2nd National Level Students' Research Conference on
"Innovative Ideas and Invention with Sustainability in Computer Science and IT-2021"
In association with International Journal of Scientific Research in Computer Science, Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Regulating the Usage of Social Media Using Artificial Intelligence

Mr. Harshal Ogale, Mr. Anand Varrier, Mr. Pranay Ranjan, Prof. Swapnil Goje, Mr. Anuj Phalke
School of Comp. Sci., MIT-World Peace University, Pune, Maharashtra, India

ABSTRACT

Due to the advent of Social Media in the last decade humankind has seen a lot of technological advancements which has made life better. Social media started with the intention of connecting people around the world within seconds. An update of a discovery made by the researchers far in the Antarctic continent or the latest happenings of a protest in Myanmar is just a click away from us. But at the same time the amount of hatred and negativity spread through social media is a matter of great concern. Our research paper is inspired by a documentary 'The Social Dilemma' which opened the eyes of the public by revealing how the social media business is run around the world. The paper is mainly concerned with Online Content Moderation with the help of Artificial Intelligence tools which will help in curbing Cyber bullying and Fake News along with the need for Artificial Intelligence in India's National Security. The paper also covers how to reduce the screen time of the users.

Keywords - The Social Dilemma, Artificial Intelligence, Online Content Moderation, Cyber bullying, Fake News.

I. INTRODUCTION

The Internet has become an integral part of human society in the past few decades. It has impacted organizations across all sectors thus positively impacting the economy of a country. With time there has been an exponential growth in the number of people benefiting from the internet which indicates huge amounts of data being stored and transferred around the globe every second. An extensive study by Cisco describes the growth of users and traffic around the globe in their Annual Internet Report.

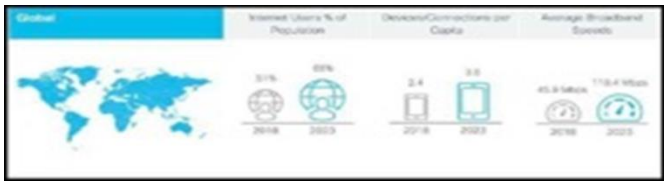

Fig 1: Cisco Annual Internet White Paper 2020

However, as the number of users are increasing every year there has been a potential concern over a negative influence of the data being transferred. This has triggered a wider discussion among the policy makers, company stakeholders and the government. The negative influence mainly is spread by the harmful content over the internet which includes child abuse content.

content which promotes terrorist activities, violent or illegal or offensive material concerning any cultural or religious beliefs. The fact is determining or classifying such content by the machine is challenging. Such scenarios have given rise to a growing awareness within the internet companies of their responsibilities to filter such content before it goes over the internet. This will not only protect the privacy of the users but also help in filtering undesirable content. In the present day, classifying and determining harmful

content is labour intensive in many internet companies which is expensive and time consuming but with AI in perspective algorithms or applications must be developed to scrutinize such content before it goes public.

II. LITERATURE REVIEW

The authors have observed that in the existing internet world the data to be it in any form is not filtered or moderated in such a manner that they still hurt or offend or have a negative influence per say in the society. Social media plays a vital role in influencing the mass on taking an action. Spreading of fake news which leads to protest on the streets is one of the major drawbacks of the influence of social media firms like Facbeook, Instagram, Twitter etc. have started taking the responsibility to moderate but despite all these efforts the problem still exists. Therefore, in this research paper the authors have found that there is an increasse in the need of moderation on the social media content being shared. In this paper the authors have tried to do an extensive research on various ways in which the content can be moderated and resolve the mentioned problems.

III. NEED FOR ONLINE MODERATION

A governance system also known as moderation or filtration is required over the content sharing platforms to prevent abuse, aggression or content that the users prefer to avoid. The practice of Moderation on websites like Wikipedia would not be beneficial as it is an open encyclopedia which gives information on any topic but on the other hand moderation on websites like Facebook, Instagram, Twitter etc. is necessary because these are websites where harmful content is shared or posted deliberately in the form of advertisements, memes or just general posts. The two examples are not necessarily a conflict but are a matter of concern when it comes to companies to filter what actually matters and what content if removed won't make a big difference.

Perhaps, Keep moderation is necessary to mitigate harm and support a pro-social behaviour in the society which ultimately motivates economic goals for such companies. Online Content Moderation must be considered essential to avert Cyberbullying and spreading of Fake News.

IV. SCREEN TIME AND ITS IMPACT

A screen could be of any device be it a TV, mobile phone or a tablet. In this era of technology these devices have governed our lives and without these devices our world would turn upside down. Screen time defines the number of hours or minutes we spend in front of the screen of any device for work or for entertainment purposes. Using these devices in our day to life is gradually causing health related issues, family relations and what not. According to a study at the Harvard Medical Digital devices interfere with everything from sleep to creativity. The study shows that children are the most affected by the increase in screen time. They have observed that a good night's sleep is essential to any human to work in their full potential. The study concludes that we as humans need to be more flexible with the change in technology and develop a habit of using the devices in the right manner.

Fig 2: Google - Ideal Screen Time

V. AI SPECIFIC CHALLENGES IN ONLINE CONTENT MODERATION

Online Content Moderation using AI has its own set of challenges and complexities. It is generally expected to deliver higher standards of results as compared to humans. In addition to this there exists multiple techniques ranging from random forest and Markov Models to support vector machines and neural networks. Neural Networks and Deep learning are inherently unexplainable as they replicate how the human brain learns such that even the AI developers find it difficult to comprehend. Deploying AI in filtering has its own pros and cons as every model or technique gives unique results and there are times that developers themselves do not understand why a particular output was given. The difficulty is illustrated in a test case scenario where the machine was given to differentiate huskies from wolves. The study showed that the algorithm was seemingly accurate at identifying the two animals but a deeper analysis revealed that it was simply learning that photographs of wolves are clicked in snowy regions. Due to such complexity of Neural networks and Deep learning where there are millions of interconnected neurons it is extremely difficult to come up with an explainable result set.

However, incorporating an explainable data set has been an ongoing research itself. Neural Saliency Techniques is considered to increase the explain ability of image and video moderation. These attention mechanisms can identify regions and features in an image which was highlighted during classification. Generally measuring the moderation of any online content is difficult but if powered by AI there could be substantial results. Nevertheless, AI has the potential to have a significant impact on Online Content Moderation in 3 ways.

1) Advanced AI based algorithms can be used to increase pre-moderation stage to improve the accuracy.
2) AI can be used to analyze the training data to improve pre-moderation performance.
3) AI can augment human moderators to increase productivity and to reduce the harmful effects of content moderation.

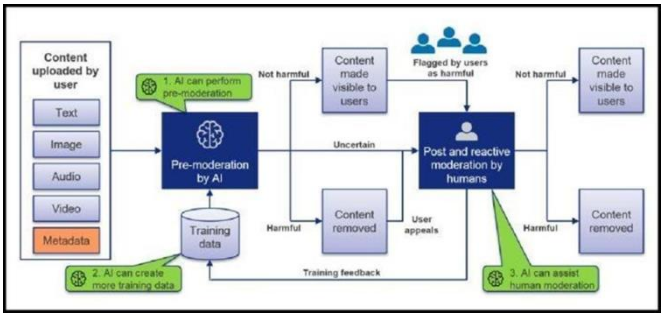

Fig 3: Cambridge Consultants - workflow of Content Moderation

VI. TYPES OF FILTER ALGORITHMS AND ITS WORKING

The techniques used to filter content differ depending on the media to be analyzed. A filter can work at different levels of complexity, spanning from simply comparing contents against a blacklist, to more sophisticated techniques employing complex AI techniques.

A. Machine Learning (ML)

AI deployed systems in moderation generally use Machine Learning (ML) which is becoming more dominant in recent times. They adopt the following 3 approaches:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning.

All 3 mentioned ways have their own pros and cons but with the help of all 3 approaches the machine can be trained to filter the content.

B. Neural Networks

Neural networks consist of a set of nodes called neurons which are laid out in multiple layers and interconnected. They depict the biological neurons which transmit information. In case of filtering, the input may represent any point on an image or a text. The training of the network starts from telling the network whether the output is right or wrong based on which the algorithm is updated by making few changes in the calculations and interconnection of neurons.

C. Metadata Filtering

Metadata filtering uses the information (metadata) that accompanies another set of data, providing information about that item. Typical examples of metadata are as follows: book's title, author and publisher; a song's title, performer, and length; a movie's title, performers; However, it is considered that metadata filtering of unwanted content is often inaccurate and easily manipulated to avoid detection.

D. Hash Function Algorithm

A hash function/hashing method takes a file as input and generates a small record that is uniquely linked to the file. For a different file every time a different hash would be produced. Major drawback of Hash based filtering is that if there is any change in a data item like changing the file format, compressing it, deleting a few words from a text or even shortening a song by second will produce an entirely different hash. Hash-based filtering can only identify exact matches of unwanted files.

E. Blacklisting

The blacklist approach is a text-based filtering technique which involves creating and maintaining a dataset of unwanted textual content. Incoming texts are then compared with the dataset to spot similarities and then they are either rejected, deleted, or flagged. Blacklisting approaches can be hacked by misspelling the undesired words or combining texts with graphics, such as emoji's.

F. Natural Language Processing

Natural Language Processing (NLP) is the subfield of computer science which studies how to equip computer systems to handle the language naturally spoken by humans. In textual filtering, natural language processing is needed whenever the simple occurrence of certain word patterns is insufficient to classify the relevant textual items as needed. Just like Neural Networks, Natural Language Processing is complex to understand and comprehend.

VII. PRECECISION OF FILTER ALGORITHMS

The evaluation of Content Filtering Systems on standard metrics is in order to determine their accuracy. Filtering can be viewed as a binary classification task whose purpose is to assess whether a given item belongs to a positive class or to a negative class (e.g., the message is harmful or non-harmful, infringing IP or not). A filtering system's positive or negative answer can be evaluated as follows: 1) true positive (TP): Item is classified as harmful. 2)true negative (TN): Item is classified as non-harmful. 3)false positive (FP): Item is classified as harmful but is not. 4)false negative (FN): Item is classified as non-harmful but is harmful.

VIII.NEED FOR AI IN NATIONAL SECURITY - THE INDIA PERSPECTIVE

As we know recently in a controversial move, the cybercrime cell of the Ministry of Home Affairs (MHA) has started a new programme under which: - Citizens can participate as volunteers to identify, flag and report to the Government Illegal and unlawful content. Which includes terrorism, radicalisation, rapes, unlawful and outrageous sexual content and anti-national activities. Under this programme, the

MHA's INDIAN CYBER CELL COORDINATION CENTRE (I4C) acts as a nodal point. It has been blatantly mentioned that the personal details of these volunteers will be kept confidential. This move is welcome especially after recent soar in use of social media to beget anti- national sentiments amongst masses and to exhort turmoil in our peace loving country. Social media is also being used to circulate deceptive knowledge to influence masses.

Currently the components of the I4C scheme includes:

1) National Cyber Crime Threat Analytics Unit.
2) National Cybercrime Unit.
3) Platform for Joint Cybercrime Investigation Team.
4) National Cybercrime Forensic Laboratory Ecosystem.
5) National Cyber Training Center.
6) Cybercrime Ecosystem Management Unit.

However there is a myriad of shortcomings that is still to be addressed.

1. There is no legal definition of anti-national content or activity, either by the government or the judiciary.
2. Giving people the option to report fellow citizens gives too much power without adequate checks and balances.
3. Government must impute a fair share in the national budget to development of and research in application of Artificial Intelligence to effectively counter Cybercrime.

IX. AI AND COUNTERTERRORISM

Terrorism has its root spread throughout human history. However, with time, the channels to harness terror in the socialized world have been undergoing radical renaissance. The madmen with dreadful visions have left no stone unturned to perpetrate horrors on the civilized world. If the USA has it's 9/11 then India has its own 26/11 that constantly reminds us that perpetual and impregnable security of our borders via land, sea and air is indispensable. Especially for a country like India, which shares 3,323 kilometers of its border with a hostile nation "PAKSTAN" considered as "Cradle of Global Terrorism". Extensive research in Artificial Intelligence, Deep Learning and Robotics has allowed and will continue to allow new capabilities that will improve military strategies assertively. Implementation of autonomous weapons and surveillance systems based on AI and robotics will not only improve accuracy of results but will also save precious human lives. Major global superpowers like the USA and China continue aggressively, to compete in the sphere of Artificial Intelligence. In India, as of now, D.R.D.O. has a laboratory specifically dedicated to artificial intelligence called CENTER FOR ARTIFICIAL INTELLIGENCE AND ROBOTICS.

X. COLLECTION AND ANALYSIS OF DATA

There is a myriad of surveillance equipment like the security cameras and satellites capturing an infinite array of photos and videos every day. India has perpetual surveillance over its massive land boundary of 15,200 kilometres and coastline of 7,516.6 kilometres. This results in an overflow of data which is too extensive for any human force to analyse and determine suspicious or hostile activities. For intelligence agencies it creates both a challenge and an opportunity. Computer-assisted intelligence analysis, leveraging machine learning, could be a game changer in such a scenario. Automated analysis based on machine learning has already given accurate results in variegated experiments. In 2015, image recognition systems developed by Microsoft and Google outperformed human competitors. These machine- learning based techniques are already being adopted by U.S. intelligence agencies to Automatically analyse satellite reconnaissance photographs.

Research and implementation of analogous technologies by Indian Intelligence on its massive reconnaissance archives could be a game changer. It would enable us to keep impregnable eyes on our 15,200 kilometres land boundary and 7,516.6 kilometres of coastline everyday 24/7.

XI. DIGITAL DATABSE FOR AI IMPLEMENTATION

India has established a wide range of surveillance and data gathering advanced tools. And, through database-centric schemes like National Intelligence Grid (NATGRID), Network Traffic Analysis System (NETRA) and the Crime and Criminal Tracking Network & Systems (CCTNS), law enforcement agencies have achieved centrally a
Lawful Intercept and Monitoring (LIM) system. Post 2008 Mumbai terror attack, these mechanisms got a big boost. However, India must keep progressing and analysing to maintain a state-of-the-art infrastructure, knowledge as well as proficient work force.

XII. CONCLUSION

The above-mentioned approaches are currently being used in filtering out all the unwanted content from the internet by few Social Media Giants but not many. The time taken to train the machine and then to re-correct the algorithm is a time-consuming process but with time or over a few years these approaches if used by the social media firms could make a difference in the society. This
will not only reduce crimes but also have an impact over the health of an individual. Every method is complex in its own way but if the right algorithms are used the moderation of online content is possible. For all this to happen the world needs volunteers who will willingly work on such problems and come up with more efficient ways of filtering the data.

XIII. REFERENCES

[1]. https://emerj.com/ai-sector-overviews/ai-social-media- censorship-works-facebook-youtube-twitter/
[2]. https://hms.harvard.edu/news/screen-time-brain.
[3]. https://www.rallyhealth.com/health/unexpected-effects-screentime
[4]. https://www.ofcom.org.uk/data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf
[5]. https://www.researchgate.net/publication/328838624_No_More_FOMO_Limiting_Social_Media_Decreases_Loneliness_and_D epressio