Mitigating hate speech in Nigeria: The possibilities of artificial intelligence

Joseph Wilson
Rahila Jibrin
University of Maiduguri. Department of Mass Communication
joeweee2003@gmail.com
rayjib2013@gmail.com

Submission date: November 2018
Accepted date: November 2019
Published in: December 2019

Recommended citation: WILSON, J. and JIBRIN, R. (2019). "Mitigating hate speech in Nigeria: The possibilities of artificial intelligence". *Anàlisi: Quaderns de Comunicació i Cultura*, 61, 17-30. DOI: <https://doi.org/10.5565/rev/analisi.3188>

Abstract

Hate speech has become a global concern. Nations worldwide in one way or another have to grapple with the enormous problem of this phenomenon, which is predominantly perpetrated through new media or online media platforms. In Nigeria, the situation is such that governments at the federal and state levels have continued to express concern over the growing wave of hate speech in the country. While technology propels this phenomenon, technology may just be the solution. Technology has no doubt continued to offer humanity several possibilities to better human existence. Increasingly, it is becoming an indispensable part of the daily life of individuals. The mobile phone, for instance, is used as a typewriter, a calculator, a calendar, a time piece, a communication system, an interactive database, a decision-support system and much more. In recent times, the insatiable drive for technology has reached a point where devices act intelligently. These intelligent systems are rapidly developing for use to enhance human endeavours. Artificial intelligence (AI) technologies, driven by big data, are fuelling unprecedented changes in many facets of human endeavours. Many achievements using AI techniques surpass human capabilities. If machines can recognise speech and transcribe it – just like typists did in the past – if computers can accurately identify faces or fingerprints from among millions, cars drive themselves and robots fight wars, among other remarkable things, there is no doubt there would be a way round the complex challenge of hate speech. Therefore, this study examines the inherent possibilities of AI for mitigating hate speech in Nigeria.

Keywords: Artificial Intelligence; hate speech; mitigation; Nigeria

Resum. *Mitigar el discurs d'odi a Nigèria: les possibilitats de la intel·ligència artificial*

El discurs d'odi s'ha convertit en una preocupació global. Les nacions mundials d'una manera o d'una altra han d'afrontar l'enorme problema d'aquest fenomen, perpetrat majoritàriament a través de nous mitjans de comunicació o plataformes de comunicació en línia. A Nigèria, la situació és tal que els governs dels nivells federal i estatal han continuat manifestant la seva preocupació per la creixent onada de discurs d'odi al país. Si bé la tecnologia propicia aquest fenomen, la tecnologia podria ser la solució. La tecnologia, sens dubte, ha continuat oferint a la humanitat diverses possibilitats per a una millor existència humana. Cada vegada s'està convertint en una part indispensable del dia a dia dels individus. El telèfon mòbil, per exemple, ha esdevingut imprescindible en la nostra vida quotidiana. S'utilitza com a màquina d'escriure, calculadora, calendari, rellotge, sistema de comunicació, base de dades interactiva, sistema de suport de decisions i molt més. En els darrers temps, l'impuls insaciable per la tecnologia ha arribat a un punt en què els dispositius actuen de manera intel·ligent. Els sistemes intel·ligents es desenvolupen ràpidament per utilitzar-los per millorar els esforços humans. Les tecnologies d'intel·ligència artificial (IA), impulsades per dades massives, estan alimentant un canvi sense precedents en moltes facetes dels esforços humans. Molts assoliments que utilitzen tècniques d'IA superen les capacitats humanes. Si les màquines poden reconèixer la parla i transcriure-la, tal com ho feien els mecanògrafs del passat, si els ordinadors poden identificar amb exactitud rostres o empremtes dactilars entre milions, si els cotxes es condueixen ells mateixos i si els robots estan lluitant en guerres, entre altres coses destacables, no hi ha dubte que hi ha una manera d'evitar el complex desafiament del discurs d'odi. Per tant, aquest estudi examina les possibilitats inherents a la intel·ligència artificial per mitigar el discurs d'odi a Nigèria.

Paraules clau: mitigació; discurs d'odi; intel·ligència artificial; Nigèria

Resumen. *Mitigación del discurso de odio en Nigeria: las posibilidades de la inteligencia artificial*

El discurso de odio se ha convertido en una preocupación mundial. Las naciones de todo el mundo de una forma u otra tienen que lidiar con el enorme problema de este fenómeno, predominantemente perpetrado a través de nuevos medios o plataformas de medios en línea. En Nigeria, la situación es tal que los gobiernos de los niveles federal y estatal han seguido expresando preocupación por la creciente ola de discurso de odio en el país. Si bien la tecnología impulsa este fenómeno, la tecnología podría ser la solución. La tecnología, sin duda, ha continuado ofreciendo a la humanidad varias posibilidades para mejorar la existencia humana. Cada vez más se está convirtiendo en una parte indispensable de la vida diaria de las personas. El teléfono móvil, por ejemplo, se ha vuelto indispensable en nuestra vida cotidiana. Se utiliza como máquina de escribir, calculadora, calendario, reloj, sistema de comunicación, base de datos interactiva, sistema de soporte de decisiones y mucho más. En los últimos tiempos, el impulso insaciable por la tecnología ha llegado a un punto en el que los dispositivos actúan de manera inteligente. Los sistemas inteligentes se están desarrollando rápidamente para mejorar los esfuerzos humanos. Las tecnologías de inteligencia artificial (IA), impulsadas por macrodatos, están impulsando cambios sin precedentes en muchas facetas de los esfuerzos humanos. Muchos logros que utilizan técnicas de IA superan las capacidades humanas. Si las máquinas pueden reconocer el discurso y transcribirlo, tal como lo hicieron los mecanógrafos en el pasado, si las computadoras pueden identificar con precisión rostros o huellas dactilares de entre millones, si los coches se conducen solos y si los robots están peleando en guerras, entre otras cosas notables, no hay duda de que hay una forma de sortear el complejo desafío del

discurso de odio. Por lo tanto, este estudio examina las posibilidades inherentes a la inteligencia artificial para mitigar el discurso de odio en Nigeria.

Palabras clave: mitigación; discurso de odio; inteligencia artificial; Nigeria

1. Introduction

There is global concern about the hate speech phenomenon. Nations are increasingly mindful of communications that are considered to express hatred for individuals or groups in terms of social characteristics such as race, ethnicity, gender, religion, sexual orientation and other defining attributes of human beings. Widely known as hate speech, this kind of communication has become common with the ubiquitous nature of new media that makes information dissemination so easy, snappy and with global reach. Online platforms such as social media, private and public messaging forums and blogs, as well as the conventional media platforms such as print and broadcast media serve as channels for hate speech to occur, be disseminated and amplified. Ring (2013: 1) noted that hate speech is widespread on social media such that "A quick glance through the comments section of a racially charged YouTube video demonstrates how pervasive the problem is". Galeon (2017: 1) pointed out that online hate speech has become more common than real life hate speech because it is easier to be behind the computer screen.

The daily reality of hate speech has bedevilled communities for a long time and can have harmful effects even in a global context (Palfrey, 2018; Oloja, 2018). The worries associated with hate speech is not far from the high possibility it has in instigating or triggering violent conflict. The Nigeria Stability and Reconciliation Programme (NSRP, 2017: 1) has described hate speech as a catalyst for violence because it is considered generally offensive communication, noting that hate speech content "can create a vicious cycle as audiences convene around it and by acting as an alternative source of information that neutralises positive information". This has compelled nations into adopting measures to ensure that the menace of the phenomenon is addressed. Opusunju (2017) reported that governments in various countries are worried by hate speech perpetrated through various online platforms and described it as 'digital menace' of social media. He further noted that nations, such as France, Germany and Kenya, among others, have introduced some forms of measures to curb online and social media hate speech.

Nigeria, like other nations of the world, is not immune to hate speech. Indeed, the country's diversity has made it even more vulnerable to hate speech. Its multi-ethnic, multi-religious and multi-cultural characteristics has made it prone to sharp divides that influence political and social affiliations in addition to other issues instigating hate speech. The situation is such that government officials at federal and state levels have continued to express concern over the proliferation of hate speeches in the country, especially in social and mainstream media (Ehikioya, 2018; Falana, 2017). Opusunju (2017: 1)

noted that Nigerian "authorities are worried that if stiff penalties are not imposed on online hate propagation, violent conflicts will grow unchecked". Therefore, efforts to minimise hate speech are necessary at this point in time.

In this regard, some countries have taken steps to curb this dangerous trend of communication. For example, France, Germany and Kenya, among others, have passed legislation on hate speech. Moreover, there are reports that a security meeting held in 2017 led to a directive for "security agencies to monitor conversations and posts of prominent social media buffs as part of the processes of putting hate mongers to check" (Opusunju, 2017). The Nigerian legislature also made an effort to enact a law to check hate speech. Among other measures, the proposed bill sought the establishment of an Independent National Commission for Hate Speech, which shall enforce hate speech laws across the country to ensure that any person found guilty of any form of hate speech that results in the death of another person shall die by hanging upon conviction (Utomi, 2018). How effective these efforts are across nations remains debatable, as the phenomenon continues to be a concern (Palfrey, 2018; Oloja, 2018).

Onyibe (2017) noted that "desperate actions require dynamic responses, but not equally desperate reactions such as the introduction of draconian laws like the monitoring of hate speeches in the social media". He suggested that the simple mitigation of hate speeches perpetrated via online platforms lies in "information managers' ability to be ahead of the game or even just a few steps behind the hate speech purveyors". The increasing wave of hate speech on social media in recent years has made it imperative to recognise that effective counter-measures rely on automated data mining techniques (Zhang et al., 2018). Similarly, Galeon (2017) noted that stopping the flurry of hate speech, especially on social media, is difficult, and thus we are turning to artificial intelligence (AI) for a possible solution. It is against this backdrop that this paper examines the potentials of AI in mitigating hate speech in Nigeria, especially hate speech online. The objective of the paper is to identify the ways AI can be used to mitigate hate speech in Nigeria.

2. Literature review

2.1. Hate speech

Hate speech has gained significant global attention in recent times. However, its definition has remained contestable. Gagliardone et al. (2015) argued that the concept of hate speech lies in a complex link with freedom of expression, individual, group and minority rights, as well as terms related to dignity, liberty and equality. According to Brown (2017), significant attention has been focused on critical discourse and arguments about the various efforts towards checking hate speech as opposed to the earlier attempt to conceptualise hate speech itself. Gagliardone et al. (2015: 1) argued that "hate speech is a broad and contested term". They further noted that:

> Multilateral treaties such as the International Covenant on Civil and Political Rights (ICCPR) have sought to define its contours. Multi-stakeholders processes (e.g. the Rabat Plan of Action) have been initiated to bring greater clarity and suggest mechanisms to identify hateful messages.

Zhang et al. (2018: 3) noted that there has been an increasing number of research on hate speech detection as well as other related areas; thus, the "term 'hate speech' is often seen to co-exist or become mixed with other terms such as 'offensive', 'profane', and 'abusive languages', and 'cyber bullying'". To distinguish them, the authors identified hate speech as that which targets individuals or groups on the basis of their characteristics with a clear intention to incite, harm or to promote hatred which may or may not use offensive or profane words. The British Institute of Human Rights (2012: 8) defines hate speech as a term:

> covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.

According to Gagliardone et al. (2015), Internet platforms that mediate online communication, especially social media platforms such as Facebook, Google, Twitter or Instagram, have advanced their own definitions of hate speech that provide some kind of rules that guide the content on these platforms, thus limiting certain forms of expression.

While the argument continues in various academic and legal forums as to what constitutes hate speech, there are reasonable conceptualisations that serve as frames for understanding the meaning of hate speech, which we adopt in this paper. Onanuga (2018) defines hate speech as any online or offline communication that "expresses hatred for some group, in terms of race, ethnicity, gender, religious [sic], sexual orientation and others [sic] defining attributes of mankind". It is also defined as a term mixing concrete threats to individuals and groups' security with cases in which people may be simply expressing their anger against authority (Gagliardone et al., 2015). Hate speech is any expression conveyed through text, images or sound that contains degrading or dehumanising expressions or content targeted at individuals or groups and functions to dehumanise and diminish such an individual or group (Waldron, 2012). This kind of expression usually results in violence. Hopko (2018) noted that the most extreme examples of hate speeches are direct threats to individuals or groups, such as doxing, where people with malicious intent publish personal information that puts someone in harm's way and leaves them vulnerable to possible attacks or unwanted attention. The broadcasts of genocidal instructions by Radio Télévision Libre des Mille Collines in Rwanda in 1994 influenced citizens and increased

violence in communities with complete radio coverage, resulting in 9% of all deaths, of a total of 45,000 Tutsi people, which were attributed to violent acts incited by the station (NSRP, 2017). This also goes to show that although the mainstream media (print and broadcast) might have clear codes of conduct regarding content that incites violence, there are loopholes in the regulation that could lead to hate speech, not to mention the fewer available means to control such speech on social media, which is rife with inadequate regulation.

According to Galeon (2017), in an ideal world, the best way to mitigate hate speech is an individual's good sense of decency and respect for fellow human beings, irrespective of the diversity of socio-cultural characteristics (opinion, race, gender, ethnic and religious affiliation). As he states, "however, we don't live in an ideal world. As such, hate speech abounds, and the relatively free space social media offers us has given it a platform that's equally destructive – or perhaps even more". Interestingly, technology has continued to offer humanity several possibilities to better human existence and to protect humanity when the need arises. The Internet, for instance, has become indispensable in our everyday life. Along with other devices, it is used to shape global communication systems. These intelligent systems are rapidly emerging for use to enhance human endeavours. AI technologies, driven by big data, are fuelling unprecedented changes in many facets of human life. In recent years, there have been impressive advances in the field of AI that have resulted in inventions that were probably never imagined or thought possible. For instance, computers and other smart devices now have the capacity to learn how to improve performance and make decisions in various sectors of society using algorithms. In fact, a number of achievements using AI techniques already surpass human capabilities (Ganascia, 2018).

Developments in the communications field, made visible through the numerous emerging communication technologies such as mobile technologies and Internet with its speed and reach, make it difficult for governments to effectively enforce legislative curbs in the virtual sphere (Gagliardone et al., 2015). Freedom in the virtual world has made it very easy for the proliferation of hate speech globally. This is further heightened by the low awareness and understanding of hate speech amongst both media workers and members of the public. This has led to the unconscious publication of hate (NSRP, 2017). The NSRP reported that 76% of hate speech messages in Nigeria are

> transmitted through Facebook, either as a post on a private page or in a group, a post n a public page or group or as a response to a post or forum. The remainder of messages are transmitted through online articles or on Twitter (NSRP, 2017: 3).

The most prevalent messages call for discrimination (45%), for war (38%) and advocate the killing of others (10%). This online speech tends to

be actively recirculated by audiences with over 75 percent of messages receiving moderate to significant responses and observation (NSRP, 2017). According to Zhang et al. (2018: 1):

> The exponential growth of social media such as Twitter and community forums has revolutionised communication and content publishing, but is also increasingly exploited for the propagation of hate speech and the organisation of hate based activities. The anonymity and mobility afforded by such media has made the breeding and spread of hate speech – eventually leading to hate crime – effortless in a virtual landscape beyond the realms of traditional law enforcement.

This disturbing trend certainly calls for innovations to curb hate speech, especially hate speech that is propagated through online platforms because of its relative freedom from absolute government control. It is in the light of this that organisations have suggested how to mitigate hate speech both online and offline. The NSRP (2017), for example, suggested that a framework should be established that allows for easy identification of statements that constitute hate speech through a refined automated methodology by creating an accessible platform for regulatory agencies to share analysis actions needed to counter and mitigate hate speech in the short, medium and long terms. The proliferation of hate speech online, observed by the UN Human Rights Council, poses a new set of challenges. The report pointed to the fact that both social networking platforms and organisations created to combat hate speech have recognised that hateful messages disseminated online are increasingly common and have elicited unprecedented attention to develop adequate responses (British Institute of Human Rights, 2012). Similarly, Zhang et al. (2018) pointed out that over the years, the increasing propagation of hate speech on social media and the urgent need for effective mitigation efforts have drawn significant investment from governments, companies and empirical research.

Zhang et al. (2018) noted that there is an increasing pressure for scalable, automated methods to detect hate speech and this has continued to attract research from various perspectives, especially from the natural language processing (NLP) and machine learning (ML) communities. Similarly, in an effort to curb online hate speech, Hopko (2018) noted that researchers in California and the Anti-Defamation League have adopted a computer-based approach by teaching computers to recognise hate speech on social media platforms using artificial intelligence.

2.2. Artificial intelligence

The term artificial intelligence was coined in 1956. Research into AI dates back to the 1950s, which focused on issues tied to areas such as problem solving and symbolic methods. The US Department of Defense took interest in

the development of AI in the 1960s and began training computers to mimic basic human reasoning. It has become a buzz term in recent years due to the increased data volumes, advanced algorithms and improvements in computing power and storage (SAS, n.d). Exploring AI paved the way for the automation and the formal reasoning that are common features of computers designed to complement and augment human abilities (SAS, n.d).

AI is an area of computer science that emphasises the creation of intelligent machines, among them computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages (Techopedia, n.d). AI "makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks. Most AI examples that you hear about today – from chess-playing computers to self-driving cars – rely heavily on deep learning and natural language processing" (SAS n.d.: 1). AI enables computers to be programmed to accomplish specific tasks by processing large amounts of data and recognising patterns in the data and is as enabled with certain traits such as knowledge, reasoning, problem solving, perception, learning, planning and the ability to manipulate and move objects (Techopedia, n.d.; SAS, n.d.).

AI functions by combining large amounts of data with fast, repetitive processing and intelligent algorithms (a step-by-step technique used to get the job done by a computer), allowing the software to learn automatically from features in the data (SAS, n.d.; Howstuffworks, n.d.). SAS (n.d) noted that among other fields, AI is visible in its area of the ability of computers to analyse, understand and generate human language, including speech; process, analyse and understand images; capture images or videos in real time and interpret their surroundings; interpret images and speech – and then speak coherently in response. AI is widely used in question answering systems for risk notification in various fields, legal assistance, patent searches, etc.

AI applications affect almost all fields of human endeavour and are most visible in the industry, banking, insurance, health, media and defence sectors. Several tasks are now automated, transforming many sectors. In the health care sector, for example, AI applications can provide personal health care assistance through life coaching, X-ray readings or reminding patients to take their medicines, do exercise or eat healthier, among others. In the area of marketing, AI provides virtual shopping capabilities that offer personalised recommendations and discuss purchase options with the consumer. The manufacturing sector uses AI to analyse factory Internet of Things data as it streams from connected equipment to forecast expected load and demand using recurrent networks. Smart technologies can also help manufacturers diversify into offering both manufactured products and complementary services. In the sports sector, AI is used to capture images of a game and provide coaches with options on how to better organise the game, including suitable strategy (SAS, n.d.; Ganascia, 2018).

In the same vein, the author pointed out that:

> In 1997, a computer programme defeated the reigning world chess champion, and more recently, in 2016, other computer programmes have beaten the world's best Go (an ancient Chinese board game) players and some top poker players. Computers are proving, or helping to prove, mathematical theorems; knowledge is being automatically constructed from huge masses of data, in terabytes (1012 bytes), or even petabytes (1015 bytes), using machine learning techniques. (Ganascia, 2018: 9)

He further noted that:

> Machines can recognize speech and transcribe it – just like typists did in the past. Computers can accurately identify faces or fingerprints from among tens of millions, or understand texts written in natural languages. Using machine learning techniques, cars drive themselves; machines are better than dermatologists at diagnosing melanomas using photographs of skin moles taken with mobile phone cameras; robots are fighting wars instead of humans and factory production lines are becoming increasingly automated. (Ganascia, 2018: 9)

Similarly, Wolverton (2018a, 2018b) noted that there is increasing fear of how AI will affect society, but at the moment it is a lot better at some things than others. For example, Facebook is increasingly relying on AI to monitor its service and identify content that violates its policies and guidelines. According to Zuckerberg (quoted in Wolverton, 2018b), AI "is having varying degrees of success. Terrorism-related posts from the likes of ISIS and Al Qaeda are easy to police with AI. So too is nudity", but there is still a lot to be done for hate speech. Similarly, Facebook is using AI to proactively detect eight categories of content: nudity, graphic violence, terrorism, hate speech, spam, fake accounts and suicide prevention (Terdiman, 2018).

3. Theoretical Framework

The theoretical framework for this study is the Diffusion of Innovation (DOI) Theory and the Push-ICT Theory. The DOI Theory, developed by E. M. Rogers in 1962 (Rogers, 1995), explains the adoption of a new idea, behaviour or product (i.e. "innovation") in stages and established adopter categories (initiator, early adopter, early majority, late majority and laggards). The introduction of AI in Nigeria would be an innovation and product to address a problem that affects the country. The key to adopting AI as an innovation in mitigating hate speech is that the government must perceive the idea or product as new or innovative. It is through this that diffusion is possible. The way this would be accepted is through the Push-ICT Theory, which holds that in a situation where information and communication technologies are considered important or relevant to development of individuals or a community, such technologies should be deployed by the relevant organisation (government, non-governmental organisations or individuals). In our case, international

organisations, such as the International Telecommunication Union, can ensure the availability of AI software for use by the Nigerian government.

- The relevant technology or services (e.g. training) should be made affordable. It can be deployed free of charge or highly subsidised
- The deploying organisation or individual would clearly identify the workable benefits of the technology and subsequently coerce the individuals or communities to use the deployed technologies or services.
- The push is usually through policy framework, the cheap and affordable deployment of ICT facilities, social status push and ICT-user push.
- Where ICT remains unaffordable (perhaps as a result of poor policy implementation), users also push ICT providers to offer affordable facilities and services.

When there is easy access to or the availability of AI systems, resistance to adopting the use of ICT is highly reduced, while acceptance of use is greatly enhanced and the possibility to use the technology is high (Wilson, 2017).

3.1. AI approach to mitigating hate speech in Nigeria

Mitigating hate speech through AI would require the partnership of the Nigerian government, other stakeholders (i.e. bloggers, online news media and security agencies) and the hosting service providers to swiftly take down hate speech. As pointed out in Starr (2004) for the case of the UK, the Internet Watch Foundation (IWF) and police work in partnership with the hosting service provider to swiftly remove any reported or identified hate speech.

The AI approach depends largely on the availability of software (which is primarily the responsibility of programmers) based on standard statistics with machine learning and algorithms (a popular machine learning technique that relies on training a very large simulated neural network to recognise subtle patterns using a large amount of data) to ensure the following:

1. Automatic Detect and Delete Hate Speech AI system: Online platforms in Nigeria depend on host companies to float their websites. Therefore, owners of such websites should ensure that the Detect and Delete software forms part of the site development to automatically and swiftly delete any word that falls within the categories of hate speech. For example, according to Thomas (2018), "Facebook seems to be internally testing new automatic detection via a 'hate speech button' that briefly allowed users to report hate speech on individual posts before it was removed". Facebook has also trained its AI systems to look for fake accounts using "kinds of signals that would indicate illegitimacy: an account reaching out to many more other accounts than usual; a large volume of activity that seems automated" (Terdiman, 2018) However, the position of this paper is the adoption of

automatic detect and delete programme. This can be applied to mainstream media platforms, especially the comments section. In this line, Thomas (2018) noted that "Stop PropagHate has received funding from the Digital News Initiative (DNI) to use AI to help detect and reduce hate speech in online news media".

2. Automatic Detect and Hide Hate Speech AI system: This system is similar to what Twitter has begun to do on its site. Through AI, hate speech is automatically detected and posts from certain accounts that detract from conversation are hidden (Zhang et al., 2018). This can be applied to all websites to detect hate speeches and hide them swiftly, especially sites used by Nigerians that provide a platform for user comments.
3. Automatic Detect and Block/Remove User AI system: This programme detects hate speech and blocks/removes the user or owner of the site or account when hate speech is recurrent and blocked or hidden daily for a period of time by the user itself or the site.
   According to Thomas (2018), "the European Commissioner for Justice, Consumers and Gender Equality, Věra Jourová, is examining how to have hateful content removed swiftly by social media platforms, with tough legislation being one option that could replace the current system". Thomas (2018) and Nolan (2017) pointed out that Google recently launched an AI tool that identifies abusive comments online and publications such as *The New York Times*, *The Guardian* and *The Economist* are testing the new software as a way of policing comments sections. This measure can also be applied by Nigerian news media.
4. Filter and Review Hate Speech AI system: News organisations worldwide want to encourage interactivity (engagement and discussion) around their news content, but sorting through millions of comments to find those that are trolling or abusive will certainly take a lot of time. The filter and review programme can be used to filter and compile comments on websites and review them for use especially by media organisations. This would help in the swift moderation of comments and decisions on whether to allow or discard such comments.
5. Detect Signs/Emotions and Block Hate Speech AI system: This system is similar to that used by Twitter, which has "developed a new artificial intelligence system that can detect sarcasm in tweets better than humans, an advance that may help computers automatically spot and remove online hate speech and abusive comments" (New AI system can detect sarcasm, 2018). In fact, the system was better than humans at detecting sarcasm and other emotions on Twitter and would also help detect emojis and signs that are sarcastic in nature and delete or block them.

If not completely, these AI-based approaches certainly have the potential to mitigate hate speech in Nigeria to a large extent. Considering that the social media and comments sections of mainstream media are the major out-

lets for hate speech in Nigeria, the present AI system has an enormous potential for mitigating this type of speech in the country. However, certain challenges must be addressed to fine tune the use of AI in mitigating hate speech in Nigeria and beyond.

3.2. Challenges

1. Human rights challenges: The United Nations Charter reaffirmed their faith in fundamental human rights globally, part of which guarantees freedom of speech for all without distinction. Concern has been expressed that the emerging scientific consensus on the censorship of law-abiding content is that it actually amplifies violence and extremism for a variety of different reasons, as well as infringing on people's rights.
2. Subjectivity of hate speech: Taylor (cited in Coles, 2018) noted that "while AI could be trained to identify keywords or phrases and 'flag' them as potentially hateful, it is likely it would still need human intervention to review and make the determination on whether a given word or phrase is hateful," Similarly, Zuckerberg himself noted that "Hate speech is a problem for AI, because it's subject to lots of nuance. Also, because Facebook operates in numerous countries around the world, its AI needs to understand those nuances in multiple languages" (quoted in Wolverton, 2018a).
3. Socio-political factors: These factors have to do with the usual reluctance that comes with accepting change brought by leaders of nations. Ordinarily the Nigerian government can saddle the Nigerian Communication Commission with the responsibility of partnering with relevant hosting service providers to apply the already existing hate speech detectors to Nigeria cyber space activities. However, political will has always been a challenge, especially regarding programme implementation in Nigeria. As a result, important initiatives in the country often suffer neglect, particularly when the initiator is no longer in power.

4. Conclusion

The goal of AI is to provide software or systems that can reason on input and explain on output. AI will enable human-like interactions with software and provide decision support for specific tasks. AI could have a good potential in Nigeria in mitigating hate speech, especially considering the already existing AI that detects some level of inappropriate communication. There are obvious challenges that would mitigate the adoption of AI for the purpose of addressing the issue of hate speech. These challenges can be addressed if proper attention is given to the issue by the Nigerian government, especially if it has the political will to adopt and use AI to mitigate hate speech.

The bulk of the effort lies with the government in creating an enabling environment for AI to thrive in mitigating hate speech. Bringing hosting ser-

vice providers to discuss is an important first step, which should be followed by the development of human and material resources to implement and sustain the AI system.

Bibliographical references

BRITISH INSTITUTE OF HUMAN RIGHTS. (2012). *Mapping study on projects against hate speech online*. Strasbourg: Council of Europe. Retrieved from <https://rm.coe.int/16807023b4>.

BROWN, A. (2017). "What is hate speech?". *Law and Philosophy*, 36, 419-468.

COLES, T. (2018). *Hate speech presents significant challenge for Facebook AI*. ITProToday. Retrieved from <https://www.itprotoday.com/cloud-data-center/hate-speech-presents-significant-challenge-facebook-ai>.

EHIKIOYA, A. (2018). "Presidency laments hate speeches by media houses". *The Nation Newspaper*, 2 February [Online]. Retrieved from <http://thenationonlineng.net/presidency-laments-hate-speeches-media-houses/>.

FALANA, F. (2017). *Nigeria has enough laws to curb hate speeches by Femi Faana*. Sahara Reporters. Retrieved from <http://saharareporters.com/2017/08/26/nigeria-has-enough-laws-curb-hate-speeches-femi-faana>.

GALEON, D. (2017). *Researchers are trying to use AI to put an end to hate speech*. Futurism. Retrieved from <https://futurism.com/researchers-use-ai-end-hate-speech>.

GANASCIA, J.-G. (2018). "Artificial intelligence: Between myth and reality". *The UNESCO Courier*, 3, 7-9.

GAGLIARDONE, I; GAL, D; ALVES, T and MARTINEZ, G. (2015). *Countering Online Hate Speech*. Paris: UNESCO. Retrieved from <http://unesdoc.unesco.org/images/0023/002332/233231e.pdf: UNESCO>.

HOPKO, A (2018). "Can artificial intelligence recognize hate speech? Cal-Berkeley researchers think so". *Cronkite News,* 9 August [Online]. Retrieved from <https://cronkitenews.azpbs.org/2018/08/09/can-artificial-intelligence-recognize-hate-speech/>

HOWSTUFFWORKS (n.d). *What is a computer algorithm?* [Date consulted: 30/08/2018] <https://computer.howstuffworks.com/what-is-a-computer-algorithm.htm>

"NEW AI SYSTEM CAN DETECT SARCASM ON TWITTER BETTER THAN HUMANS" (2017). *The Economic Times*, 8 August [Online]. Retrieved from <https://economictimes.indiatimes.com/magazines/panache/new-ai-system-can-detect-sarcasm-on-twitter-better-than-humans/articleshow/59969985.cms>.

NIGERIA STABILITY AND RECONCILIATION PROGRAMME (NSRP) (2017). *How-to guide. Mitigating dangerous speech. Monitoring and countering dangerous speech to reduce violence.* Retrieved from <http://www.nsrp-nigeria.org/wp-content/uploads/2017/12/NSRP-How-to-Guide-Mitigating-Hate-and-Dangerous-Speech.pdf>.

NOLAN, L. (2017). "Google launches AI program to detect 'hate speech'". *Breitbart*, 23 February [Online]. Retrieved from <https://www.breitbart.com/tech/2017/02/23/google-launches-ai-program-to-detect-hate-speech>.

OLOJA, M. (2018) "Whose hate speech threatens national unity?". *The Guardian*, 22 July [Online]. Retrieved from <https://guardian.ng/opinion/whose-hate-speech-threatens-national-unity/>.

ONANUGA, B. (2018). "Roots of hate speech, Remedies". Paper presented at the *Workshop on Hate Communication in Nigeria: Identifying Its Roots and Remedies*, 22 February. Abuja: Nigerian Press Council.

ONYIBE, M. (2017). "Mr President: There must be better ways to curb hate speeches. *Vanguard*, 16 September [Online]. Retrieved from <https://www.vanguardngr.com/2017/09/mr-president-must-better-ways-curb-hate-speeches>.

OPUSUNJU. O. (2017). "Nigerian government begins monitoring social media to tame hate speech". *ITEdgeNews*, 26 January [Online]. Retrieved from <https://itedgenews.ng/2018/01/26/nigerian-government-begins-monitoring-social-media-tame-hate-speech>.

PALFREY, J. (2018). *Safe spaces, brave spaces diversity and free expression in education*. Cambridge: MIT Press.

RING, C. E. (2013). *Hate speech in social media: An exploration of the problem and its proposed solutions*, PhD thesis, University of Colorado at Boulder. Retrieved from <https://scholar.colorado.edu/jour_gradetds/15>.

ROGERS, E. (1995). *Diffusion of innovations*. New York: Free Press. 5th edition.

SAS (n.d.) *Artificial Intelligence: What it is and why it matters*. [Date consulted: 30/08/2018]. Retrieved from <https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html>.

STARR, S. (2004). "Understanding hate speech". In: C. MÖLLER, A. AMOUROUX (eds.). *The Media Freedom Internet Cookbook*. Vienna: OSCE, 125-160.

TECHOPEDIA. (n.d). *Artificial Intelligence (AI)*. [Date consulted: 30/08/2018]. Retrieved from <https://www.techopedia.com/definition/190/artificial-intelligence-ai>.

TERDIMAN, D. (2018). *Here's how Facebook uses AI to detect many kinds of bad content*. Fastcompany. Retrieved from <https://www.fastcompany.com/40566786/heres-how-facebook-uses-ai-to-detect-many-kinds-of-bad-content>.

THOMAS, A. (2018). *EJC joins forces with DataScouting to counter hate speech directed at journalists online*. Medium. Retrieved from <https://medium.com/we-are-the-european-journalism-centre/ejc-joins-forces-with-datascouting-to-counter-hate-speech-directed-at-journalists-online-60ed1c857a17>.

UTOMI, J.-M. (2018). "The controversial hate speech bill". *The Sun News*, 8 March [Online]. Retrieved from <http://sunnewsonline.com/the-controversial-hate-speech-bill>.

WALDRON, J. (2012). *The harm in hate speech.* Cambridge, MA and London: Harvard University Press.

WILSON, J. (2017). "Overcoming technophobia in communication education: The push-ICT approach". *Media and Communication / Mediji i komunikacije*, 1 (7), 19-32.

WOLVERTON, T. (2018a). "Mark Zuckerberg says AI won't be able to reliably detect hate speech for 'five to 10' years (FB)", *Pulse*, 10 April [Online]. Retrieved from <https://www.pulse.ng/bi/tech/tech-mark-zuckerberg-says-ai-wont-be-able-to-reliably-detect-hate-speech-for-five-to/fq74l83>.

WOLVERTON, T. (2018b). "AI is great at recognizing nipples, Mark Zuckerberg says (FB)". *Pulse*, 25 April [Online]. Retrieved from <https://www.pulse.ng/bi/tech/ai-is-great-at-recognizing-nipples-mark-zuckerberg-says-fb-id8304794.html>.

ZHANG, Z.; ROBINSON, D. and TEPPER J. (2018). "Detecting hate speech on Twitter using a convolution-GRU based deep neural network". In: A. GANGEMI et al. (eds.). *The Semantic Web. ESWC 2018 Satellite Events. 15th Extended Semantic Web Conference Heraklion, Crete, Greece, 3-7 June 2018*. Cham: Springer, 745-760.