An Automated Framework to Identify and Eliminate Systemic Racial Bias in the Media

Lamogha Chiazor[1], Geeth de Mel[2], Graham White[3], Gwilym Newton[4], Joe Pavitt[5], Richard Tomsett[6]

IBM Research Europe (UK), Hursley, Winchester, UK

{lamogha.chiazor[1],geeth.demel[2],gwhite[3], gwilnew[4], joepavitt[5], rtomsett[6]}@uk.ibm.com

Abstract

The impact of the media on the world's stage is evident. It has the power to narrate the discourse—be it political or entertainment. Given the evolving landscape of bias in the world and the crucial role and power the media plays, we argue for technology playing its critical part in being a civic society's gatekeeper. This paper aims to propose and discuss a set of techniques that can come together as a technical framework to address the issues of systemic racism in the media. We have identified a set of data sources that can be useful in developing and evaluating the proposed framework. We have acknowledged the potential pitfalls of the approach in some contexts and means to mitigate them.

Introduction

Impact of the media (e.g., CNN, BBC, The Times and Wall Street Journal) on the world is phenomenal and has the power to influence people's opinions, emotions and actions (McCombs and Reynolds 2002; Kennedy and Prat 2019). This power makes it crucial that we continue to tackle systemic racism in information from the media. For example, in recent times, we continue to witness more commercials showcasing the abject poverty and sufferings of people in Africa and other black communities worldwide (Kendi 2020); as opposed to the number of times we see in the news the representation of affluent areas and inspiring stories of those same communities and their world-leading successes. We recognise that the sufferings are real, and the commercials run by media outlets to solicit for funding to help those suffering might be well intended. However, we believe there is some level of poverty, suffering, and goodness across all countries and nations worldwide. Therefore, we propose and discuss an automated technical analysis framework that could be designed to assist media workers in representing a balanced view of non-Caucasians in these Caucasian communities around the world.

The rest of the paper is structured as follows: In the next section, we present some related work highlighting some issues of systemic racial bias in the media and discuss some technological solutions to mitigate those problems. Then, we briefly discuss a collection of techniques that can come together to address the problem more holistically. Following that, we discuss potential data sources to test the proposed framework and present potential pitfalls of such a system and some remedies for them. We discuss the potential expansion of the framework in the penultimate section and conclude the document by providing concluding remarks.

Background: A survey on systemic racial bias in the media

Works from the Social Sciences

Fair (1993) focused on addressing and highlighting some systemic racial issues arising from the construction of Africa and Africans as the *Other* in American news media. Reviewing how the selection of news happens as a manual process involving sorting or sifting through incoming information and then media workers deciding what becomes news based on personal motives or organisational profits—which is one of the root causes of racial bias in the media—was discussed. Research scholars and working journalists concerned with the process by which media organisations select news for coverage, postulates that political, social, economic, cultural and geographic attributes of a country will often determine or predict the amount of coverage that a country might receive in the press of another country. Further highlighted are some examples of American news coverage over the years that uses some terminology that instigates systemic racial separations and bias in news media e.g., the use of the phrase *black-on-black violence* instead of just violence OR how the term *tribal* was used in reports by news outlets when referring to civil war.

Other works of literature (Gruley and Duvall 2012; Baker 2015; Adegbola, Skarda-Mitchell, and Gearhart 2018; Hammett 2011) highlight several issues such as how certain terminology is used and how news and information from media organisations are framed to instigate or encourage racial bias. On the flip side, the authors in (Nothias 2018) empirically analyse how these issues have improved over time. These background information sources serve as reference points in the design of the technical solution framework we propose.

Bunce, Franks, and Paterson (2016) discuss how provocative work by leading media researchers and experts in the complexity of African societies and politics have come together to discuss and consider change and continuity in the

portrayal of Africa. They discuss how today's news media with little funding are under more pressure than ever before to meet high standards of accuracy. However, what we do not see from the book are technical solutions that have the potential to help ease this pressure.

Hypotheses (e.g., negative valence tending to appear more often than positive or neutral valence over time periods OR episodically framed stories appearing more often than thematically framed stories across time periods) and the research question (how did coverage of issues vary between later and more recent time periods?) brought forward by the authors in (Adegbola, Skarda-Mitchell, and Gearhart 2018) — though focused on the portrayal of Nigerian news coverage on US broadcast networks over specific historical periods — also forms part of the base points we have taken into consideration for the technical solutions we propose.

Works from the movie industry

There are a few analytical frameworks and tests we can learn from that are currently used to address bias and inequality in the movie industry. For example, as discussed in (Hickey et al. 2017), the Bechdel-Wallace test mainly addresses gender inequality in the movie industry. However, by exploring and expanding to 11 more tests[1] (e.g., in *the Waithe test*, *the KO test* and *the Villalobos test*), the authors begin to address some racial issues within the movie industry, such as the portrayal or dehumanisation of non-Caucasian characters as stereotypes. The current Bechdel-Wallace test has many shortcomings, both as a test of gender equality and bias in general. For example, the proposed test attempts to look at fictional characters' relationships and behaviours, thus allowing a much higher bias towards the fictional character than the real world. Although, we acknowledge that this sets a precedent of its own for influencing the perceptions and subconscious nature of the human mind in terms of representation.

The Portrayal Vs. Betrayal study (Council and Interactive 2011) by the UK film council used a qualitative approach asking various audiences how they feel about different groups being represented in films. From their approach, the authors established and concluded several points including: "films having the power to influence mindsets" or "filmmakers needing to spend more time to reflect and strategize on how all parts of society are portrayed" fairly.

Madaan et al. (2018) analyse, detect and remove gender stereotyping biases from Bollywood movies. The authors do so by leveraging techniques such as Natural Language Processing/Understanding (NLP/NLU), image understanding and semantic graphs, whilst considering features such as occupation, associated actions and descriptions. The algorithm they proposed for removing stereotypes enabled them to analyse movie plots and posters from the 1970s until 2017 and show how bias in Bollywood movies decreased over the last three years of that time period. The work proves some positive changes and improvements over the years towards eliminating bias in the media. This gives us some hope that with a bit more focus and work in this space, e.g., by extending and building more technical solutions for not just the movie industries but across all media sectors, we can get to the point where these stereotyping (which is a form of systemic racial bias) is eliminated. It is also interesting how they capture and analyse the gender biases from images using deep image analysis of promotional posters for the said movies. They also do not just rely on the textual intra-sentence (per sentence analysis, no context used) and inter-sentence (a sentence analysed in the context of another sentence) analysis of the associated text data (which contains mainly fictional characters), but they go further to capture details about the casts (non-fictional) to build a complete view on the movie as a whole. Tasks at the intra-sentence level involved the analysis of: how many times a female cast is referred to in the plot versus a male cast; using verbs and adjectives to determine how male and females casts are addressed; the introductions of male and females casts in a plot; occupation as a stereotype and how gender diverse singers in soundtracks are. At the inter-sentence levels they construct a context flow between sentences via a word based knowledge graph using dependency parsers. On the knowledge graphs they perform: a technical node based analysis to determine how much a cast is focused on in a plot (they called this the 'Centrality of each cast node'); and use word embeddings to study bias patterns in the knowledge graphs - via a joint modelling of verbs, adjectives and relations in the graph. Finally, for the bias removal system that they propose (called DeCogTeller), the authors actually made use of a news article data set to train word embeddings using *word2vec* (Mikolov et al. 2013), action extraction, word pairs classification (e.g., into *gender neutral* or *gender specific*) and developed a specialised module for handling occupations. Their knowledge base datasets consisted of fact-based and biased data points. DeCogTeller (Madaan et al. 2018) attempts to eliminate gender bias in the movie scripts by switching gender roles in a movie plot when a gender bias has been identified from the co-referenced based knowledge graphs constructed - via the help of NLU techniques.

Works from the field of computer science

Kiela et al. (2020) technically construct a challenging data set for multi-modal classification with a focus to detect hate speech in multi-modal memes (i.e., memes with both textual and visual characteristics). They reiterate how common memes are on the internet, especially on social media and although subtle, the true meaning of a meme might be more straightforward for a human to detect but difficult for AI systems. Their challenge set was designed so that only models successful at complex multi-modal reasoning or understanding can accurately detect hate speech. By flipping images and text contents of a meme with alternates, they could reconstruct several sourced memes (originally 162k memes posted on public social media groups from within the United States) from scratch. They outsourced to an external company the annotations into categories 'yes' or 'no' for whether a meme is hateful. Their goal was not to use these challenge sets to train their models from scratch but mainly to fine-tune and test large-scale, pre-trained, multi-modal

[1] https://projects.fivethirtyeight.com/next-bechdel/

models. By collecting contrasting or counterfactual examples of memes annotated as hateful, they can make the data set much more challenging. The results of analysing several text-based or visual-based models on their challenge set determine that with human accuracy at 84.7% and the best multi-modal models still performing at 64.73%, is telling of how much improvement is needed for the state of the art multi-modal models.

An interesting comparison to our discussions and proposed technical solution is the work of (Hamborg, Donnay, and Gipp 2019), who also consider expertise knowledge on the topic from the social sciences whilst analysing ways in which computer science can contribute to the identification of bias in the media—thus calling for the design of an inter-disciplinary solution for media bias research. By contrasting and comparing known social science research methods about bias in the media with technical approaches from computer science, they are able to draw conclusions and highlight ways in which research in computer science (e.g., NLP/NLU) can be used to make distinctive contributions in the study of media bias. They propose that just like the manual analyses carried out in the social sciences, automated solutions will need to consist of methods to obtain news articles relevant to the topic, link the articles to a baseline or other articles, and compute some statistics on the linked data. Considering techniques such as *event detection* or *document clustering* or *news aggregation*, the authors suggest current limitations and promise in their usage for news linkage to recognise patterns of bias in the media. They also agree with our discussions that there is a current lack of technical research approaches that specifically try to resolve media bias caused as a result of commission or omission, and that graph analysis amidst some other techniques like *Centering resonance analysis (CRA)* are up-and-coming candidates for this. Nonetheless, we observe that a lot of their proposed technical solutions in comparison to ours will facilitate the identification of bias in the media without any potential technical solutions proposed that will help mitigate such biases after it has been identified or in future situations.

Potential technical solutions to reduce systemic racial biases in the media

With the knowledge and information about the news selection process, or specific attributes that can be used to determine or predict the amount of coverage a country might have in another country and issues around some terminology used when referring to non-Caucasian communities in news media – we propose a technical framework made up of but not limited to:

Main aim developing an automated system for news selection and analysis.

Predictive machine learning (ML) models trained using historical data sets of news media that contain political, social, economic, cultural and geographic attributes of a country – to predict if a particular piece of news will instigate any racial bias or portrayal of minority groups. Moreover, a weekly or monthly predicted probability of news media outlets meeting a balanced ratio for the positive vs negative portrayal of minority groups will help the continuous process of eliminating systemic racism in the media.

Automated knowledge graph construction that will depict the semantic relationships of selected news with a positive, negative or neutral portrayal of the minority or related systemic racial terms or strategies.

NLP/NLU to analyse and fine-tune terminology used by news media when reporting. We propose having a technical solution where if a term or phrase is analysed or predicted to instigate any racial bias or negative portrayal of minority groups, then suggestions to replace such a term or phrase is provided automatically.

Potential data sources to facilitate the analysis

The analysis of systemic racial bias in movies can be achieved by sourcing the movie scripts through websites like "Internet Movie Screenplay Database (IMSDB)" or the likes.

Similar to the data collection and analysis of the Darfur conflict news in (Gruley and Duvall 2012), we propose the use of tools such as LexisNexis Academic Knowledge Centre (Knapp 2018) to access full-text news articles, as well as the use of content retrieval Application Programming Interfaces (APIs) from news media organisations such as the American Broadcasting Company resource API or msnbc.com APIs. Rapidapi.com, for example, in a recent blog entry (Janet Wagner 2021)[2] provided an updated review of the top 10 best news APIs out of over 61 news APIs for accessing various media data sets.

Potential issues elevated due to the proposal and means to resolve them

A critical area for bias in systems and models' design often stem from a given human's intrinsic biases. They are usually a reflection of ourselves. One possible solution to this problem is to ensure that a diverse group of individuals are involved from the inception of the solutions design to the testing phase of such technical solutions—this is one reason why we gathered a diverse group of authors to be involved in the discussions presented in this paper.

A second area for bias surrounds the definition of *systemic racism*. Although several attempts are being made to define this, this is still somewhat of a grey area. A possible solution to this will be to develop any technical solutions within a narrowly and well-defined context of what *systemic racism* is within the media context. For example, it could be a case of starting the definition as:

> *one aspect of systemic racial bias in the media can manifest as news that attempts to portray in a negative light any minority group more than it portrays in a negative light any majority group and vice versa.*

A third potential area for bias may be the data sets used for implementing or developing the proposed solution. For example, any historical data set sourced might not contain enough information (including semantic information) that

[2]https://rapidapi.com/blog/rapidapi-featured-news-apis/

depicts our example definition of systemic racial bias in the media. One possible way to combat this is to investigate ways to build more robust training data sets or investigate ways to introduce learning methods that do not rely on existing historical data sets to build such a system.

Broader problems may arise in any situation where technology is naively applied to solve a societal issue. As envisaged, our framework should be applied as a means to help people working in the media improve their output with respect to racial bias and diversity of representation. However, as warned by Goodhart's law (Manheim and Garrabrant 2018), if the measures and metrics suggested here become targets, they will cease to be useful. This is challenging to prevent and will require careful consideration before deploying the proposed framework in the wild.

Finally, we must ensure that the framework takes a suitable nuanced, inter-sectional view of bias. While in this paper, we have focused on race, characteristics determined by other factors including gender, sexual orientation, disability and so forth will result in potentially complex interactions in their media representations. This must be carefully considered when any technical framework is implemented.

Discussion: future work

Below, we sketch the potential future extensions of this proposed solution.

In our opinion, literature shares similar issues as movies, though its likely targets would need to be adjusted. An average screenplay is around 7,500 — 10,000 words, whereas a novel will come in around 100,000 to 150,000 words. Suppose something like a variant on the Bechdel-Wallance test would be applied to a novel its more likely to pass as there is a higher chance the required *events*. In this case, two characters of colour (CoCs) talking would occur just by pure chance. In a situation like this, we would suggest that we would need to assess both the semantic importance of conversation to the plot and the ratio of these to base factors (i.e., the ratio of CoC conversations to other conversations). It is worth noting that we are more likely to experience issues directly getting the metadata required for analysis as IMDB-like services do not exist for books. However, we would have more raw text, enabling the use of an NLP based approach to infer this data. Many books also have a glossary and characters that would make this type of test easier.

Another potential avenue is video games as they provide a vast opportunity for new tests as they offer two things that neither books nor movies can: (1) control of a character; and (2) choices made by the player during the game. In the case of the player-character, it would further depend if the character is authored or not. Examples of authored characters would include Aloy from Horizon Zero Dawn[3] or Booker from Bioshock[4]. They are created by the developers and have a set race, background, personality, and path. Unauthored characters would include The Farmer from Stardew Valley[5] or even Steve from Minecraft[6] and are for the most part blank slates. Given these variations, we can test a multitude of hypothesis via our proposed framework, if extended to this domain. For example,

1. from authored characters, we can explore the developers' choices, about race, voice, background and opinions.
2. from unauthored characters, we can test the players' options to build CoCs or express other aspects of identity that are overlooked in minority groups. A recent example of this is 2020's Cyberpunk 2077 (Eklundh 2020) that allowed players to be transgender when creating their characters, the first mainstream title to do so.
3. are there non-player characters (NPCs) of colour or other poorly represented groups? We should examine their presentation and dialogue, does it fall into dangerous stereotypes? e.g., CoCs in abusive relationships and only working in lower-status jobs.
4. can we interact with NPCs, and what is the ratio of them that can be interacted with compared to others? Moreover, when we interact, are we the character given the possibility of a positive reaction? Furthermore, if the NPC is harmed, is it played the same as when a white NPC is harmed? Is it treated less seriously?

Conclusions

Our hypothesis is that using extensive qualitative research methods to collect and analyse text data will facilitate better understanding and more in-depth insights into concepts or experiences of systemic racism in question.

Automating the process for analysing and interpreting the data collected above will eliminate or reduce qualitative research issues of unreliability, subjectivity, limited generalisability or labour-intensity. For example, approaches for qualitative data analysis which includes *content analysis* (i.e., describing and categorising common words, phrases and ideas in qualitative data), *thematic analysis* (i.e., identifying and interpreting patterns and themes in qualitative data) or *textual analysis* (i.e., examining content, structure and design of texts) can be achieved effectively and efficiently using NLU and knowledge graphs (Bhandari 2020). Creating predictive ML models that media outlets can use before a news article or report is released, to predict or automatically check for any racially biased perception occurring as a result of such a report been released, will aide any necessary adjustments to be made to the information prior to its public release.

References

Adegbola, O.; Skarda-Mitchell, J.; and Gearhart, S. 2018. Everything's negative about Nigeria: A study of US media reporting on Nigeria. *Global Media and Communication* 14(1): 47–63.

[3]https://horizon.fandom.com/wiki/Aloy

[4]https://bioshock.fandom.com/wiki/Booker_DeWitt

[5]https://stardewcommunitywiki.com/The_Farm

[6]https://minecraft.gamepedia.com/Player

Baker, A. 2015. Race, paternalism, and foreign aid: Evidence from US public opinion. *American Political Science Review* 109(1): 93–109.

Bhandari, P. 2020. An introduction to qualitative research. *Noudettu osoitteesta https://www. scribbr. com/methodology/qualitativeresearch/#:˜: text= Qualitative% 20research% 20involves% 20collecting* 20.

Bunce, M.; Franks, S.; and Paterson, C. 2016. *Africa's Media Image in the 21st Century: From the" Heart of Darkness" to" Africa Rising"*. Routledge.

Council, U. F.; and Interactive, H. 2011. Portrayal vs. betrayal: An investigation of diverse and mainstream UK film audiences.

Eklundh, H. 2020. Cyberpunk 2077. In *ACM SIGGRAPH 2020 Computer Animation Festival*, 1–1.

Fair, J. E. 1993. War, famine, and poverty: Race in the construction of Africa's media image. *Journal of Communication Inquiry* 17(2): 5–22.

Gruley, J.; and Duvall, C. S. 2012. The evolving narrative of the Darfur conflict as represented in The New York Times and The Washington Post, 2003–2009. *GeoJournal* 77(1): 29–46.

Hamborg, F.; Donnay, K.; and Gipp, B. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20(4): 391–415.

Hammett, D. 2011. British media representations of South Africa and the 2010 FIFA World Cup. *South African geographical journal* 93(1): 63–74.

Hickey, W.; Koeze, E.; Dottle, R.; and Wezerek, G. 2017. We Pitted 50 Movies against 12 New Ways of Measuring Hollywood's Gender Imbalance. *FiveThirtyEight. com* .

Janet Wagner. 2021. Top 10 Best News APIs (Updated for 2021). https://rapidapi.com/blog/rapidapi-featured-news-apis/.

Kendi, I. X. 2020. Stop blaming Black people for dying of the coronavirus. *Atlantic* .

Kennedy, P. J.; and Prat, A. 2019. Where do people get their news? *Economic Policy* 34(97): 5–47.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.

Knapp, J. A. 2018. Nexis Uni. *The Charleston Advisor* 19(3): 31–34.

Madaan, N.; Mehta, S.; Agrawaal, T.; Malhotra, V.; Aggarwal, A.; Gupta, Y.; and Saxena, M. 2018. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 92–105. New York, NY, USA: PMLR. URL http://proceedings.mlr.press/v81/madaan18a.html.

Manheim, D.; and Garrabrant, S. 2018. Categorizing variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585* .

McCombs, M.; and Reynolds, A. 2002. News influence on our pictures of the world. .

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations: Workshops Track* URL https://arxiv.org/abs/1301.3781v3.

Nothias, T. 2018. How Western journalists actually write about Africa: Re-assessing the myth of representations of Africa. *Journalism Studies* 19(8): 1138–1159.