Hello and welcome to my research proposal.

Agenda

Today’s agenda will be as follows

Project Title

Research Topic

Methodology

Cyberbullying

Artificial Intelligence

Chatbot

Ethical and Risk Consideration

Artefacts

Timeline

Future Work

My Project Title will be

Can Machine Learning Reduce and Educate users on cyberbullying.

As a parent of two children, I have noticed that our digital footprint as a family has grown over the years. Restrictions were put in place due to the Covid-19 pandemic and working and studying at home became the norm for many people. Although life is slowly returning to normal, one thing that has remained consistent is our digital footprint, and this is only going to continue to grow. As a parent, this is concerning for many reasons – an important one being cyberbullying. With social media sites, video gaming and instant chats on the rise, interaction online has become more common than ever. There are many children that don’t understand the consequences of their actions and follow along with the crowd. There has been much research performed on Artificial Intelligence, Machine Learning and Chat bots. However; there seems to be limitations with the two technologies when addressing cyberbulling. This research proposal intends to merge these technologies together. The idea is a standalone micro service can be incorporated to either a front-end or back-end application such as an instant messaging application or social media platform. Allowing for the following to be achieved:

- Identify cyberbullying messages or images before they are received by a user.
- Warn and educate the user who is sending the message of the impact such messages would have.
- Alerting if a repeat offender continues with these profane messages, this is then escalated to parents and guardians of both parties and the moderator of the application.

The end result is to reduce cyberbullying and to increase awareness of cyberbullying.

Methodology

The experiment will take the form of quantitative research. Data from previous research studies such as hate speech and tweets from social media sites and offensive language will be utilised. The data will be cleaned and formated to ensure our models are able to learn from this data. This will then be replayed on a number of different models to find which has the most effective successes rate. This will allow for further analysis and allow us to compare the results to previous studies and see what the success rates are.

(Waseem and Hovy, 2016 - Davidson et al., 2017 -www.kaggle.com, n.d.)

The reason quantitative research has been selected is that it allows for a large data set to be used and to be repeated on several algorithms. We are also able to compare the success rate and determine the statical measures for a larger sample size of data.

This project will fall within the Human, Organisational & Regulatory Aspects of this MSc course. Cyber security risk assessment and management is crucial to everyone living and working in the digital world.

Cyberbullying

Cyberbullying can be a significant risk in today's digital age and is defined as the act of one or more individuals frightening another using digital communication such as social media platforms, instant messaging, online forums, or text messages. It comprises using technology to target someone to inflict harm on them. This can be further categorised in the following way:-

- Cyber-harassment - To send repeated and offensive messages to a specific person, causing mental and emotional distress.
- Cyber-stalking - Harassment, violence or threats toward a person in order to isolate and frighten them.
- Denigration - The distribution of false or derogatory messages.
- Impersonation - Access to the personal information of the victim in order to impersonate them.

- Tricky – To befriend the victim, allowing them to share private information and then spreading the information or threatening to do so.
- Flaming – Send harmful messages to provoke verbal conflicts.

(Vismara et al., 2022)

A major concern for parents of this generation is bullying. A global study of 20,000 parents showed that cyberbullying on social media sites was one of their highest concerns, with 65% claiming that this worried them (digital cooperation, 2022)

According to ‘Teens and Cyberbullying 2022’ research conducted in the USA found that teenagers between the ages of 13 - 17 have experienced cyberbullying in one of the following forms

• Offensive name-calling	32%
• Spreading or false rumours	22 %
• Receiving explicit images	17%
• Harassment	15%
• Physical Threats	10%
• Images shared without consent	7%

(Anderson, 2018)

Research performed by Ofcom in the United Kingdom found that parents had a genuine worry about keeping their children safe online. The following percentage of parents (based on children’s age range) were concerned about their children:

- 76% for 8 – 11 year olds

- 71 % for 12 – 15 year olds
- Another finding of the same study was that, 39 % of 8 – 17 year olds said they had been bullied both on-line and face to face.
- The most common form of bullying experienced by these individuals was by:
- Text or messaging app 56%
- Social media 43%

(Ofcom, 2022)

The same study also found that although 91% of parents knew safety controls were available, only on one in seven parents actually implemented these rules.

Another survey performed in the United Kingdom in 2022 from the period of January 2022 to February 2022 that had a total of 534 respondents showed that the most common online abuse was cyberbullying. The study broke down cyberbullying into the following categories:

- Online Harassment 45%
- Trolling – 36 %
- Cyber Stalking 33%
- Accounts Hacked 31%
- Brigading 30%
- Virtual Mobbing 24%
- Physical sexual violence 24%
- Cyber Flashing 24%

(Dixon, 2022)

To highlight this as a global problem, research was conducted where 20,793 interviews were conducted and found that India had the highest percentage of cyberbullying at 37% , Brazil at 29 % and United States at 26% This was broken down into age groups as follows the highesr being Ages between 14 – 18

The form of bullying took place over

- Social media sites and app 19.2%
- Text Messages 11%
- Video Gaming 7.9%
- Non –social media websites 6.8%
- Email – 3.3%

(Cook, 2018)

The impact of cyberbullying can have devastating effects on the victim as well as the offenders. These effects can range from the following.

- Depression
- Suicidal Thoughts
- Drug abuse
- Mental health issues

(Maurya et al., 2022)

This list indicates that we need to pay special attention on the Human Factor and ensure the risk is communicated.

Artificial intelligence

For this we will look at Artificial intelligence, which 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 models can be categorised into the following:

- 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 structure that recursively partitions 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.
- 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)

Chatbot

A chatbot is a program that mimics human conversation via text or voice interactions. An automated program that executes instructions based on specified inputs and provides feedback that mimics natural conversational interaction. Depending on the needs, platforms, and technology, chatbots can perform a wide range of communication and interaction functions. They are frequently created utilising artificial intelligence (AI) technologies and can be deployed on a variety of platforms, including messaging applications. (Kumar, 2021 - Adamopoulou and Moussiades, 2020)

There are several chat bots available which can be categorised as follows.

- A tree-based bot that only responds to question that are pre-defined in the database. Able to respond to key words or catchphrases rather than free text
- Artificial Intelligence (AI) bot which has the ability to update and learn from previous conversations using a Natural Learning Process
- Hybrid a combination of both tree bases and AI

Purpose of chat bots can range from fun interactive games and can commonly be seen in the household like Alexa or Google Echo. These bots are designed for a specific purpose to ease admin tasks such as adding items to on-line store. (Haristiani, 2019, - Caldarini, Jaf and McGarry, 2022)

Previous studies have been conducted in this area where a Cyberbullying Awareness and Prevention Through Artificial Intelligence chat bot was implemented and its focus was to identify cyberbullying messages, answer questions regarding cyberbullying and finally provide tips on how to stop cyberbullying. The data was sourced from previous studies. The data was then trained in various models including.

- Stochastic Gradient Descent (SGD) Classifier.
- Logistic Regression.
- Random Forest.

The results showed that SGD Classifier achieved the highest accuracy of 89.13%

(Lian, Alfredo Costilla Reyes and Hu, 2023)

Ethical and risk Click

To ensure the experiment is not skewed in any way we need to consider ethical and risk assessments:

- Bias

Machine learning models rely on training data, and if the data used to train the model is biased then the results produced will be incorrect. The data needs to be raw and labelled in ambiguous way.

- Learning

Just like the technology around us is constantly changing, so are phases and vocabulary. We need to take into account slang words, non-English words, phrases in both English and non-English and a mix of both in a sentence. This means the model should be constantly learning.

- Privacy Concerns

The models being implemented will require access to personal data, conversations and online activities. This raises concerns about privacy and the potential misuse of sensitive information.

- Complexity

Machine learning models used for cyberbullying detection can be complex and difficult to interpret. Lack of transparency can make it challenging to understand how the models make decisions and identify potential biases.

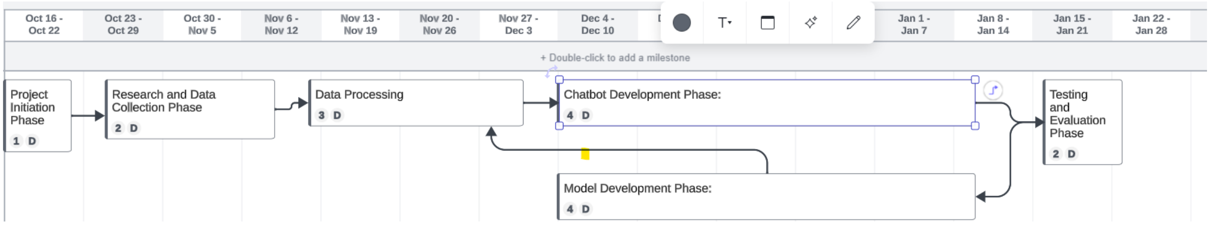
(Milosevic, Van Royen and Davis, 2022)

Artefacts

For this experiment, the proposal is to build a full stack consisting of

- Front End – This is where the messages will be sent and received between individuals and would require a Chat Interface or a Plug-in for a messaging apps like FaceBook Messenger, WhatsApp, Slack or SMS
- Back end – Python based program where the logic of the chat bot will reside, Handle responses from the front end. Deal with database interactions.
- Libraries – To allow for Flask (API) and other bundled code.
- Websocket – This is for applications that require real time communication.
- Database - This is where the data will be stored that the model will use to learn.

Timeline



1. Project Initiation:

2. Data Collection:

- Gather relevant data on cyberbullying.

- Identify and collect labelled datasets for training the chatbot.

3. Data Processing:

- Clean and preprocess the collected data.

- Perform feature selection and extraction.

- Prepare the data for training the machine learning model.

4. Model Development:

- Select appropriate machine learning algorithms for the chatbot.

- Train the model using labels.

- Fine-tune the model.

- Evaluate of the model.

5. Development:

- Develop the chatbot interface.

- Integrate the trained machine learning model into the chatbot.

- Implement natural language processing (NLP) techniques

6. Testing and Evaluation Phase:

- Conduct testing of the chatbot's functionality and performance.
- Evaluate the effectiveness of the chatbot in detecting incidents.

Future Work

Depending on the success of the model, it would be beneficial to simulate this on real life individuals. For this, it would be beneficial to run and a pre and pro- testing analysis with a mixed method of research consisting of a both qualitative and quantitative test. Having a sample of random people participate in a questionnaire to see if any cyberbullying has taken place. Run the algorithm models while the sample participate with each other over a period of 6 – 12 months over an online chat. At the end of the trial ask them to repeat the questionnaire. Any sample that has confirmed that they experienced cyberbullying are then interviewed further with a psychological test which will then be reviewed to find the true impact of cyberbullying.

Reference list

Adamopoulou, E. and Moussiades, L. (2020). An Overview of Chatbot Technology. IFIP Advances in Information and Communication Technology, 584(1), pp.373–383. doi:https://doi.org/10.1007/978-3-030-49186-4_31.

Anderson, M. (2018). A Majority of Teens Have Experienced Some Form of Cyberbullying. Pew Research Center: Internet, Science & Tech. Available from: https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/ [Accessed 8 Oct. 2023].

Caldarini, G., Jaf, S. and McGarry, K. (2022). A Literature Survey of Recent Advances in Chatbots. Information, 13(1), p.41. doi:https://doi.org/10.3390/info13010041.

Cook, S. (2018). Cyberbullying Statistics and Facts for 2016 - 2018 | Comparitech. Comparitech. Available from: https://www.comparitech.com/internet-providers/cyberbullying-statistics/ [Accessed 8 Oct. 2023].

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. arXiv:1703.04009 [cs]. Available from: https://arxiv.org/abs/1703.04009.

digital cooperation (2022). Cyberbullying Statistics and Data for 2022. Digital Cooperation. Available from: https://digitalcooperation.org/research/cyberbullying/ [Accessed 8 Oct. 2023].

Dixon, S. (2022). UK victims on types of online abuse 2022. Statista. Available from: https://www.statista.com/statistics/1319815/uk-online-abuse-experienced/ [Accessed 8 Oct. 2023].

Haristiani, N. (2019). Artificial Intelligence (AI) Chatbot as Language Learning Medium: An inquiry. Journal of Physics: Conference Series, 1387(1), p.012020. doi:https://doi.org/10.1088/1742-6596/1387/1/012020.

Hinduja, S. (2021). Cyberbullying in 2021 by Age, Gender, Sexual Orientation, and Race. Cyberbullying Research Center. Available from: https://cyberbullying.org/cyberbullying-statistics-age-gender-sexual-orientation-race [Accessed 8 Oct. 2023].

Kumar, J.A. (2021). Educational chatbots for project-based learning: investigating learning outcomes for a team-based design course. International Journal of Educational Technology in Higher Education, 18(1). doi:https://doi.org/10.1186/s41239-021-00302-w.

Lian, A.T., Alfredo Costilla Reyes and Hu, X. (2023). CAPTAIN: An AI-Based Chatbot for Cyberbullying Prevention and Intervention. Lecture Notes in Computer Science, 14051, pp.98–107. doi:https://doi.org/10.1007/978-3-031-35894-4_7.

Maurya, C., Muhammad, T., Dhillon, P. and Maurya, P. (2022). The effects of cyberbullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from India. BMC Psychiatry, 22(1). doi:https://doi.org/10.1186/s12888-022-04238-x.

Milosevic, T., Van Royen, K. and Davis, B. (2022). Artificial Intelligence to Address Cyberbullying, Harassment and Abuse: New Directions in the Midst of Complexity. International Journal of Bullying Prevention. doi:https://doi.org/10.1007/s42380-022-00117-x.

Ofcom (2022). Children and parents: Media use and attitudes report 2022. OFCOM. Ofcom. Available from: https://www.ofcom.org.uk/__data/assets/pdf_file/0024/234609/childrens-media-use-and-attitudes-report-2022.pdf [Accessed 8 Oct. 2023].

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

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, 131(131), pp.323–340. doi:https://doi.org/10.1007/978-981-19-1844-5_25.

Vismara, M., Girone, N., Conti, D., Nicolini, G. and Dell’Osso, B. (2022). The current status of Cyberbullying research: a short review of the literature. Current Opinion in

Behavioral Sciences, 46(101152), p.101152. doi:https://doi.org/10.1016/j.cobeha.2022.101152.

Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL Student Research Workshop. doi:https://doi.org/10.18653/v1/n16-2013.

www.kaggle.com. (n.d.). Tweets Dataset for Detection of Cyber-Trolls. Available at: https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls [Accessed 9 Oct. 2023].