Can Artificial Intelligence reduce cyberbullying?

This Literature review is based on a topic that I have a personal interest in. As a parent of two children under the age of 6, I have seen firsthand the impact of screentime - our digital footprint has grown over the years. Over the past few years strict restrictions were put in place due to the Covid-19 pandemic and we began working and studying from home and using devices to connect to work or to the classrooms. Although these restrictions have now being lifted, one thing that has remained constant is the use of devices over the internet being used daily in the classroom and in the home environment. It is through these devices that cyberbullying occurs.

The definition of cyberbullying is the act of one or more individuals intimidating another by using digital communication such as social media platforms, instant messaging, online forums, or text messages. It entails the use of technology and the internet to target someone to cause harm. (Smith 2015, Englander et al, 2017 and Leung et al. 2023) The term cyberbullying was introduced in the late 90's early 2000's and overlaps with traditional bullying. However, one of the main differences between cyberbullying and in person bullying is that the victim is unable to hide away from the attacker in a way that may be possible with in person bullying. The effects of cyberbullying were more severe as it led to the rise of psychological harm, substance abuse and depression.   (Englander 2017 and Li, 2010)

Online bullying are categorised as follows: -

- Flaming – Sending angry, harsh, or profane remarks to a person or people privately or through an internet group.
- Harassment – Sending offensive messages to someone repeatedly.

- Denigration – Posting rude, untrue statements about an individual.
- Masquerading – Pretending to be someone else and posting something that discredits an individual or puts them in danger.
- Outing and trickery - Posting material about a person that contains sensitive, private, or embarrassing information.
- Social exclusion - Intentionally exclude a person from an online group.

Social media such as Twitter Facebook, TikTok, WhatsApp and Instagram usage has increased over the years for individuals between the ages of 10 – 30. Due to this cyberbullying has increased specially during the pandemic. They also highlight that traditional upbringing cannot monitor the behavioural patterns from these platforms. (Martinez, 2023 and Hamiza Wan Ali, Mohd and Fauzi, 2018))

To give some context on the number numbers of active users that each platform Facebook has the highest active users reaching 2.94 billion as of 2022. YouTube has a total of 2.48 billion users. WhatsApp with 2 billion users and Instagram has 1.44 billion users worldwide. Of course, not all these users will experience cyberbullying but given the large number of users the likelihood of some form of bullying being experienced by individual users is quite high. Once an individual experiences cyberbullying the likelihood of developing depression and social anxiety also increases (Selfhout et al., 2008 -Noori, Sayes and Anwari, 2023).

The question then becomes, can Artificial Intelligence or Machine Learning help detect and reduce to incidence of cyberbullying as well as ensure that social media platforms are used appropriately and for the purpose they were intended for.

Artificial intelligence simulates human intelligence. Running algorithms and methodologies on data, which are labelled and classified (Panch, Szolovits and Atun, 2018 and Shakeel and Dwivedi, 2022)

Artificial Intelligence are categorised into the following:

- Machine Learning - Where algorithms learn from examples of data.
- Deep Learning - Allows for large quantities of raw data to discover representation necessary for detection or classification.
- Supervised Learning – A technique in which an algorithm learns from labelled training data to make predictions or decisions. The data consists of input features and their corresponding labels or target values.
- Unsupervised Learning – A technique that allows algorithm to learn patterns or structures in the data without any explicit supervision or labelled examples.
- Reinforcement Learning – The focuses on how an agent can learn to make sequential decisions in an environment.

This can be broken down further into Classification models, which are machine learning models used to predict the class or category of a given input based on its attributes. They are a subcategory of supervised learning algorithms, where the training data consists of labelled examples with known classes. The goal of a classification model is to learn a set of rules that can accurately assign new, unseen instances to their respective classes. There are various types of classification models, the most common ones are as follows:

- Logistic regression - A model that determines the probability of an instance belonging to a particular class. It works well for binary classification.

- Decision trees – A hierarchical structures that recursively partition the feature space based on the values of different features. Each internal node represents a decision based on a feature, and each leaf node represents a class assignment.
- Random Forest - Creates a collection of decision trees trained on different subsets of the data and features and aggregates their predictions.
- Support Vector Machines (SVM) A binary classification model that finds an optimal hyperplane to separate instances of different classes
- K-Nearest Neighbours (KNN) - KNN classifies instances based on their proximity to the labelled instances in the training data..
- Naive Bayes is a probabilistic classification model based on Bayes' theorem.
- Neural networks, particularly deep learning models, have gained popularity in classification tasks. They consist of multiple interconnected layers of artificial neurons that can learn complex patterns and relationships from the data. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) image classification and natural language processing.

(Hamiza Wan Ali, Mohd and Fauzi, 2018 -Khurram, 2023 -Panch, Szolovits and Atun, 2018 Sandya Venu et al., 2023)

For any of these to be successful they need to identify bullying actions. Bullying actions could consist of social exclusion, sharing private information, or direct abuse. This entails collecting data, which could include:

- Profane words that would be used describe an ethnic group or labelling someone by their sexual orientation or gender identity.
- Phrases that individuals may found offensive such as labelling a person by race, religion, or sexual preference.
- Collecting images combined with words that maybe target an individual or group of people an example of this would be a grave with a caption 'DIE'

Once the data is obtained it then needs to be cleaned to ensure there is no missing values or ambiguity and has a suitable format for training (Salawu, He and Limsden, 2017 Ea et al., 2023 Islam, 2023).

Experimental results would normally take two forms of classification and use the data to run the algorithms. Various studies have shown Support Vector Machines (SVM) have a higher success rate with more accuracy compared to other Classification models. (Ea et al., 2023 -Dalvi, Baliram Chavan and Halbe, 2020 -Hamiza Wan Ali, Mohd and Fauzi, 2018).

One of the most interesting points to highlight is that the majority of research and classification models have only focused on words from the English language; however, global platforms such as Twitter or Facebook are typically used in the language an individual feels most comfortable communicating in or the country they reside in.

To add complexity, many individuals utilise multilingual words made up from two languages, some examples of this may be combining Chinese and English words to make "Chinglish" or combining Hindi and English words to make "Hinglish". Chinglish for example, is frequently the result of a direct translation from Chinese to English that fails to account for differences in grammar, syntax, or cultural background between the two languages. The same is with Hinglish, it incorporates elements of both languages, combining Hindi vocabulary and grammar with English words and phrases.

This can result in awkward or incorrect phrases that may be difficult for native English speakers to understand. Sometimes, some thought and process of elimination you can understand what is being communicated. It is important to note that these are not a reflection of the English skills or intelligence of the speaker or writer but rather a result of the complexities of language translation and cultural differences. If humans have issues understanding them fully, Machine Learning will also struggle to identify them. Research has been conducted not only on the translation of these words but also on text in languages such as Arabic, Urdu, Bangla, and Hindi, but there seem to be a limitation on other countries such a Kenya where they speak Swahili and English and known as ‘Swanglish’ or ‘sheng’ and ‘Franglais’ which is a combination of French and English. It is important to note that words and phrases for these and native languages are constantly changing, and new words are being added modified and abbreviated into slang. For humans and machines to understand they would need to keep learning and understand the new phases, words and meaning, then add them into context of cyberbullying.

(Bhowmik Shanto, Jahirul Islam and Abdus Samad, 2023 -Singla et al., 2023 - Raj et al., 2022)

In conclusion research within Artificial Intelligence and Machine Learning have made great strides forward with positive results. It has been refreshing to see so many studies have been conducted. With the large number of users on the platforms its important that the models can analyse large amount of data such as words and images and in some cases both. This will assist with real time detection of cyber bullying and allow for intervention before the victim is impacted. With a rigorous rule base the human element such a moderator workload would ease, and any emotions and bias views the moderator would have should be removed from the equation. With cloud base technology and the ability to increase Central Processing Unit (CPU) can be achieved with ease so algorithms can process huge amount of date and with continues improvement, adjusting and data the algorithms can become more effective in reducing cyberbullying.

However with every positive, there is a negative and currently algorithms may struggle to understand the complexities of language especially when there is a mix of English and another native language. There seems to be a lack of data to support the algorithms from determining the context in which cyberbullying happens and differentiate between sarcasm, irony, or different types of bullying can be difficult. In the research reviews there has been a subset of data that has been used. With over a billion active users on the top three platforms it is hard to know how many tweets, posts or messages each platform produces and if organisations are willing to pay for the extra process power and working hours needed to collect the data and process it. Organisation’s such a Facebook and Twitter may have their own in-house team to ensure cyberbullying is detected and dealt with appropriately; however, new apps that are starting up my not be able to afford the overhead or plan for cyberbullying on the offset.

In the ever-evolving society we live in cyberbullying techniques and vocabulary are continuously changing, with new platforms being developed training the models on historical data may be unable to detect new kinds of cyberbullying until they are fed new labelled data. This could potentially lead to more false positives and false negatives and failing to detect actual instances of cyberbullying and example of this could be symbol of the swastika a Hindu symbol which means to welcome could misrepresented for a Nazi Hakenkreuz symbol. An important factor to also consider is that models can be influenced by biases in training data. If the training data is biased or has insufficient representation of particular groups, the model's predictions may be prejudiced as well. Detecting cyberbullying would also mean that the data of individuals must be studied. The data being used could be of personal information and falls into the data privacy laws such as the General Data Protection Regulation (GDPR)

There has been some great research performed on the topic and although some models do not have 100% accuracy it feels as if we are at a turning point and this technology can be effective in ensuring that cyberbullying is reduced.

Reference list:

Bauman, S. (2007). Cyberbullying: a Virtual Menace. Available from:: https://www.researchgate.net/profile/Sheri-Bauman/publication/265937264_Cyberbullying_a_Virtual_Menace/links/553e25b10cf2522f1835efc3/Cyberbullying-a-Virtual-Menace.pdf [Accessed 22 Sep. 2023].

Bhowmik Shanto, S., Jahirul Islam, M. and Abdus Samad, Md. (2023). Cyberbullying Detection using Deep Learning Techniques on Bangla Facebook Comments. ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/abstract/document/10083690 [Accessed 24 Sep. 2023].

Dalvi, R.R., Baliram Chavan, S. and Halbe, A. (2020). *Detecting A Twitter Cyberbullying Using Machine Learning*. IEEE Xplore. Available from: doi:https://doi.org/10.1109/ICICCS48265.2020.9120893. [Accessed 24 Sep. 2023].

Ea, P., Vidart, P., Salem, O. and Mehaoua, A. (2023). Cyberbullying Messages Detection: A Comparative Study of Machine Learning Algorithms. ieeexplore.ieee.org. Available from: https://ieeexplore.ieee.org/abstract/document/10223394 [Accessed 23 Sep. 2023].

Englander, E., Donnerstein, E., Kowalski, R., Lin, C.A. and Parti, K. (2017). Defining Cyberbullying. *Pediatrics*, 140(Supplement 2), pp.S148–S151. Available from: doi:https://doi.org/10.1542/peds.2016-1758u. [Accessed 23 Sep. 2023].

Hamiza Wan Ali, W.N., Mohd, M. and Fauzi, F. (2018). Cyberbullying Detection: An Overview. *2018 Cyber Resilience Conference (CRC)*. Available from: doi:https://doi.org/10.1109/cr.2018.8626869. [Accessed 23 Sep. 2023].

Islam, M. (2023). Detection of Cyberbullying in Social Media Texts using Explainable Artificial Intelligence. Available from: https://qspace.library.queensu.ca/server/api/core/bitstreams/d5b6be95-b5d5-47b4-b87b-13447d0d1546/content. [Accessed 23 Sep. 2023].

Jaber ALjohani, E., M.S Yafooz, W. and Alsaeedi, A. (2023). Cyberbullying Detection Approaches: A Review. ieeexplore.ieee.org. Available from: https://ieeexplore.ieee.org/abstract/document/10220688 [Accessed 23 Sep. 2023].

Karmakar, S. and Das, S. (2021). Understanding the Rise of Twitter-Based Cyberbullying Due to COVID-19 through Comprehensive Statistical Evaluation. papers.ssrn.com. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3768839.

Khurram, A. (2023). Anomaly Detection in Electric Power Systems using Machine Learning Methods. etheses.whiterose.ac.uk. Available from: https://etheses.whiterose.ac.uk/33439/ [Accessed 23 Sep. 2023].

Leung, A.N.M., Chan, K.K.S., Ng, C.S.M. and Lee, J.C.-K. (2023a). Cyberbullying and Values Education: Implications for Family and School Education. Google Books. Taylor & Francis. Available from: https://books.google.com.hk/books?hl=en&lr=&id=OVrXEAAAQBAJ&oi=fnd&pg=PT155&dq=what+is+cyberbullying&ots=xsul8upioD&sig=D9pAjJedVwBW8J6NfPdYlL9CZKU&redir_esc=y#v=onepage&q=what%20is%20cyberbullying&f=false [Accessed 22 Sep. 2023].

Leung, A.N.M., Chan, K.K.S., Ng, C.S.M. and Lee, J.C.-K. (2023b). Cyberbullying and Values Education: Implications for Family and School Education. Google Books. Taylor & Francis. Available from: https://books.google.com.hk/books?hl=en&lr=&id=OVrXEAAAQBAJ&oi=fnd&pg=PT155&dq=what+is+cyberbullying&ots=xsul8upioD&sig=D9pAjJedVwBW8J6NfPdYlL9CZKU&redir_esc=y#v=onepage&q=what%20is%20cyberbullying&f=false.

Li, Q. (2010). Cyberbullying in High Schools: a Study of Students' Behaviors and Beliefs about This New Phenomenon. Journal of Aggression, Maltreatment & Trauma, 19(4), pp.372–392. Available from: doi:https://doi.org/10.1080/10926771003788979. [Accessed 22 Sep. 2023].

Martinez, L. (2023). *The Impact of Social Media Usage Among Teens During COVID-19*. [online] www.acquaintpublications.com. Available from: https://www.acquaintpublications.com/article/the-impact-of-social-media-usage-among-teens-during-covid-19 [Accessed 23 Sep. 2023].

Noori, N., Sayes, A. and Anwari, G. (2023). The Negative Impact of Social Media on Youth's Social Lives. International Journal of Humanities Education and Social Sciences, 3(1). Available from: doi:https://doi.org/10.55227/ijhess.v3i1.613. [Accessed 22 Sep. 2023].

Panch, T. Szolovits, P. and Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of Global Health, 8(2). Available from: doi:https://doi.org/10.7189/jogh.08.020303. [Accessed 22 Sep. 2023].

Pandey, C. (2023). Detection of Cyberbullying and Abusive Language on Social Media Using Supervised ML & NLP Techniques. International Journal of Mechanical Engineering, Vol.7 No. 1 January 2022. Available from: doi:https://doi.org/10.56452/7-788. [Accessed 22 Sep. 2023].

Raj, M., Singh, S., Solanki, K. and Selvanambi, R. (2022). An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. SN Computer Science, 3(5). Available from:: https doi:https://doi.org/10.1007/s42979-022-01308-5. [Accessed 23 Sep. 2023].

Salawu, S., He, Y. and Limsden, J. (2017). Approaches to Automated Detection of Cyberbullying: A Survey. ieeexplore.ieee.org. Available from: https://ieeexplore.ieee.org/document/8063898 [Accessed 23 Sep. 2023].

Sandya Venu, V., Shanmugasundaram, H., Reddy Seelam, M., Vardhan Reddy Kotha, V., Snehith Rayudu Muthyala, S. and Kansal, S. (2023). Detection of Cyberbullying on User Tweets and Wikipedia Text using Machine Learning. [online] ieeexplore.ieee.org. Available from: https://ieeexplore.ieee.org/abstract/document/10140252 [Accessed 23 Sep. 2023].

Selfhout, M.H.W., Branje, S.J.T., Delsing, M., ter Bogt, T.F.M. and Meeus, W.H.J. (2008). Different types of Internet use, depression, and social anxiety: The role of perceived friendship quality. Journal of Adolescence, 32(4), pp.819–833. doi:https://doi.org/10.1016/j.adolescence.2008.10.011. [Accessed 23 Sep. 2023].

Shakeel, N. and Dwivedi, R.K. (2022). A Survey on Detection of Cyberbullying in Social Media Using Machine Learning Techniques. Intelligent Communication

Technologies and Virtual Mobile Networks, pp.323–340. Available from: https doi:https://doi.org/10.1007/978-981-19-1844-5_25. [Accessed 23 Sep. 2023].

Singla, S., Lal, R., Sharma, K., Solanki, A. and Kumar, J. (2023). Machine Learning Techniques to Detect Cyber-Bullying. ieeexplore.ieee.org. Available from: https://ieeexplore.ieee.org/abstract/document/10220908 [Accessed 23 Sep. 2023].

Smith, P.K. (2015). The nature of cyberbullying and what we can do about it. Journal of Research in Special Educational Needs, 15(3), pp.176–184. Available from: doi:https://doi.org/10.1111/1471-3802.12114. [Accessed 23 Sep. 2023].