Utilization of Artificial Intelligence for Social Media and Gaming Moderation

Heba Saleous
Department of Information Systems & Security,
United Arab Emirates University
Al Ain, United Arab Emirates
201670187@uaeu.ac.ae

Marton Gergely
Department of Information Systems & Security,
United Arab Emirates University,
Al Ain, United Arab Emirates
mgergely@uaeu.ac.ae

Khaled Shuaib
Department of Information Systems & Security,
United Arab Emirates University,
Al Ain, United Arab Emirates
k.Shuaib@uaeu.ac.ae

2023 15th International Conference on Innovations in Information Technology (IIT) | 979-8-3503-8239-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IIT59782.2023.10366468

***Abstract*— As the world continues to evolve, technology has proven to be a necessity in the lives of everyone. Evolving beyond professional use, cyberspace is now populated by online communities being used for communication, learning, and entertainment. However, the increased online presence exposes users to a variety of cultures, personalities, and levels of maturity. Some may also seek to cause harm to others through cyberbullying or may display toxic behaviors. This research aims to tackle the growing problem of toxicity and harassment in online environments. The proposed solution will utilize Artificial Intelligence (AI), and more specifically Natural Language Processing (NLP), to moderate communication and detect malicious language and behavior. The efforts shared in this paper specifically present a work-in-progress. For the time being, two models have been tested with a single dataset from Twitter: a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). The results of experimentation show a promising start for the use of NLP in moderation with an 83% accuracy using an RNN.**

Keywords—NLP, Sentiment Analysis, Online Harassment, User Moderation

INTRODUCTION

Social media has had a major presence around the world for the past decade under the guise of connecting friends and family wherever, whenever. However, this service has evolved to become more than just a communication tool for those that know each other. Social media platforms have become media for meeting new people, creating communities of like-minded individuals, and allowing groups of people to participate in various activities together. Even organizations take advantage of social media platforms for marketing purposes.

While social media platforms have proven to be beneficial in terms of communication, there are users that take advantage of the increasing online presence. Social media can be used negatively in several ways. Firstly, adversaries can create fake profiles in an attempt to phish for user information and commit cybercrimes. If the fake profile was made carefully enough, most users would not differentiate between the false and legitimate pages.

Another negative use of social media is to abuse the simplicity and anonymity to harass other users. While most people will follow the rules of whichever platform they are using to maintain the privilege to use it, others may misbehave and act out against others. As such, the type of material shared by the more malicious population may be deemed harmful, disturbing, or explicit. To maintain the sanctity of these platforms, moderators must review the content being shared to ensure that rules are being followed and that material is safe for everyone. However, this means that they are exposed to whatever material the more toxic users are sharing, which can range from extremely vulgar language to disturbing photos or videos. Additionally, the massive amount of social activity that occurs within just one platform can become too overwhelming to moderate, leaving some toxic behaviors unpunished. This can have a negative impact on other, more innocent users, especially if they are underage or sensitive to certain material.

NLP can be used to assist with moderation by automating this task while effectively filtering out content. Over the past year, NLP models have evolved greatly, creating a boom in the use of artificial intelligence (AI) for numerous applications, especially with the release of ChatGPT by OpenAI in November 2022. Additionally, rival companies, such as Google (hosts of both LaMDA and BERT), are announcing improvements to their own models to match OpenAI's feat. As a result, developers all around the world of all ages and skill levels are now creating applications revolving around these NLP models for daily use.

In this work, we present a work-in-progress (WiP) potential solution to online environment moderation challenges. We begin approaching the possibility of using NLP to moderate online interactions by selecting two commonly used models, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), and training them using a dataset compiled from Twitter. The results of training and validation are used to determine the success of these models with such tasks. The results will also be compared to each other to determine the effectiveness of each model and determine the more "ideal" one.

The purpose of starting with this small experiment is to begin preparing the models for the overall project. The first task of the project is to decide on which models to use and to test them with different datasets from various social media platforms. This

paper covers the work completed so far, as well as the current and next steps for the project.

The remainder of this work is organized as follows: Section 2 will go over the background and context of this work. Section 3 describes the methodology followed in the overall project that the work done in this paper is a part of to address the problems mentioned. Section 4 presents the results gathered by training and testing the two models. Section 5 will discuss the implications of these results and the remaining goals and tasks of the project. Finally, the paper will conclude in Section 6.

BACKGROUND

The dependency on social media and online communities, while beneficial in terms of closing the gaps encountered mere decades ago, is becoming concerning. New challenges are arising as a result of users taking advantage of the opportunities offered by online environments. Different cybercrimes, such as fraud, stalking, or cyberbullying, have become common in online environments. Over the past 5 years, since 2018, the number of cyber incident reports per years has increased by 449,007 in total, with $10.3 billion financial loss in 2022 alone [1].

Hosts of online platforms offer users the ability to report any foul play. These reports are then presented to moderators who review the case to determine their severity and the kind of punishment appropriate. Depending on the platforms and the type of content involved, reports may include material that are deemed disturbing. Depending on the platforms and the type of content involved, reports may include material that is deemed disturbing. TikTok moderators, for example, have claimed that they were being trained and forced to moderate the platform using uncensored, explicit material [2], [3].

Along with the increased popularity of online environment, artificial intelligence has also been growing as a field. The uses of this technology are vast, from being used to detect health problems [4], [5] to improving education [6], [7]. The release of ChatGPT in November 2022 has inspired a boom in AI learning and usage with the same company that released it, OpenAI, shortly following up with the announcement of GPT-4. Meanwhile, rival companies, such as Google (who host both LaMDA and BERT), are also announcing improvements to their own models to match OpenAI's feat. Developers all around the globe of all ages and skill levels are now creating applications revolving around these NLP models to use daily. Given the advancements of modern NLP models, online environment moderation can be improved using AI while also preserving the wellbeing of employees.

A. Video Game Harassment Survey

One of the most popular online environments that users flock to is an online multiplayer video game. When thinking about online platforms, most people might think of social media websites, such as Facebook or Twitter/X, or communication tools like Discord or Zoom. A lesser studied online environment is an online-based video game.

Video games have become a big part of the digital world, evolving from mere sources of entertainment to entire career paths. Combined with improved device accessibility and implemented network-based features, gamers all around the world can play together any time they want. While the increased social aspects of video games mean that players can make new friends wherever they are in the world, some gamers take advantage of anonymity to harass others for various reasons. As the number of gamers grows over the years, so does the number of toxic behaviors found in online video games.

A study was conducted within the United Arab Emirates University (UAEU) to collect their experiences with toxicity and harassment in online video games. The questionnaire was sent to the UAEU community and included questions about various aspects of gaming, such as gaming habits, devices, the emotional/mental effects of in-game harassment, and player reporting.

After filtering out users that spent little-to-no time in online games (based on responses to questions about gaming habits), a total of 494 responses considered for this survey. Approximately 71% of respondents stated that they have been the victim of harassment while playing games, with 83% stating that they have witnessed others being bullied.

Respondents were also asked about the effect experiencing toxicity in-game had on their mood. The results of this question are summarized in Fig. 1. While 35% of participants stated that there was no effect on their mood, more responses (approximately 44%) stated that experiencing harassment in-game had a negative effect on their mood for the remainder of the day.

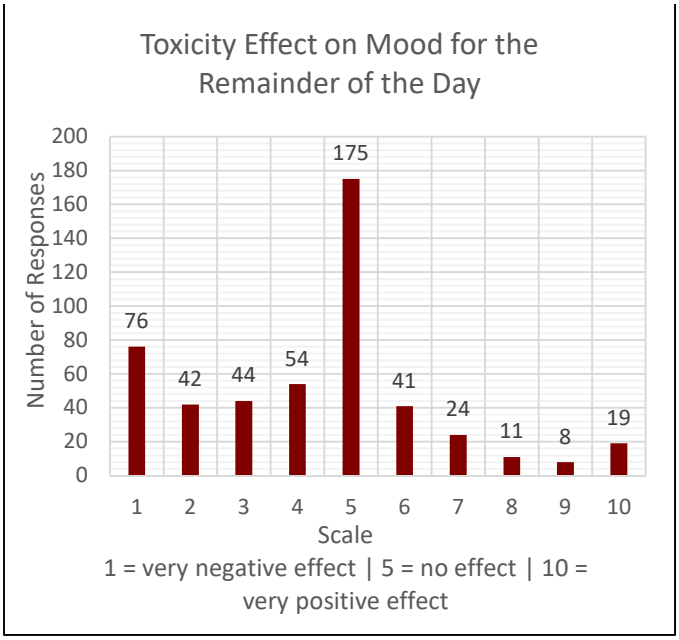

Fig. 1. The effect of harassment on respondents' moods.

When asked about their opinion on whether game companies are doing a good job addressing these issues in-game, respondents had stronger opinions. A great majority of the responses leaned negatively, stating that they did not think companies were handling the toxicity problem well at all. The overall distribution of responses is visualized in Fig. 2.

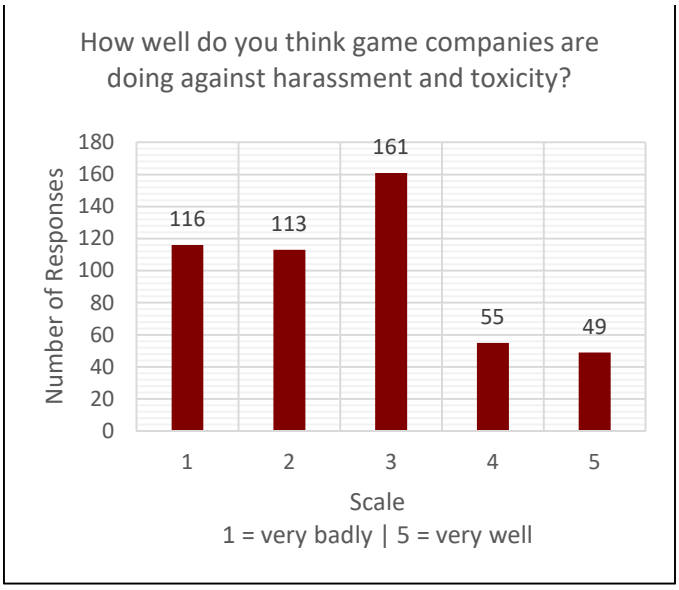

Fig. 2. Opinions on how well in-game toxicity is being handled.

Based on these responses, there is clear room for improvement with regards to approaching the toxicity problem in online video games. Additionally, this is just one online platform being studied to provide an example of the extent of this problem. More work will need to be done to improve the atmosphere of online environments.

METHODOLOGY

To address the challenges posed by online environments and the toxic behaviors that may occur, the goals of the overall project must first be clarified. As such, the following are the research questions asked in the overall project:

1. How does toxic behavior effect other players in online environments?
2. What can be done to assist game companies with moderating online interactions?
3. In what ways can users be held accountable for their actions against others?

The work presented in this paper aims to begin answering research question two. The first question was answered by the survey that was reviewed in the previous section. The third question will be addressed once the second question has been fully answered.

To answer research question two, the most popular models used for sentiment analysis are explored, namely CNN, RNN, BERT, GPT, and XLNet. The work presented in this paper specifically focuses on CNN and RNN. Secondly, the data to be studied must be discovered, saved, and cleaned. Finally, the model must be implemented in an environment with sufficient resources for testing.

The datasets used in this project come from various online platforms, such as Twitter/X, Facebook, Wikipedia, or gaming chats from multiplayer online games, such as DotA2, Valorant, League of Legends, and Overwatch. The contents of the datasets are cleansed to ensure that they are consistent with one another. More specifically, the content to be classified is sanitized of emojis and special characters for current testing. Additionally, the categories for classification are currently limited to 'Sexism,' 'Racism,' and 'None.' For this paper, only one dataset[1] has been used to ensure that the models are able to analyze the datasets and successfully give results. Once the models have been configured properly, multiple datasets will be used for training/testing.

The two models presented in this work are the CNN and the RNN. The environment used to create and configure the models is the Google Colaboratory using a T4 GPU. Colab was used to ensure that the models were being tested in an "even" testbed; the challenges of manually downloading libraries and ensuring that device constraints are met are avoided by using Google's environment.

The configurations were kept as consistent as possible between the two models. More specifically, the following are common configurations:

- The train/test division is 70/30
- Random State = 42
- 5 Epochs were tested
- Batch Size = 20

These common configurations were decided after testing out various combinations of each and determining the ideal options for both models. "Ideal" in this case refers to the accuracy of the model and the computational resources consumed.

RESULTS

The experiment presented in this paper was run with one version of a CNN and two different versions of an RNN (Simple/Vanilla and Long Short-Term Memory). Each were tested with the same dataset (one from Twitter/X) in the same environment (Google Colab using a T4 GPU). For each model, the precision, recall, and F1 Score were collected to determine how each model behaved with the dataset used. The results of each in terms of average precision, recall, and F1 score are summarized in Table 1.

Overall, the LSTM-based RNN achieved a higher precision, recall, and F1 Score using the single Twitter/X dataset, deeming this model the preferable one of the three that were tested.

TABLE 1: THE AVERAGE PERFORMANCE METRICS FOR EACH MODEL.

Model	Precision	Recall	F1 Score
CNN	0.82	0.82	0.82
SimpleRNN	0.79	0.7	0.73
LSTM RNN	**0.83**	**0.82**	**0.83**

In addition to the precision, recall, and F1 score, the loss and accuracy throughout the duration of the runtime was also

[1] Source: https://www.kaggle.com/datasets/virajtapkir/twitter-parsed-dataset

recorded for each model. Fig. 3-Fig. 5 depict the loss and accuracy for the CNN, Vanilla RNN, and LSTM RNN, respectively.

Each of the graphs depict a decreasing loss and increasing accuracy over the five epochs (blue line). However, the validation loss and accuracy (orange line) seem to suffer, leaving room for potential improvement. These improvements

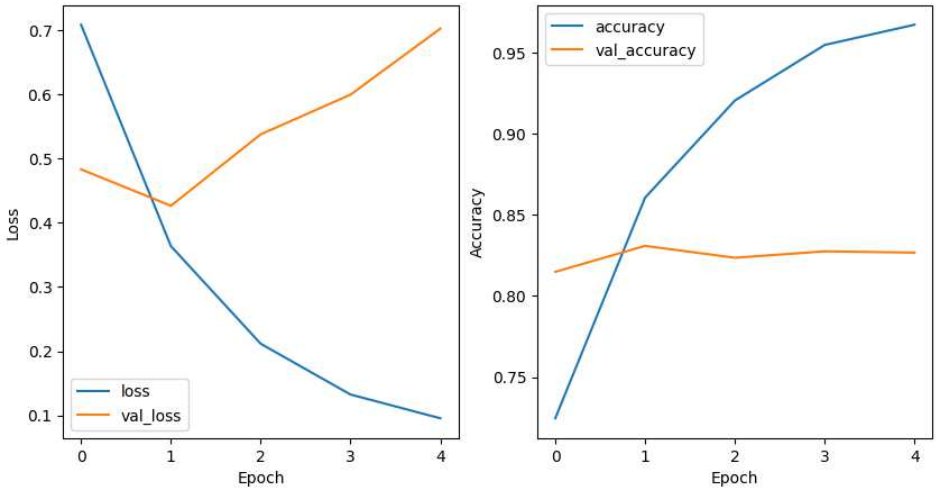

Fig. 3. The loss and accuracy of the CNN.

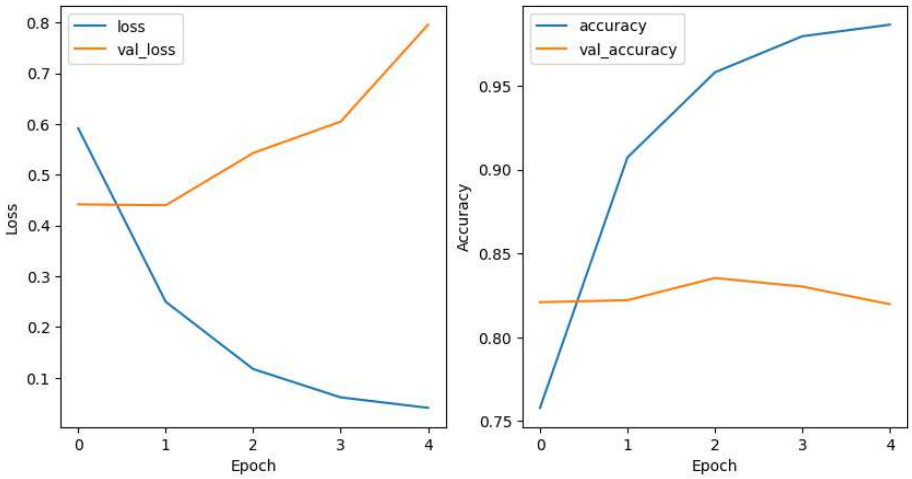

Fig. 4. The loss and accuracy of the Vanilla RNN.

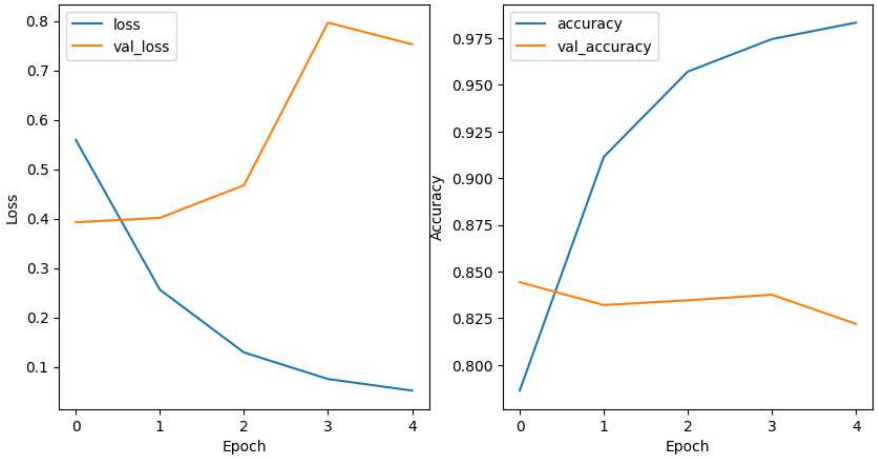

Fig. 5. The loss and accuracy of the LSTM RNN.

potentially come in the form of modified model configurations or the addition of more datasets for training and testing.

DISCUSSION

Toxic behavior online has become a cyberpandemic that affects users all around the world. Game companies have already begun to address this issue by making the conditions for punishment stricter and encouraging users to report offenders. However, based on the results of the questionnaire summarized in this work and the potential effects negative content may have on those witnessing it, there is room for improvement with regards to content moderation. The project that the work presented in this paper is part of seeks to address these issues, namely text communication between users in online environments.

AI can be used to assist with user moderation while, at the same time, reducing the amount of human contact with potentially harmful material. The objective of the project is to take advantage of NLP models and train them against social platform data to determine their effectiveness with text classification. The progress presented in this paper shows the results of testing a CNN and two types of RNNs against a Twitter/X dataset. Based on the results shared, the LSTM-based RNN performed the best in terms of precision, recall, and F1 score.

While the initial results presented in this WiP gave a positive impression of NLP use for user moderation, there is still room for improvement. There are some fallbacks to the progress described throughout this paper. The configurations of models can still be improved further to improve performance metric values. While 83% is a decent achievement for these models, there is plenty of room to improve this rate of success. One configuration that could be modified is the number of epochs that the code runs through while working. For this WiP, only 5 epochs were completed. This small value was selected because it allowed the code to run quickly enough to determine if it would yield results while beginning to present some classification pattern. Increasing the number of epochs the model completes would likely improve the final results. Additionally, the batch size, the train/test ratio, and the random state could be modified further to determine more "ideal" values.

Now that the CNN and RNN codes have successfully yielded results, the next step is to focus on the other NLP models to be tested for this project. The remaining models to be configured are BERT, GPT, and XLNet. These are more advanced models compared to the NNs presented in this work and could provide more promising results.

Once each of the models has been tested and compared, a prototype moderation system can be built revolving around NLP. This system would focus on text flagging and classification, as well as user accountability according to the extremity of the offense. The end goal of the project that this work is a part of is to provide a proof-of-concept NLP-based user moderation system. Such a system would not only automate the process of analyzing flagged texts for punishment but also (1) reduce the likelihood of human error and (2) reduce the amount of contact human moderators would have with potentially harmful material. The summarized processes that would occur throughout this project are shown in the training cycle shown in Fig. 6 and the functional model in Fig. 7.

CONCLUSION

Online environments have evolved beyond simply enabling easy communication between people all around the world. These platforms are now used for healthcare, education, and entertainment. While there are many benefits to the rise in popularity of online environments, there are also several

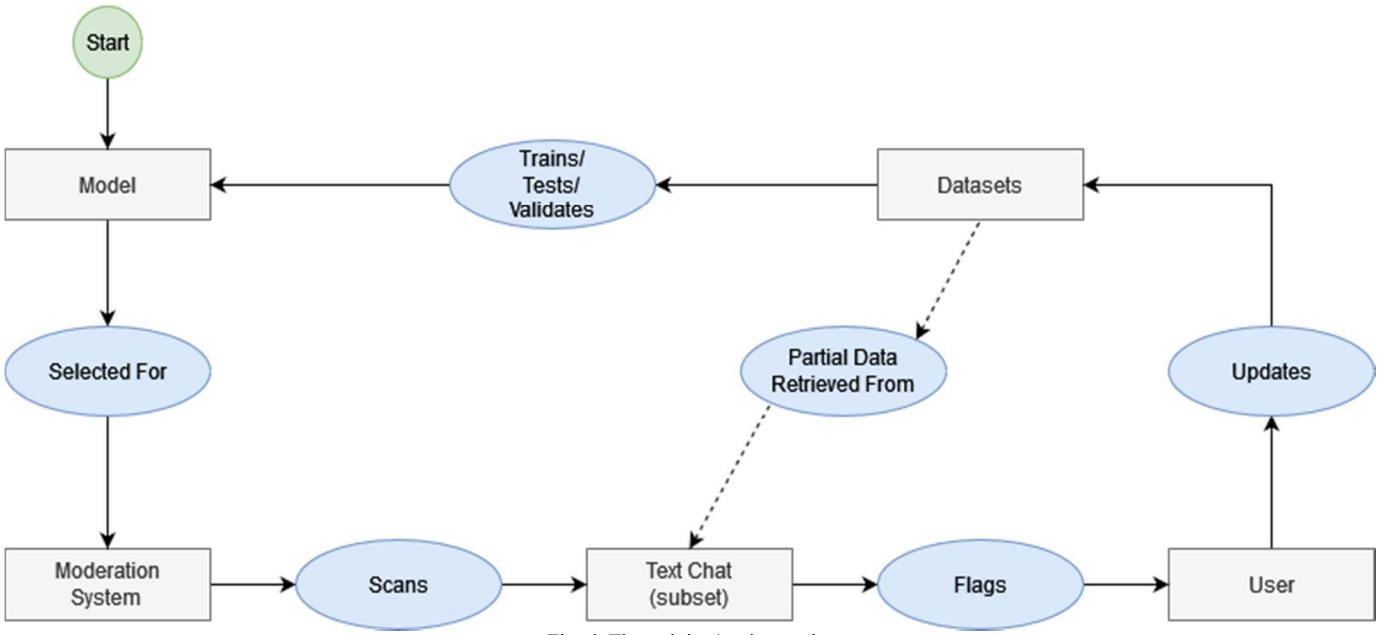

Fig. 6. The training/testing cycle.

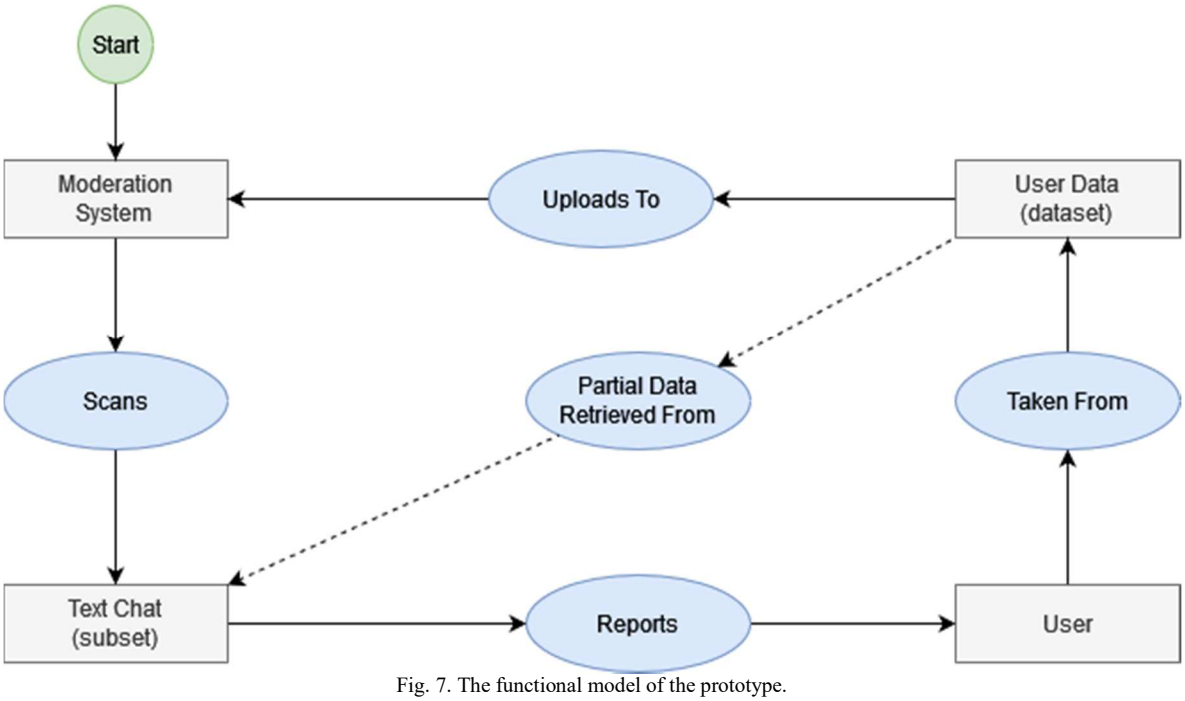

Fig. 7. The functional model of the prototype.

challenges that remain to be addressed. More specifically, these works target the prevalence of online harassment and toxic behaviors.

This work specifically presents current progress that is part of a bigger project aiming to target this challenge. In this WiP, two well-known AI models - a CNN, a Vanilla RNN, and an LSTM RNN - are tested against a Twitter/X dataset to determine their initial behaviors with similarly formatted text data. The models were configured in such a way that would ensure the code works while also providing preemptive impressions on the use of AI for user moderation. The results show that, with very basic configurations, the LSTM-based RNN performed better than the other two models with an F1 score of 0.83. While not high, this result shows promise that, with more experimentation and better configurations, AI could be used to moderate user communication in online social platforms.

Given these results, there remains room for improvement. The next step is to run similar experimentation on other, more complex NLP models, such as BERT, GPT, and XLNet. Once the code for each model is confirmed to be successful, different configurations of each model will be tested to achieve the ideal results. Successfully completing this task will offer room for training these models against multiple datasets at once to provide more comprehensive "knowledge" for classification. Once all of this is complete, a prototype user moderation system can be developed to test the automation of user flagging and accountability.

REFERENCES

[1] Federal Bureau of Investigation, "Internet Crime Report 2022." Internet Crime Complaint Center.

[2] A. Levine, "TikTok Moderators Are Being Trained Using Graphic Images Of Child Sexual Abuse," Forbes, Aug. 04, 2022. [Online]. Available: https://www.forbes.com/sites/alexandralevine/2022/08/04/tiktok-is-storing-uncensored-images-of-child-sexual-abuse-and-using-them-to-train-moderators/?sh=641173935acb

[3] B. Berthelot and H. Ren, "TikTok's Moderators Still Review Child Abuse Despite Vow to Exit," Bloomberg, Mar. 20, 2023. [Online]. Available: https://www.bloomberg.com/news/articles/2023-03-20/tiktok-s-moderators-still-review-child-abuse-despite-vow-to-exit

[4] D. Nguyen, "How AI Can Help Diagnose Rare Diseases," Harvard Medical School. [Online]. Available: https://hms.harvard.edu/news/how-ai-can-help-diagnose-rare-diseases

[5] Y. Kumar, A. Koul, R. Singla, and M. Fazal Ijaz, "Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review, Synthesizing Framework and Future Research Agenda," J Ambient Intell Humaniz Comput, Jan. 2022, doi: 10.1007/s12652-021-03612-z.

[6] Unesco, "Artificial Intelligence and the Futures of Learning," Unesco, Feb. 2023.

[7] Office of Educational Technology, "AI and the Future of Teaching and Learning: New Interactions, New Choices," Medium. [Online]. Available: https://medium.com/ai-and-the-future-of-teaching-and-learning/ai-and-the-future-of-teaching-and-learning-new-interactions-new-choices-c726bcf03012