CAPTAIN: An AI-Based Chatbot for Cyberbullying Prevention and Intervention

Andrew T. Lian[1](✉), Alfredo Costilla Reyes[2], and Xia Hu[2]

[1] The Kinkaid School, Houston, TX 77024, USA
andrew.lian@kinkaid.org
[2] Department of Computer Science, Rice University, Houston, TX 77005, USA

Abstract. Cyberbullying is a widespread and growing problem that can cause various psychological and health well-being outcomes in youth and is considered a serious public health threat. Cutting-edge informatics technology would enable us to identify and stop cyberbullying to prevent harm, death, and privacy violations. However, current cyberbullying prevention approaches offer limited interactions, individualized education, and in-time intervention. With the current emerging technologies in Artificial Intelligence (AI), the use of chatbots have become increasingly popular in health promotion. However, there are current technological challenges that need to be addressed, such as detecting and preventing cyberbullying in real-time, providing personalized responses and intervention, as well as developing the chatbot with a user-friendly interface. This paper introduces CAPTAIN (Cyberbullying Awareness and Prevention Through Artificial INtelligence), an AI-based chatbot for cyberbullying prevention that can provide anytime interaction for personalized intervention.

Keywords: Cyberbullying prevention · Machine Learning · Topic Modeling · Chatbot

1 Introduction

The widespread use of digital technologies has contributed to the rise of cyberbullying, which has become a significant problem in the digital age. Cyberbullying refers to using information technology and digital platforms to harass, intimidate, or harm people in any format. According to a recent survey, about 10% of the teens have experienced cyberbullying during their lifetime and 59% of them reported that it has happened within the past year [1]. Many people also didn't realize that they were bullying others before the consequences. Around 23% of teens claimed that they have done something cruel online to others and almost 60% of children online have seen someone getting cyberbullied and most of them have not intervened [2]. In addition to self-harm and suicidal behaviors, cyberbullying can cause several psychological well-being outcomes, including emotional distress, violence, family conflict, relationship problems, substance abuse, and learning differences [3].

Different from traditional bullying, cyberbullying has some unique features such as publicity, anonymity, and the lack of supervision, which brings additional challenges

H. Degen and S. Ntoa (Eds.): HCII 2023, LNAI 14051, pp. 98–107, 2023.
https://doi.org/10.1007/978-3-031-35894-4_7

for cyberbullying prevention. Current methods for cyberbullying intervention and prevention include pamphlets, websites and social media, as well as school programs to educate people about cyberbullying. A recent survey indicates that most online education resources appear to target parents [4]. In addition, these methods lack interaction, provide no individualized education or require extensive resources to provide in-time intervention [5].

Cutting-edge Artificial Intelligence (AI) technologies would enable us to identify and stop cyberbullying. One way is by using machine learning methods to automatically detect and flag potentially harmful or abusive language in online conversations and social media posts. This can help identify cyberbullying before it escalates. In addition, AI-powered chatbots can provide real time interaction and communication with the users who need help on cyberbullying prevention.

Much research has focused on automatic cyberbully detention using machine learning methods. Different types of social media data have been used for this task including YouTube, Instagram, Whatsapp, and Twitter [6]. Scientists usually focus on one data source because each community can act differently [7]. Different research also focus on different tasks that cover various perspectives on cyberbullying. Many efforts focus on binary classification to detect whether a message is harmful or not [6]. Other efforts focus on more specific tasks. For example, Van Hee et. al. Developed an annotation guideline to annotate online messages with respect to the cyberbullying participant roles (e.g., harasser, bystander assistant, bystander Defender, victim, not cyberbullying) and cyberbullying subcategories (e.g., curse, defamation, defense, encouragement, insult, and sexual) [8]. However, they still focused on binary classifications for automatic detection. Other researchers have focused on simple types of subcategories such as racism, sexism [7] and severity (low, medium and high) [9].

Machine learning techniques have been widely applied in the field of automatic bullying message detection. A recent research survey summarized two different categories of techniques, conventional machine learning and deep learning [6]. While conventional machine learning methods perform relatively well for single source (e.g., Twitter) binary classification, deep learning models such as Recurrent Neural Network (RNN), Current Neural Network (CNN), and Multi-layered perceptron (MLP) have demonstrated improved performance in detecting cyberbullying, especially when used in combination with contextual language models like Bidirectional Encoder Representations from Transformers (BERT) [6].

With the current emerging technologies in Artificial Intelligence (AI), the use of chatbots have become increasingly popular in health promotion especially after the beginning of the COVID-19 pandemic [10]. Chatbots have been developed to help youth and adolescent mental health. For example, Deshpande and Warren developed a machine learning based self-harm detection method for mental health chatbots [11]. A recent study evaluated the perceptions of a chatbot developed to psychoeducate adolescents on depression in a small group of participants [12]. Another example is Vivbot, a chatbot that can deliver positive psychology skills and promote well-being among young people after cancer treatment [13]. The results of both studies indicate that chatbots can be potentially more engageable and acceptable to adolescents, who tend to be reluctant to traditional mental service. In addition to mental health, chatbots are also being used in

other health interventions such as vaccine promotion [14], cancer risk triangle [15], and life skill coaching [16].

In this paper, we introduce CAPTAIN (Cyberbullying Awareness and Prevention Through Artificial INtelligence), an AI-based chatbot for cyberbullying prevention. More specifically, CAPTAIN can (1) automatically detect cyberbullying messages, (2) answer questions regarding bullying, and (3) provide tips to the users on how to prevent/stop cyberbullying. CAPTAIN combines machine learning based tools with a chatbot to enable real time bullying message detection and cyberbullying prevention. To the best of our knowledge, this is the first chatbot that targets AI-based personalized intervention for cyberbullying prevention.

2 Method

The CAPTAIN system includes three components: cyberbullying detection, promoting cyberbullying prevention, and cyberbullying data analysis. In the first component, we implemented a machine learning based model to automatically classify online messages as either a bullying sentence or a non-bullying sentence. This component also facilitates the second component which is the interface of the chatbot. The chatbot aims to provide tailored interaction with users. It can interact with the users to answer their questions, provide information and resources about cyberbullying, and detect cyberbullying messages using our machine learning model. This is where users communicate and have conversations with our AI through the interface. The third component of CAPTAIN, the cyberbullying detection system, uses topic modeling to cluster the data into topics and super topics. It can provide a more in-depth understanding of the types of cyberbullying that are occurring and the reasons behind them. This component informs the interface by identifying the topic of the bullying messages, based on which more effective interventions can be designed and provided (Fig. 1).

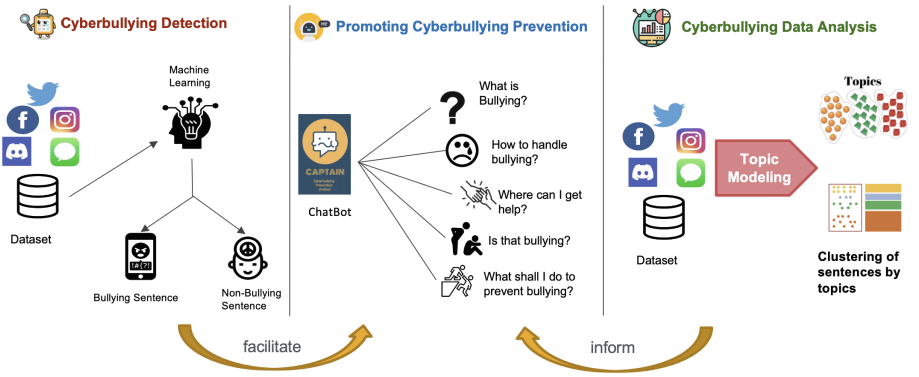

Fig. 1. System Overview of the Three Major Components of the CAPTAIN Chatbot: (1) Machine Learning based Cyberbullying Detection (2) Personalized Cyberbullying Prevention; and (3) Cyberbullying Data Analysis using Topic Modeling

Datasets and Data Preprocessing

We collected a set of annotated messages, including both bullying and non-bullying from three different sources [17–19]. These data sets were used to train the machine learning models and perform data analysis. In order to prepare the data for machine learning, we preprocessed it by removing common english stop words, slang words, URLs, emojis, and acronyms that are commonly found in web and SMS messages. This preprocessing step helped to improve the accuracy of our models and the overall effectiveness of our analysis.

Automatic Cyberbullying Detection

We utilized the scikit-learn machine learning library on the collected data set to implement, test, and evaluate nine different machine learning algorithms, including Stochastic Gradient Descent (SGD) Classifier, Logistic Regression, Random Forest, to determine their performance on automatically classifying messages into categories of bullying or non-bullying. We split the data sets into a training set and a test det, with the training set comprising 80% of the data and the test set comprising 20% of the data. We used this data set to train and test the nine different machine learning models and compare their performance. our objective here was to identify the best model in regard to accuracy and speed to be used in the CAPTAIN chatbot for real time bullying message identification.

Cyberbullying Data Analysis

We applied the bertopic, a topic modeling technique that leverages transformers to cluster words and/or sentences. BERTopic is a topic modeling technique that vectorizes text data into a low-dimensional space using word embeddings to discover semantically comparable words, sentences, or documents. BERTopic generates topic representations through a pre-trained transformer-based language model and then subsequently clusters the embeddings using a hierarchical clustering approach [20]. This component allowed us to summarize the bullying messages into different topics, providing valuable insight into which topics were most prevalent among cyberbullies. This information can be used to inform the chatbot and create more targeted responses to users who may be experiencing cyberbullying related to specific topics.

Chatbot Implementation

To develop the chatbot, we leveraged chatterbot [21], a machine-learning-based framework for conversational agent implementation. On top of the chatterbot framework, we developed the cyberbullying detection adaptor, which embedded our machine learning algorithms. We also built a question-answering library for cyberbullying prevention education based on reliable online resources such as Microsoft, UNICEF, and endcyberbullying.org. using the library, we implemented the education adaptor, which can initiate a conversation and answer users' questions about cyberbullying. To create a user-friendly interface, we used chat-bubble [22], a chatbot UI for the Web with JSON scripting. Chat-bubble allowed us to create a web-based interface where users can easily interact with CAPTAIN and assess information and Feedback.

Evaluation

We evaluated the performance of the machine learning models using measures including accuracy, precision, recall, and F-measures. We also considered the performance time of each model since speed is important to support real time bullying message detections.

We conducted an evaluation of the CAPTAIN chatbot using a set of scenario-based tests that encompassed various use cases, including bullying messages, questions regarding cyberbullying, cyberbullying intervention, and small talk. First, we tested the chatbot's ability to identify bullying messages and to respond appropriately with support and intervention. Second, we evaluated the chatbot's capacity to provide accurate and relevant answers to users' inquiries about cyberbullying. Third, we assessed the chatbot's ability to engage with users with small talk.

3 Result

We have collected 71,350 messages, 31,300 of which are classified as offensive. Table 1 shows the results of different machine learning algorithms for the automatic classification of bullying messages. For the machine learning classifier, SGD Classifier achieved the highest accuracy of 89.13%, F1 score of 88.93%, and the second highest precision of 94.46% with a relatively fast training time (0.13 S) and perdition time (0.0025 S). Decision Tree reached the highest recall of 87.09%, but its training time is slow. Adaptive boosting has the highest precision score of 95.68% however, the training time is relatively slow. Therefore, we used the SGD classifier in CAPTAIN for its highest accuracy and fast speed.

After running a topic modeling analysis, we discovered that the most frequently mentioned super-topics include race, gender, appearance, drinking/smoking, and financial status. We also identified the most commonly appeared terms within each category as well. Figure 2 shows these categories and their associated terms. These categories will potentially inform CAPTAIN for more tailored intervention. By understanding the specific topic that is mostly brought up in bullying messages, CAPTAIN can provide more targeted support to users who are experiencing bullying related to the specific topic. For example, if the bullying message judges someone based on their appearance (e.g., "you are fat and ugly"), the chatbot will provide intervention based on the topic (e.g., "it is uncool to judge people by their looks"; "Beauty can be subjective"). This can ultimately lead to more effective interventions and support for those who are being bullied (Table 1).

We further evaluated the functionality of the CAPTAIN chatbot with a list of competency scenarios, including bullying messages, questions regarding cyberbullying, cyberbullying intervention, and small talk. Figure 3 shows a demo of the CAPTAIN interface. When opening the interface, CAPTAIN first greeted the user and provided a self introduction. It can also answer questions on cyberbullying related topics. In addition, it can determine if a sentence could be flagged as bullying, inappropriate, or appropriate. As Fig. 3 shows, it successfully detected possible mean messages and provided feedback accordingly. Furthermore, CAPTAIN interacted with the users in a natural way. It can answer questions that are not necessarily about bullying such simple math, sports, etc. This can provide companionship to its users when needed.

Table 1. Comparison of 9 Different Machine Learning Algorithms for Cyberbullying Detection.

Algorithm	Accuracy	Precision	Recall	F1 Score	Prediction Time	Training Time
SGD Classifier	**89.13%**	**94.46%**	84.01%	**88.93%**	**0.0025 SEC**	**0.13 SEC**
Logistic Regression	88.74%	93.25%	84.45%	88.63%	0.0043 SEC	1.06 SEC
Random Forest	87.52%	89.07%	86.62%	87.83%	6.48 SEC	79.92 SEC
Bagging Classifier	87.44%	89.23%	86.24%	87.71%	0.55 SEC	70.33 SEC
Linear SVC	87.54%	90.14%	85.37%	87.69%	0.0033 SEC	3.70 SEC
Decision Tree	86.77%	87.40%	**87.09%**	87.25%	0.067 SEC	10.54 SEC
Adaptive Boosting	87.57%	95.68%	79.69%	86.96%	0.29 SEC	1.41 SEC
Multinomial Naïve Bayes	84.15%	84.28%	85.45%	84.86%	0.0052 SEC	0.0089 SEC
K-nearest Neighbors	76.49%	87.55%	63.84%	73.84%	127.78 SEC	0.0033 SEC

Fig. 2. Topic Clustering of Different Topics Based on Popularity

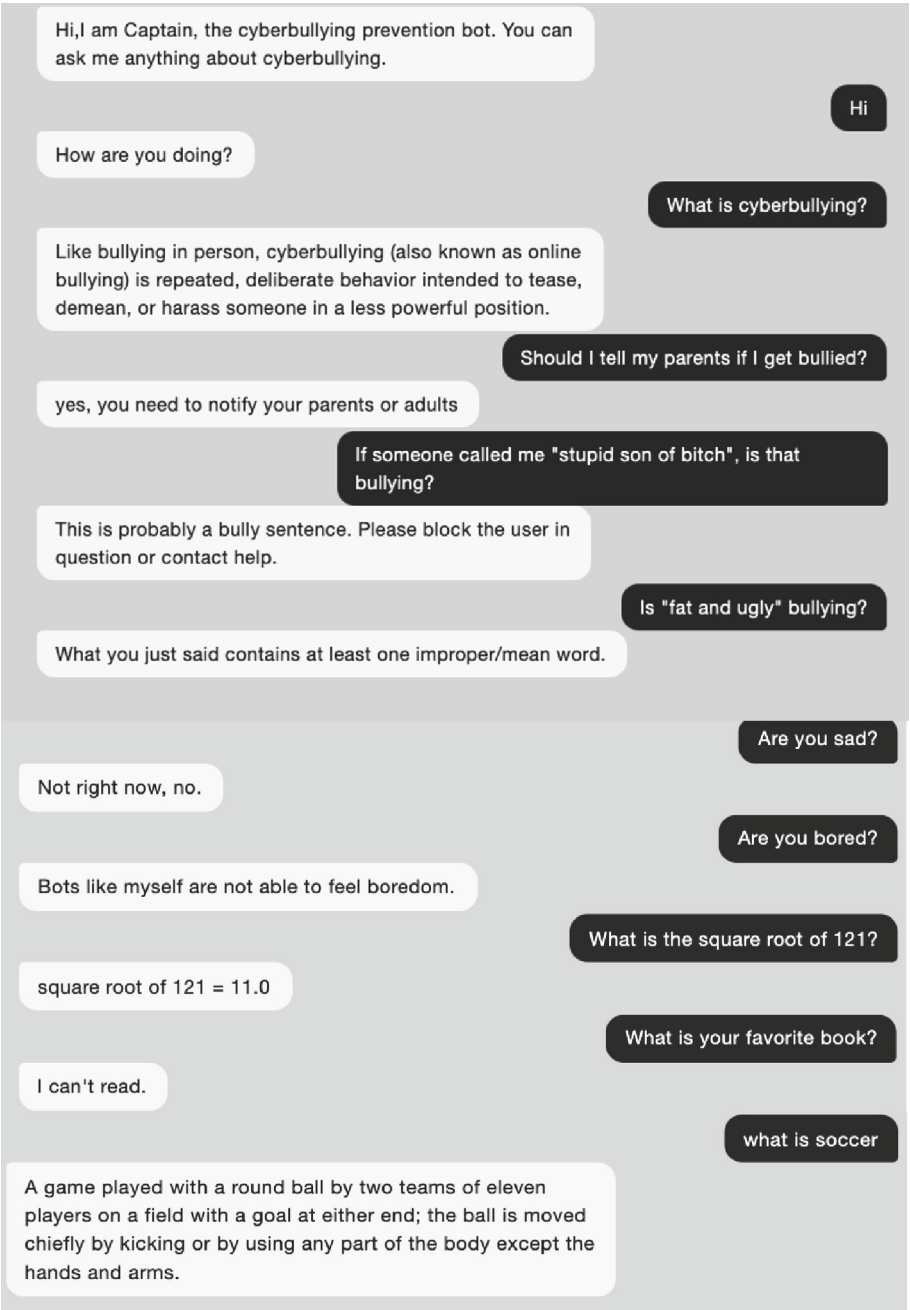

Fig. 3. Demonstration of Chatbot with interaction of Cyberbullying-related Topics and Small Talks

4 Discussion and Conclusion

In summary, CAPTAIN is an AI-based platform that aims to address the growing problem of cyberbullying. The main goal of this chatbot is to educate users about the negative effects of cyberbullying, provide tailored responses to those who may be experiencing cyberbullying, and automatically detect cyberbullying messages when needed. It uses a combination of natural language processing and machine learning techniques to classify messages as bullying or non-bullying. It also used a data-driven approach based on Topic Modeling to identify common topics. One of the unique features of CAPTAIN is that it offers personalized messages based on the machine learning classifier results. This allows the chatbot to provide targeted and relevant responses to users, which will potentially increase the effectiveness of the intervention. To the best of our knowledge, this is the first chatbot that targets AI-based personalized intervention for cyberbullying prevention.

The CAPTAIN system has the potential to reduce mental health problems caused by cyberbullying and enhance the overall online community. By providing targeted support to individuals who may be experiencing cyberbullying, CAPTAIN has the potential to reduce the negative effects of cyberbullying on mental health. The chatbot's personalized responses can help users feel heard and understood, provide timely interaction and information, which can in turn increase their resilience and empower them to seek additional support when needed. In addition, by reducing the incidence of cyberbullying, CAPTAIN has the potential to enhance the overall online community. A reduction in bullying behavior can create a safer and more positive online environment, where individuals feel free to express themselves and engage with others without fear of being targeted. This can lead to a more inclusive and supportive online community, where people can build meaningful connections and engage in constructive discussions.

Of course, there are still challenges associated with developing a chatbot for cyberbullying. Firstly, one of the main challenges is to develop a chatbot that can accurately identify and classify bullying behaviors. Every person has a different personality and may perceive bullying differently than others. This requires a sophisticated understanding of the language used in cyberbullying messages, as well as the context in which they are used. Secondly, Personalized messages are a critical component in the design of an effective intervention for cyberbullying. By understanding the underlying causes and motivations behind cyberbullying behavior, these messages can be tailored to address specific psychological and behavioral factors, leading to more successful outcomes. This is achieved by incorporating theoretical models from psychology and behavioral science into the design process. However, it is challenging for the AI algorithms to link to these models to dynamically generate responses. Thirdly, developing a chatbot that addresses cyberbullying raises a number of ethical considerations, such as privacy and confidentiality, particularly when dealing with sensitive and potentially traumatizing experiences.

In the future, the authors would like to further improve CAPTAIN in several directions. First, our machine learning cyberbullying data classification only focuses on binary boolean classification so far. We will extend the work and implement a multi-class classifier that can categorize messages into different topics. Second, although several existing papers discussed classifying messages into basic categories, there is no standard way

to reasonably categorize bullying messages. We will leverage the data-driven approach that uses Topic Modeling to identify common topics which can serve as categories to classify bullying messages. In addition, we will increase the chatbot knowledge base and add emotional support to the chatbot so it can connect with users better. In addition, we would also like to test feasibility, usability, and intervention effectiveness on teens to further improve CAPTAIN and deploy it to the public.

References

1. Nagata, J.M., et al.: Social epidemiology of early adolescent cyberbullying in the United States. Acad. Pediatr. **22**(8), 1287–1293 (2022). https://doi.org/10.1016/j.acap.2022.07.003
2. Patchin, J.: Summary of Our Cyberbullying Research (2007–2019). Cyberbullying Research Center (2019). https://cyberbullying.org/summary-of-our-cyberbullying-research
3. JAMA Netw. Open **4**(9), e2125860, September 2021. https://doi.org/10.1001/jamanetworkopen.2021.25860
4. Espelage, D.L., Hong, J.S.: Cyberbullying prevention and intervention efforts: current knowledge and future directions. Can. J. Psychiatry **62**(6), 374–380 (2017). https://doi.org/10.1177/0706743716684793
5. Smith, P.K., Bauman, S., Wong, D.: Challenges and opportunities of anti-bullying intervention programs. Int. J. Environ. Res. Public. Health **16**(10), 1810 (2019). https://doi.org/10.3390/ijerph16101810
6. Elsafoury, F., Katsigiannis, S., Pervez, Z., Ramzan, N.: When the timeline meets the pipeline: a survey on automated cyberbullying detection. IEEE Access **9**, 103541–103563 (2021). https://doi.org/10.1109/ACCESS.2021.3098979
7. Dadvar, M., Eckert, K.: Cyberbullying detection in social networks using deep learning based models; a reproducibility study (2018). https://doi.org/10.48550/ARXIV.1812.08046
8. Van Hee, C., et al.: Automatic detection of cyberbullying in social media text. PLoS ONE **13**(10), e0203794 (2018). https://doi.org/10.1371/journal.pone.0203794
9. Talpur, B.A., O'Sullivan, D.: Cyberbullying severity detection: a machine learning approach. PLoS ONE **15**(10), e0240924 (2020). https://doi.org/10.1371/journal.pone.0240924
10. Wilson, L., Marasoiu, M.: The development and use of chatbots in public health: scoping review. JMIR Hum. Factors **9**(4), e35882 (2022). https://doi.org/10.2196/35882
11. Deshpande, S., Warren, J.: Self-harm detection for mental health chatbots. In: Mantas, J. (eds.) Studies in Health Technology and Informatics, IOS Press (2021). https://doi.org/10.3233/SHTI210118
12. Dosovitsky, G., Bunge, E.: Development of a chatbot for depression: adolescent perceptions and recommendations. Child Adolesc. Ment. Health, p. camh.12627, December 2022. https://doi.org/10.1111/camh.12627
13. Greer, S., Ramo, D., Chang, Y.-J., Fu, M., Moskowitz, J., Haritatos, J.: Use of the Chatbot 'Vivibot' to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR MHealth UHealth **7**(10), e15018 (2019). https://doi.org/10.2196/15018
14. Weeks, R., et al.: Chatbot-delivered COVID-19 vaccine communication message preferences of young adults and public health workers in urban American communities: qualitative study. J. Med. Internet Res. **24**(7), e38418 (2022). https://doi.org/10.2196/38418
15. Nazareth, S., et al.: Hereditary Cancer risk using a genetic chatbot before routine care visits. Obstet. Gynecol. **138**(6), 860–870 (2021). https://doi.org/10.1097/AOG.0000000000004596
16. Gabrielli, S., Rizzi, S., Carbone, S., Donisi, V.: A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. JMIR Hum. Factors **7**(1), e16762 (2020). https://doi.org/10.2196/16762

17. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop, San Diego, California, pp. 88–93 (2016). https://doi.org/10.18653/v1/N16-2013
18. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. arXiv 11 March 2017. Accessed 14 Jan 2023. http://arxiv.org/abs/1703.04009
19. Tweets Dataset for Detection of Cyber-Trolls. https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls
20. Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv, March 11 2022. Accessed 28 December 2022. http://arxiv.org/abs/2203.05794
21. ChatterBot (2022). https://pypi.org/project/ChatterBot/
22. Chat-Bubble: Simple chatbot UI for the Web with JSON scripting. https://github.com/dmitrizzle/chat-bubble