subjected to a "loop effect" of biased video recommendations on platforms such as YouTube and TikTok may become more radicalized politically because of the algorithms' ability to trap them. The researchers discovered that both left- and right-wing consumers may get radicalized and polarized due to these algorithms. To quantify bias, videos were scored based on their alignment with left or right-wing accounts on Twitter. While the initial tests showed more radicalization on the right, the issue became roughly equivalent on both sides over time [5].

Discriminatory Practice is another concern. Facebook's owner, Meta, and the Department of Justice (DOJ) have settled claims of discriminatory advertising practices contravening federal housing law. According to the DOJ, Meta used algorithms to target consumers with housing ads based on attributes, including sex, race, and national origin, that are protected under the Fair Housing Act. The settlement, still pending court approval, would force Meta to stop utilizing a unique audience tool for housing advertisements and develop a new system that complies with government regulations and is subject to routine compliance audits by third parties [6].

When AI learns from historical data, it perpetuates these ingrained biases, reflecting and often amplifying society's problematic aspects. This could involve promoting dominant cultural narratives or misrepresenting marginalized groups [7,8]. This paper aims to highlight the significance of social media in the modern world, discuss some of its possible hazards, and investigate methods for dealing with problems like abuse, bias, and discrimination on the platform. Through case studies and examples of cutting-edge technology, the paper also demonstrates the potential of AI in media enhancement. Key findings are also summarized, and future research and development directions are discussed.

The paper commences with an introductory section highlighting the significance of social media and its potential risks. In this introduction, the practical challenges of abuse, prejudice, and discrimination within the realm of social media are examined alongside real-life instances that exemplify these issues. Moreover, the introductory segment delves into the detrimental consequences of these problems on society and individuals. The following section delves into possible solutions for these issues. It expounds upon the vital role of legal frameworks, informed users, and cutting-edge artificial intelligence technologies in mitigating social media discrimination, abuse, and bias. This section also offers illustrative examples of best practices and successful interventions. The third section employs case studies and contemporary technological illustrations to showcase how artificial intelligence can be harnessed to elevate the media realm. Here, specific examples of AI tools designed to combat abuse, bias, and discrimination on social media are provided, alongside discussions on how

DOI: 10.54254/2753-7048/36/20240443

AI can effectively address these pressing concerns. In the paper's culmination, a comprehensive summary is presented, along with an exploration of potential avenues for future research and advancement in this domain. The conclusion underscores the imperative need for ongoing research and development in this field, highlighting the substantial potential of artificial intelligence (AI) technologies to enhance the landscape of social media profoundly.

2. The Landscape of Modern Social Media Bias

It is essential to comprehend the complex processes by which these problems develop, materialize, and spread as individuals traverse this enormous virtual environment [8]. People can get a thorough understanding of the complex issue of bias and discrimination on social media platforms by looking at the stages of production, usage, and dissemination [9].

2.1. Production

Algorithmic Bias: Algorithms are frequently used by social media platforms to control what content users see. If these algorithms are designed or trained with biased data, they can perpetuate or amplify existing biases [1].

Content Creation: Content creators who harbor biases or discriminatory views may produce content that reflects those beliefs. This can lead to the propagation of stereotypes, hate speech, or misinformation [7].

Data Collection and Usage: The data upon which platforms operate might come from non-representative samples, leading to unintentional biases in recommendations and predictions. Platforms might create profiles of users based on their demographics or behavior, which can sometimes lead to reinforcing stereotypes [1].

2.2. Usage

User Behavior: Users may employ social media platforms to harass, belittle, or bully individuals, often targeting specific racial, gender, or other identifiable groups [8]. Users also may unintentionally or intentionally surround themselves with like-minded individuals, leading to reinforcement of existing beliefs and biases.

Platform Policies: Platforms might inconsistently enforce their terms of service, leading to some hate speech or discriminatory content remaining online while other benign content gets flagged or removed [9].

Interaction: Users might engage more with content that aligns with their pre-existing beliefs, causing algorithms to show them more of such content, leading to a vicious circle [9].

2.3. Dissemination

Spread of Biased Content: Due to the nature of social media, hateful or discriminatory messages can spread rapidly, causing harm to targeted groups or individuals [10].

Misinformation: Misinformation that aligns with certain biases can spread, leading to further entrenchment of incorrect beliefs or stereotypes [11].

Countermeasures and Awareness: Social media can also be a tool for spreading awareness about bias and discrimination, with activists using it to counteract harmful narratives [10].

3. Harnessing AI for Good

Artificial intelligence (AI) emerges as a first line of defense in an era marked by an abundance of information, where misinformation can spread at an alarming rate. It gives robots the ability to

DOI: 10.54254/2753-7048/36/20240443

understand and interpret human language, providing a powerful defense against false information. Moreover, AI-powered text analysis improves fact-checking tools and makes predictive analysis possible, which helps prevent potential harm from inaccurate or misleading information in advance [9]. Like this, artificial intelligence (AI) tools for image and video analysis are invaluable in the fight against manipulated content, including deepfakes, in the visual media domain because they can detect inconsistencies and analyze metadata. The adoption of diverse training data also takes center stage in this discussion, emphasizing the need for inclusive and unbiased datasets to ensure that AI models reflect a comprehensive understanding and do not perpetuate skewed perspectives [9].

3.1. Text Analysis

Natural Language Processing (NLP) is an advanced domain of artificial intelligence that seeks to make machines understand and interpret human language [1,8]. With the digital age comes an information overload. Misinformation can spread rapidly, making it challenging for individuals and organizations to discern truth from falsehood. Here is where AI-driven text analysis shines:

- First Line of Defense: NLP can be a sentinel against misinformation [12]. By analyzing the semantic and syntactic structures of statements, AI can determine the likelihood of a piece of news being false or misleading.
- Supercharging Fact-checking Platforms: Organizations like Snopes and PolitiFact have been at the forefront of verifying claims made by public figures, news outlets, and viral internet posts. Integrating AI into their operations can exponentially increase their efficiency, allowing for real-time analysis of viral news, monitoring trends, and flagging inconsistencies [12].
- Predictive Analysis: Beyond fact-checking, AI can predict which pieces of information are likely to go viral. This can enable proactive measures, ensuring that potential misinformation is addressed before it can cause substantial harm [12].

3.2. Image and Video Analysis

In the visually driven digital age, images and videos carry immense power. However, the rise of deepfakes and manipulated visual content threatens to undermine public trust in visual media.

- Deepfake Detection: Deepfakes employ deep learning techniques to generate realistic-looking video footage of real people saying or doing things they never did. These can be potentially damaging, especially when used to misrepresent public figures. AI models, when trained on vast datasets of real vs. manipulated footage, can detect subtle inconsistencies that the human eye might miss [8].
- Metadata Analysis: Every digital image or video carries metadata, information about the file's origins, modifications, and more. AI can quickly sift through this data to flag potential discrepancies, hinting at tampering [12].
- Collaborative Filtering: AI can cross-reference an image or video across the internet, identifying its earliest sources and tracking its modifications over time. This can help in establishing the authenticity of viral visual content [13].

3.3. Adopting Diverse Training Data

The power and efficiency of an AI model largely depend on the data it's trained on. Biased data can result in biased outputs.

- Broadening Horizons: Data sources need to be diverse, encompassing various demographics, cultures, and perspectives. This ensures that the AI model has a holistic understanding and isn't skewed towards any group [13].
- Audits and Quality Checks: Regular introspection is essential. Third-party audits can assess the inclusivity of data sets, ensuring that no group is underrepresented or misrepresented [14].

DOI: 10.54254/2753-7048/36/20240443

- Feedback Loops: AI should be seen as a continually evolving entity. By establishing feedback loops with end-users and stakeholders, platforms can continually refine their AI models, making them more accurate and unbiased over time [12,13].

4. Tangible Outcomes with AI Integration

4.1. TikTok and AI-Driven Notifications

One of the most revolutionary aspects of our digital age is the ability to share and consume vast amounts of visual content. Platforms like TikTok, with its massive global user base, play a pivotal role in information dissemination. Recognizing its influence, TikTok recently took the step to integrate AI-driven notifications:

- Proactive Detection: Rather than waiting for user reports, TikTok's AI scans content uploaded to the platform, identifying possible manipulated or false visual and audio data. This proactive approach helps in curbing the spread of misinformation from the outset.
- User Education: When a user encounters potentially manipulated content, an AI-driven notification provides context. This not only flags questionable content but also educates users on discerning manipulated media, building a more informed user base.

4.2. Twitter's AI Efforts During Critical Times

During the COVID-19 pandemic, the world grappled not just with a health crisis but an "infodemic" – a deluge of information, both accurate and misleading. Twitter, as a primary source of real-time updates, recognized the urgency of the situation:

- Flagging Misleading Tweets: Using AI algorithms, Twitter began analyzing the content of tweets for potential misinformation regarding the virus, its spread, treatments, and implications. Flagged tweets would receive labels or warnings, guiding users to credible sources for confirmation.
- Monitoring Trending Topics: Twitter's AI kept a close watch on trending topics related to the pandemic. This ensured that misleading narratives or hoaxes didn't gain undue traction, and official health sources were more prominently displayed.

4.3. The Limits of AI

Despite the remarkable strides made in artificial intelligence, it's essential to recognize its limitations:

- Adaptive Malicious Actors: Cyber adversaries are always on the lookout for vulnerabilities. As AI models evolve, so do malicious strategies. New techniques for creating deepfakes, misleading content, or deceptive narratives can often outpace the preventive measures of current AI systems [9].
- Data Dependence: AI's strength – its reliance on vast amounts of data – can also be a weakness. Rapidly evolving situations or unique nuances might not be immediately reflected in the data AI has been trained on. This can lead to gaps in detection or false positives [9].
- Over-reliance and Complacency: Believing that AI will catch all misleading content can lead to a complacent attitude among users and platform moderators. It's crucial to maintain a balance, promoting human vigilance and critical thinking alongside automated checks [7].

5. Legal Frameworks and Policy Measures

Regulations provide clear guidelines for platforms, creators, and users, balancing freedom of expression and protection against misuse.

DOI: 10.54254/2753-7048/36/20240443

5.1. European Union's GDPR: Pioneering Data Protection

The General Data Protection Regulation (GDPR) of the European Union stands as a monumental step in ensuring the rights of users in the digital realm:

- User Consent: GDPR mandates platforms to obtain clear and explicit consent from users before collecting or processing their data. This has revolutionized how online entities approach user data, instilling a sense of responsibility and transparency [15].
- Right to Erasure: Also known as the 'right to be forgotten,' GDPR allows users to request the deletion of their data. This offers individuals control over their digital footprint, empowering them against potential misuse [15].
- Data Breach Notifications: Platforms must notify users and the relevant authorities of any data breaches promptly. This fosters trust and holds platforms accountable for lapses in data security [15].

5.2. Singapore's Anti-Fake News Approach

In response to the global concern of misinformation, Singapore enacted the Protection from Online Falsehoods and Manipulation Act (POFMA):

- Mandatory Corrections: Platforms that inadvertently host false content must issue corrections, ensuring that audiences receive accurate information [16].
- Tackling Malicious Actors: Those deliberately spreading falsehoods face stringent penalties. This is a deterrent, discouraging malicious intent and promoting responsible content creation [16].
- Transparency Requirements: Platforms need to disclose the sources of digital advertisements, enabling users to understand the origins of promotional content and assess its authenticity [16].

5.3. The Role of User Education

Legal frameworks, while essential, need to be complemented by informed users:

- Digital Literacy Programs: Countries and platforms can invest in digital literacy campaigns, teaching users to discern between genuine and manipulated content.
- Platform-Driven Initiatives: Platforms can incorporate built-in tutorials, tooltips, or pop-ups, educating users about potential risks and offering guidance on secure navigation.

6. Future Research and Innovations

6.1. Fostering Collaborative Ecosystems

The vastness and intricacy of AI demand a multi-faceted approach. No single organization, government, or community can single-handedly navigate the challenges or leverage the full potential of AI [8]. Thus, building bridges across different sectors becomes imperative:

- Inter-industry Collaborations: Companies across the tech spectrum, from startups to tech behemoths, should establish joint ventures and partnerships. Such collaborations can catalyze the pooling of resources, sharing of insights, and fostering innovations that one entity might not achieve alone.

6.2. Transparency and Public Scrutiny

Prominent tech entities, like Google and Microsoft, have taken strides towards a more open approach to AI:

- Open-sourced AI Models: By making their AI research and certain models open source, these corporations not only invite peer review but also democratize AI advancements. This transparency

DOI: 10.54254/2753-7048/36/20240443

aids in identifying biases, flaws, and potential improvements, ensuring a more resilient and robust AI infrastructure.

- Engaging the Academic Community: By allowing academic institutions access to their research, tech giants can facilitate the study and critique of AI from varied intellectual perspectives, fostering a richer dialogue and more nuanced advancements.

6.3. The Role of Grassroot Organizations

While global entities and governments play their part, the grassroots level is where the real impact of AI is felt:

- Advocacy for Marginalized Communities: Historically, many AI models have inadvertently perpetuated biases against certain communities. Grassroot organizations, with their on-ground insights, can provide valuable data and feedback to ensure AI algorithms are inclusive and representative.

- Education and Awareness: Beyond just advocacy, grassroots organizations can serve as educational hubs. By organizing workshops, seminars, and training sessions, they can empower local communities to understand, engage with, and even contribute to the AI landscape.

- Feedback Loop Creation: To ensure AI evolves with societal needs, a continuous feedback mechanism should be established [9]. Grassroot organizations can act as intermediaries, collecting and relaying feedback from end-users to tech developers, ensuring that AI tools remain relevant and user-centric.

6.4. Legislation and AI Governance

As AI's influence permeates every aspect of society, establishing clear legislative guidelines becomes crucial:

- Collaborative Lawmaking: Tech industries and legislative bodies should come together to draft regulations that strike a balance between innovation and ethical considerations. Such a collaborative approach ensures that laws are technologically sound and socially responsible.

- End-user Protection: With AI influencing choices ranging from media consumption to critical decision-making, regulations must be in place to safeguard end-users from potential harms, biases, or invasions of privacy [9].

7. Conclusion

In conclusion, this paper sheds light on the pervasive issues of Abuse, Bias, and Discrimination that have infiltrated the realm of social media. These issues encompass the spread of misinformation, online harassment, and the manipulation of public opinion, posing significant challenges to the integrity and inclusivity of these platforms. Moreover, the paper underscores the role of AI systems in perpetuating biases and exacerbating unfair information dissemination and user interactions on social media. This paper comprehensively examines real-world instances of bias, discrimination, and abuse on social media, underscoring the pressing need to address these issues. Secondly, it delves into multiple strategies for addressing these challenges, focusing on how artificial intelligence can enhance media practices, as evidenced by case studies and practical examples. The core objective of this paper is to underscore the pivotal role that social media plays in contemporary society while highlighting the associated risks. As a crucial tool for disseminating information, social media should ideally function as a public, secure, inclusive, and equitable platform. This research provides an avenue for harnessing AI technologies effectively to ameliorate the social media landscape, diminishing the impact of abuse, bias, and discrimination, augmenting the diversity of information dissemination, and elevating the overall value of social media. The central theme of this paper

revolves around the diverse methods and strategies through which Artificial Intelligence (AI) can be employed to identify and mitigate bias, discrimination, and abuse within social media platforms. It underscores the significance of data diversity and quality, advocating for implementing equitable and transparent fine-tuning algorithms. Additionally, the paper explores the necessity of implementing more stringent and precise regulations in the social media sphere while promoting user education. By deploying artificial intelligence algorithms, these measures can transform social media into a safer, more inclusive, and fairer space. As the paper concludes, it underscores the critical importance of continued research and development in this essential area, calling for a collective endeavor to establish a more responsible and equitable social media environment that benefits all.

References

[1] *Shin, D., Hameleers, M., Park, Y. J., Kim, J. N., Trielli, D., Diakopoulos, N., Helberger, N., Lewis, S. C., Westlund, O., & Baumann, S. (2022). Countering Algorithmic Bias and Disinformation and Effectively Harnessing the Power of AI in Media. Journalism & Mass Communication Quarterly, 99(4), 887–907. https://doi-org.turing.library.northwestern.edu/10.1177/10776990221129245*

[2] *Roose, K. (2018, November 4). We asked for examples of election misinformation. you delivered. The New York Times. https://www.nytimes.com/2018/11/04/us/politics/election-misinformation-facebook.html*

[3] *Reynolds, S. (2022, July 21). Cyberbullying linked with suicidal thoughts and attempts in young adolescents. National Institutes of Health. https://www.nih.gov/news-events/nih-research-matters/cyberbullying-linked-suicidal-thoughts-attempts-young-adolescents*

[4] *Social media manipulation by political actors an industrial scale. University of Oxford. (n.d.). https://www.ox.ac.uk/news/2021-01-13-social-media-manipulation-political-actors-industrial-scale-problem-oxford-report*

[5] *Pflueger-Peters, N. (2022, December 14). Do youtube recommendations foster political radicalization?. Computer Science. https://cs.ucdavis.edu/news/do-youtube-recommendations-foster-political-radicalization*

[6] *Feiner, L. (2022, June 22). DOJ settles lawsuit with Facebook over allegedly discriminatory housing advertising. CNBC. https://www.cnbc.com/2022/06/21/doj-settles-with-facebook-over-allegedly-discriminatory-housing-ads.html*

[7] *Mohseni, S., & Ragan, E. (2018, December 4). Combating fake news with Interpretable News Feed Algorithms. arXiv.org. https://arxiv.org/abs/1811.12349*

[8] *Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual Mis- and Disinformation, Social Media, and Democracy. Journalism & Mass Communication Quarterly, 98(3), 641–664. https://doi-org.turing.library.northwestern.edu/10.1177/10776990211035395*

[9] *Packin, N. G. (2020). Disability discrimination using AI systems, social media and digital platforms: Can we disable digital bias? Journal of International and Comparative Law 8.2, SSRN Electronic Journal, 487–511. https://doi.org/10.2139/ssrn.3724556*

[10] *Bossetta, M. (2018). The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. Journalism & Mass Communication Quarterly, 95(2), 471–496. https://doi-org.turing.library.northwestern.edu/10.1177/1077699018763307*

[11] *8 J. Int'l & Comp. L. 487 (2021) Disability Discrimination Using Artificial Intelligence Systems and Social Scoring: Can We Disable Digital Bias?, Packin, Nizan Geslevich [26 pages, 487 to 512]*

[12] *Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., & Subbian, V. (2020). Stigma, biomarkers, and algorithmic bias: Recommendations for Precision Behavioral Health with Artificial Intelligence. JAMIA Open, 3(1), 9–15. https://doi.org/10.1093/jamiaopen/ooz054*

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

[14] *Ibrahim, H., AlDahoul, N., Lee, S., Rahwan, T., & Zaki, Y. (2023). YouTube's recommendation algorithm is left-leaning in the United States. PNAS Nexus, 2(8), 264. https://doi.org/10.1093/pnasnexus/pgad264*

[15] *Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Review, 55(Issue 4), 1143–1185. https://doi.org/10.54648/cola2018095*

[16] *Schuldt, L. (2021). Official Truths in a War on Fake News: Governmental Fact-Checking in Malaysia, Singapore, and Thailand. Journal of Current Southeast Asian Affairs, 40(2), 340–371. https://doi.org/10.1177/18681034211008908*