TOWARDS A CLEAN SOCIAL ENVIRONMENT USING MACHINE LEARNING

STUDENT NAME:	Princewill Okosun Ehimare
STUDENT NUMBER:	201569621
COURSE NAME:	Master's in Advanced Computer Science
DEPARTMENT:	Department of Computer Science
COURSE CODE:	COMP 702
SUPERVISOR:	Dr. Jeffrey Ray
DATE OF SUBMISSION:	08-07-2022

CONTENTS

INTRODUCTION

Great strides in computer science have always led to simpler living, however, these advancements did not come without much-needed intervention through people who raised an equally much needed awareness of the problems at hand. Through this synergy of problem identification and problem resolution we have created a more convenient world. We cannot stop now. The internet has brought about connectedness and has been a key element in the advancement of technology, however, we are at a junction where the rate at which toxic data being uploaded on the internet is surpassing our ability to moderate toxic altercations [1][2]. Human moderators that regulate toxic content on the internet are having to endure challenging workloads and being diagnosed with mental health illness such as post-traumatic stress disorder and chronic anxiety. The normalization of toxicity on the internet is autocatalytic as it originates more toxicity and thus demands abrupt action [3]. This project aims to raise awareness of these alarming levels of toxicity and hate speech, highlight the impact of this toxicity on the mental health of internet users and moderators and put to light the current measures being taken to mitigate these trying state of affairs. This project would also address how Artificial Intelligence in the form of Natural Language Processing (NLP) is impacting this alarming development through the design and comparative analysis of several machine learning models to underline the progress of NLP research. Addressing real-world topics such as mental health requires the highest level of academic honesty and integrity. As such, this project would abide to the highest ethical standards to avoid unexpected negative consequences. Finally, it will aim to accomplish the objective of accentuating the importance of research on this problem and most importantly stress the need for creating a clean social environment on the internet using machine learning.

BACKGROUND STATEMENT

The internet has evidently championed a collaboration that has facilitated advancements in technology and daily human living. It has brought all of humanity closer together, improved business-to-business interaction and has provided a platform for free speech and intellectual discussions. However, due to the faux presence of an authoritative governing body there are ongoing malicious activities with negative real-world consequences such as internet fraud and network insecurity to name a few. One prevailing issue that is now almost trivialised is the level of toxicity and hate the internet breeds. To briefly highlight this, according to a Pew Research Centre survey [1] four in ten internet users have experienced online harassment with 45% of these people experiencing heightened levels of hate that could peradventure lead to mental breakdown. Moreover, the internet was designed to be egalitarian in nature, yet 41% of women who have experienced the brunt of the internet's toxicity say that their encounter had them feeling their physical safety was threatened with 56% claiming this resulted in a limited ability to carry out daily tasks [4][5]. Despite these alarming levels of toxicity there is very little research regarding the general problem of internet toxicity, this project aims to change that as it not only pays attention to possible solutions but sheds light on the menacing consequence of understating the problem at hand.

There is no clear definition of what toxicity is, even more so what entails online harassment. Online abuse can vary between minor altercations such as name calling and trolling and can grow to ugly extremes like violent threats, sexist remarks, racism and hate speech. All of these fall under the term toxic which as defined by the Oxford dictionary is anything "very harmful or unpleasant in a pervasive or insidious way" [6]. The insidious nature of toxic declarative remarks by internet users has enabled society undermine the mental health effects and pervasiveness of online bullying. Such a similar naïveté regarding the consequence of declarative remarks is possibly why the Telecommunications Act of 1996 [7] leaves the monitoring and regulation of online interactions to technology companies and doesn't hold them liable for the consequences of the "content" posted on their platforms. The pseudonymous nature of the internet is also a contributing factor to the state of toxic remarks, not surprisingly over half of online abuse victims claim to not know the identity of persons involved in their incident [1]. Interestingly, in 2007 regulators of the South Korean Government mandated all internet users within its geography to register their identity in hopes to reduce the number of internet comments after several suicides were found out to have occurred as a result of online abuse [8]. The result of this was "almost negligible change" with only a 0.9% decrease in online posts and little to no decrease in hateful comments and insults [9]. Such problems with ambiguity in definitions and anonymity make regulating internet comments,

cyberbullying and internet hate a difficult problem to address and has led to a prevalent toxic experience for internet users.

Although 76% of internet users assert the internet has been a good thing for society, 73% of adult internet users claim to have witnessed online abuse or cyberbullying of some sorts [1]. A large majority of internet users use the internet to voice their opinions and create relationships with 67% of internet users claiming the internet has played a key role in improving their relationships [10]. Unfortunately, 40% of them claim they have been a victim of online abuse with about half of this demographic experiencing more alarming levels of online abuse [1]. Social media platforms are a hotbed for online abuse with about two-thirds of internet users who have been a victim to online abuse claiming they experienced their abuse on a social media platform [1]. Women have a higher chance of experiencing online harassment on social media platforms with 26% of the victims polling their harassment involved threats of sexual and physical assault [4]. More than half of the female victims of online abuse on social media claim their bouts led to panic attacks and immense anxiety episodes [4]. Self-evidently the toxicity on the internet has far reaching consequences even outside the digital sphere with users resulting to direct confrontation or law enforcement in addressing these incidents [1]. Furthermore, the lingering effects of a racist or sexist encounter has effects on not just the persons but the communities of these persons as communal differences often lead to vicious online bouts. Although governments and social media giants have made efforts to address this issue, their efforts fall short of what is needed to create a clean social environment on the internet [8][11]. Current approaches such as the use of regulations and human moderators have fallen short of expected results, it is unlikely change would occur unless adequate awareness regarding the problem is raised.

Human moderation involves the manual moderation of internet content at times with the assistance of NLP computational models to flag offensive content. Natural Language Processing (NLP) involves the study and design of machine learning algorithms that are able to carry out an analysis of natural languages, facilitate linguistic communications and execute textual or speech processing [12]. Sentiment analysis is a subtask of NLP aimed at obtaining categorical sentiments from generalized textual data [13]. These technologies provide a means to reduce the number of content human moderators must survey. Notwithstanding, the sheer amount of content uploaded to the internet on a daily basis still surpasses the moderators' capabilities substantially [2]. Moreover, human moderators working for social media platforms spend their day buried under an immense workload of toxic content that severely affects their mental health [11][14][15]. Taking the Twitter social media platform as our test case it is observed that there are about 500 million posts uploaded on a daily basis [16]. The current system in place follows a post-moderation approach where the harmfulness of a post results in it being flagged, reported and scrutinized by an AI system which deletes the post or flags it for human review [17]. Twitter employs about 1500

moderators to moderate this daily stream of content while Meta's Facebook hires 10 times more moderators for about 10 times the daily content of the Twitter social networking site [2][18]. These numbers show that the current number of human moderators employed by these organizations does not suffice for the monitoring and regulation of posts on the internet as moderators are constantly swarmed by more content [14][15]. Breakthroughs in AI have led to more efficient machine learning algorithms which in turn have improved their capabilities to mimic and complete human tasks such as scrutinizing large datasets of social media content more efficiently than human beings. One breakthrough is Google's SOTA (state-of-the-art) NLP model BERT (Bidirectional Encoder Representations from Transformers) which produces SOTA results in various NLP tasks. BERT is pre-trained on almost two billion words taken from the English Wikipedia Dataset [19][20] and applies bidirectional, functioning in two directions, training of the novel Transformer architecture to language modelling as opposed to left-to-right or combined left-to-right right-to-left training approaches [19]. This project aims to show that increased research in NLP & sentiment analysis would bolster efforts towards a clean social environment on the internet as moderators can work with more efficient artificial intelligence systems in sorting through the large heap of online content uploaded every day.

Machine learning systems rely on large datasets to enable them learn accurate representations from data. Through being able to learn the features that make up toxic posts in training scenarios they can perform pre-moderation, moderation before content upload, and post-moderation. Different techniques of feature extraction and model architecture have led to varying results. Kouloumpis et al. [21], in an effort to facilitate Twitter sentiment analysis research, uses part-of-speech features and a sentiment lexicon. They also exclude the emojis in the textual data from removal to observe the impact of creative language on the classifiers. It was discovered that using part-of-speech features decreased the performance of the classifiers while creative language was deemed beneficial, however, its contribution was only marginal [22]. Agarwal et al. [23] carried out research on sentiment analysis of Twitter data and were able to categorize tweets using the sentimental polarity of the words used in the tweets. Agarwal et al. in exploring the use of tree kernels discovered tree kernels performed at the same level as the SOTA unigram models at the time [23] shedding light for further development. This project would carry out a comparative analysis of various categorical models on toxic twitter data, thirty-two thousand in number, in a bid to highlight that there is still room for further development in the use of AI for content moderation. Nonetheless, we must recognize the limitations of categorizing textual content as speech is innately contextual. It is also important to recognize that data analysis carried out by information systems is not faultless.

OBJECTIVES AND AIMS

Overall Objective

Research that highlights not just the capability of machines to carry out sentiment analysis on textual data but the vitality of these technologies is lacking. Following this, the objective of this project is to raise awareness of the prevailing levels of toxicity in present day internet spaces and social media primarily. This project would provide a thorough review on the levels of online abuse on the internet. Using statistical details, this project aims to show how frequently users are exposed to varying levels of toxicity and the mental health consequences of these engagements. It builds machine learning models capable of performing sentiment analysis and compares the performance metrics against breakthrough techniques and methodologies in a chronologic fashion. It also carries out data analysis on the toxic dataset being used to generate insights on the data and identify characteristics of toxic posts.

Specific Aims

- Raise awareness of the volume of hate speech on the internet.
- Highlight the consequential impact of internet hate and online abuse.
- Bring to light the shortcomings of human moderators and their mental health battles.
- Encourage research in sentiment analysis for internet moderation.
- Collect and pre-process toxic Twitter data.
- Build a word2vec or BERT model for feature extraction.
- Identify and build intelligent computational models to categorize hate speech.
- Compare model performance using evaluative metrics.
- Identify best performing model on real-world data providing detailed explanations.
- Compare best performing model with novel SOTA models.
- Promote the need for a clean social environment on the internet.

To achieve this objective, research would be carried out on the state of the internet and its impact on different demographics. The consequence of varying levels of hate speech and the lingering effects it leaves on its victims will also be dissected. Various models would also be built to classify hate speech using sentiment analysis and these models would be compared against each other using several evaluative metrics such as F-scores and Receiving Operating Characteristics (ROC). Finally, a chronological comparison of the technologies addressing this issue is shown to highlight the possibility of more efficient results, the need for elevated research on the toxicity of the internet, and the use of NLP as a means to alleviate it.

METHODOLOGY

As this project categorises hate speech, the data used is limited to textual data. The dataset [24] is obtained from “Analytics Vidyha” and contains more than thirty thousand labelled tweets. Tweets in the dataset contain hashtags and special characters however usernames have been redacted and replaced with the “@user” tag.

The dataset would undergo pre-processing which is the initial phase for NLP tasks. This pre-processing would involve cleaning the dataset in order to retain salient features that would assist in training the model to predict accurately between hate speech and regular tweets. It also prepares the textual data for mining which involves extracting meaningful information from the text disregarding noisy data which could contain features that can skew expected results [25]. Possible noise in our data includes emoticons, currency figures, numbers, punctuation marks, special characters, non-dictionary words and slangs. These do not have standard context in general conversations and can skew the machine learning algorithms understanding of text sentiment [25]. We utilise pythons Natural Language Toolkit (NLTK) library in this pre-processing stage of our data. NLTK is a powerful library used in the pre-processing of data for NLP tasks. It also comes with an in-built sentiment analyser which outputs the polarity and subjectivity of textual data [26]. We use NLTK to tokenize our data in order to separate sentences into distinct words. We also use NLTK to return words to their root form and remove contractions using one of the many NLTK word-stemming packages. We lemmatize our words using NLTK’s WordNetLemmarizer package as to allow for a morphological anatomy of our tokenized sentences. We also get rid of stop words, frequently occurring words in everyday conversations, as well as short words like “oh” and “omw”. Finally, NLTK is used to generate word frequency count in order to facilitate data analysis that would help in gaining profound insights on our data. To visualise these insights, wordclouds [27] would be used in creating visual representations of relevant words used in hate speech and normal speech. After successful pre-processing and data analysis, feature selection/extraction is carried out in order for our pre-processed data to be useful.

After the data is pre-processed, features are extracted from the textual data in numerical representations as the machine learning algorithms to be used learn patterns from numeric data. The text after data pre-processing makes up the feature space therefore our data pre-processing phase would affect the performance of our models. Feature Selection is carried out using either Word2vec [28] or GloVe [29] as they both deliver accurate and consistent results in the sentiment analysis of NLP tasks [29][30]. The clean data is split into train-test-validation fragments after which we train our machine learning models using a k-fold cross-validation approach due to the variance in the number of hateful tweets and clean speech. Following this, these models are run on test data and the results are compared against their F-scores or using a ROC graph which

performs as an excellent evaluative technique in binary classification tasks [31]. We then choose the best model and check it against our baseline model which would be a simple logistic regression model that uses the bag-of-words model for feature selection. In doing this, the aim is to highlight the advancements and current capability of AI & NLP in content moderation. Furthermore, the project would in turn highlight the importance of elevated research in using NLP for content moderation whilst raising awareness that facilitates research on the prevalence of toxicity on the internet and social media platforms at large. This project would also carry out a literature comparative analysis on its best chosen model against avant-garde deep learning approaches such as LSTM & CNN to garner support for research in the use of advanced techniques for content moderation on social media platforms and the internet.

ETHICS

In addressing a societal issue strongly related to ethics and morality, this project aims to meet the highest computer science ethical standards. Computer ethics govern hardware and software use and ensure that computer science research and developments align with societal standards.

The noxious nature of the dataset used in this project demands elevated levels of protection to prevent unwarranted access to the data as there could be damaging consequences to a person's present mental state.
In a bid to keep in line with ethical practices the dataset is:

- Protected from unwarranted access.
- Abstracted away from the program code implementation.
- Opened at timed intervals not permitting continuous absorption of toxic information

This project also addresses real-world mental health related topics and their impact on the mental health of everyday persons. Since this research could impact the lives of these entities it aims to be clear and honest through providing figures and facts even if they are of a grim nature.
Therefore, to ensure the highest ethical consideration it would:

- Hide all identities of accused and affected persons.
- Ensure elevated levels of integrity and honesty in the research.
- Underscore in thorough, complete, and accurate scientific detail the adverse mental health and physical ramifications of victims of internet abuse.

In a world filled with bias and fabricated truths, this project aims to advocate for clarity and honesty through the truthfulness of its work and as such would clearly feature the advantages and disadvantages of its significant suggestions and conclusions. As such it aims to not just raise awareness for clean living on the internet but in every aspect of scientific research as well.

PROJECT PLAN

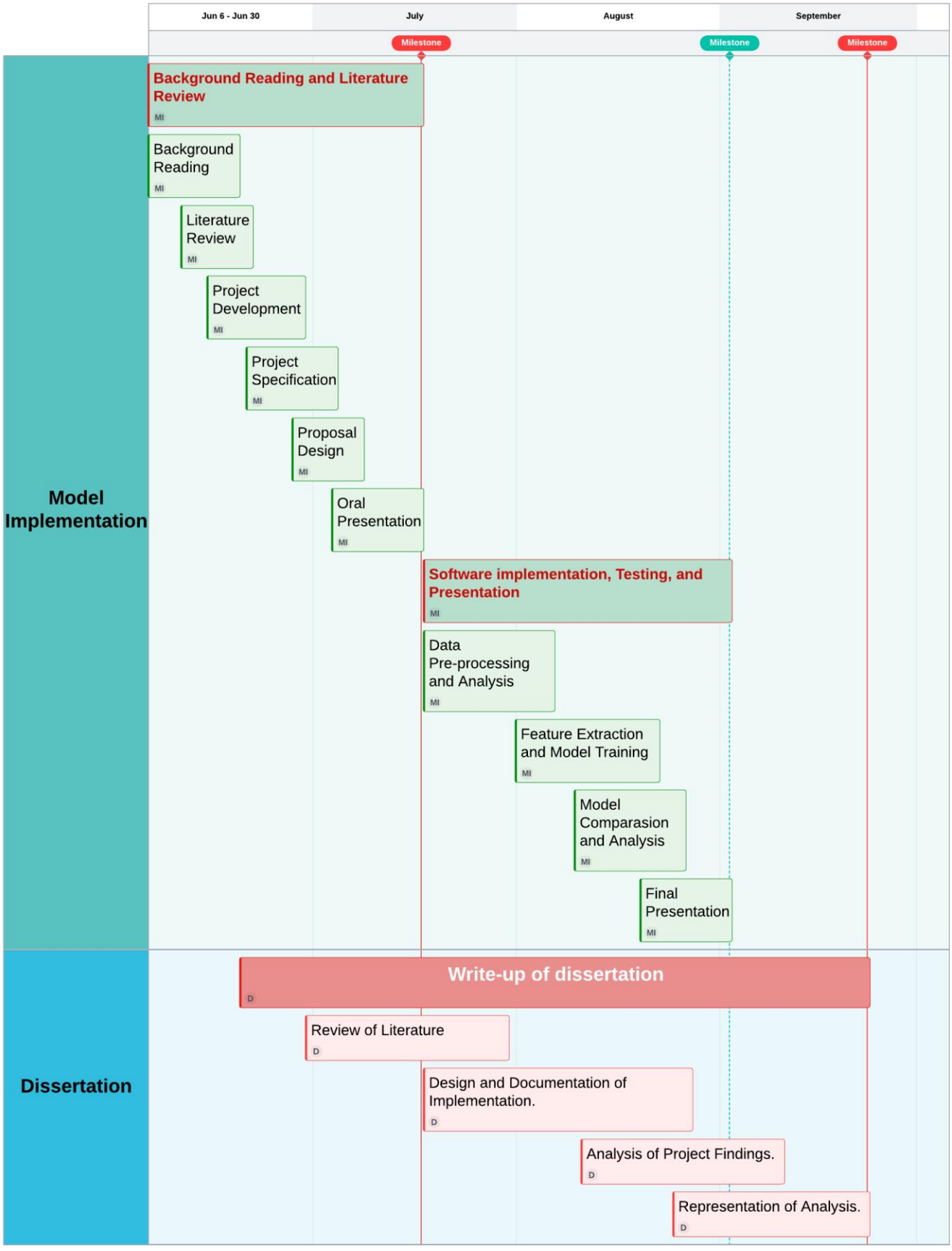

Figure 1. Project Plan Gantt Chart

RISKS AND CONTINGENCY

To ensure successful completion of this project it is paramount that all risks are identified and assessed. Planning to mitigate these risks is equally crucial, however, in the event they occur contingency plans are made available to lessen the impact of these risks.

For this project, the primary risk lies in the adverse effects from assimilating toxic data. All stakeholders involved in the implementation aspect of this project would be exposed to a dataset containing hate speech, racism and sexism which can affect their mental health. The risk of a stakeholder experiencing distress after reading multiple racist sexist and homophobic remarks is high and its mental impact even very higher. Following from this, throughout this project only authorised parties within the project evaluation committee would have access to this dataset as it would be protected. The dataset would be protected in a zipped folder to reduce the possibility of unknown persons being affected by its content. To reduce any negative mental health impact in the event of exposure, platforms like the NHS mental health services [32] would be utilised for medical assessment and counselling.

The exposure of a risk represents its threat and severity in the event of its occurrence. We calculate this exposure as the probability of the risk occurring multiplied by its impact on the project in the event it occurs. We represent the probability and impact of risks occurring as follows:

- If the probability/impact of a risk occurring is **high**, we assign it a probability value above 0.6.
- If the probability/impact of a risk occurring is **medium**, we assign it a probability value between 0.3 and 0.6.
- If the probability/impact of a risk occurring is **low**, we assign it a probability value below 0.3 but not below 0.

All risks with an exposure index over 0.5 are monitored and tracked through the entirety of this project. The risks considered in this project are stated below:

- Imbalance in dataset leading to inferior performance.
- Non-linear separability of dataset.
- Computer system failure and risk of program failure during the presentation.
- Implementation schedule being inaccurate.
- Ill health and mental health safety of myself and project supervisors.
- Requirement of paid resources.
- Poor or inadequate flow of communication during project implementation.

The following table addresses the consequences, probability, impact, exposure, mitigating actions and contingency plan for each identified risk.

Risk	Consequence	Probability	Impact	Exposure	Mitigation	Contingency
Imbalance in dataset leading to inferior performance.	Poor performance because of inaccurate predictions.	0.4	0.5	0.2	The use of k-cross validation in training the models.	Obtain or create a different dataset from a social media platform.
Non-linear separability of dataset leading to inferior performance.	SVM and linear models would perform poorly.	0.5	0.3	0.1	Basis function representations can be utilised in making data linearly separable [33].	Use alternative techniques for model implementation such as neural networks.
Computer system failure and risk of program failure during the presentation	Incomplete project, loss of information and inaccurate depiction of findings.	0.1	0.7	0.0	All files are backed up on cloud drives and program implementation is tested before presentation.	Backup files and program implementation are used.
Incorrect software implementation schedule.	Failure to deliver project in due time.	0.3	0.7	0.2	Ensure deliverables are completed before deadline and evaluate schedule as project progresses.	Request an extension for submission deadline.
Ailing physical or mental health.	Poor completion of project and/or adverse mental health issues.	0.7	0.9	0.6	Protected and limited access to contents of the dataset.	Use of NHS [32] services for medical screening and counselling.
Requirement of high-cost pay-to-use online resources.	Possible change in technologies used and implementation architecture.	0.7	0.6	0.4	Extensions to project scope are kept at a minimum to avoid external costs.	Platforms that provide cheap or free alternatives would be explored.
Poor or inadequate flow of communication during project implementation.	Possible miscommunication leading to unsuccessful project completion.	0.1	0.7	0.0	Regularly scheduled meetings and information sharing.	Restructure of meeting schedule. Adherence to project plan.

Table 1. Risks and Contingency Table

REFERENCES

[1] M. DUGGAN, “Pew Research Center Online Harassment,” Pew Research Center, Washington, DC, 2014.

[2] P. M. BARRETT, “Who Moderates the Social Media Giants? A Call to End Outsourcing,” NYU Stern Center for Business and Human Rights, New York, 2020.

[3] C. Vinopal, “Why online communities that breed hate and violence are so hard to control,” PBS, 7 August 2019. [Online]. Available: https://www.pbs.org/newshour/nation/why-online-communities-that-breed-hate-and-violence-are-so-hard-to-control. [Accessed 2 July 2022].

[4] A. Dhrodia, “Unsocial media: A toxic place for women,” IPPR Progressive Review , 2018.

[5] A. International, “ Ipsos MORI survey for Amnesty International on online abuse and harassment.,” Amnesty International, 2017.

[6] The Oxford English Dictionary, Oxford : Oxford University Press, 2010.

[7] C. L. School, “47 U.S. Code § 230 - Protection for private blocking and screening of offensive material,” [Online]. Available: https://www.law.cornell.edu/uscode/text/47/230. [Accessed 4 July 2022].

[8] Catalysts for Collaboration, “Case study: South Korea’s Internet Identity Verification System,” Catalysts for Collaboration, [Online]. Available: https://catalystsforcollaboration.org/case-study-internet-identity-verification-system/. [Accessed 27 June 2022].

[9] M. Szydlowski, “The science of toxic online comments,” Columbia Daily Tribune, [Online]. Available: https://eu.columbiatribune.com/story/news/2020/09/30/science-of-toxic-online-comments/114167254/. [Accessed 27 June 2022].

[10] S. FOX and L. RAINIE, “The Web at 25 in the U.S.,” Pew Research Center, Washington, DC, 2014.

[11] C. Newton, “BODIES IN SEATS,” The Verge, 19 June 2019. [Online]. Available: https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa. [Accessed 10 June 2022].

[12] G. Chowdhury, “Natural language processing,” *Annual Review of Information Science and Technology,* pp. 51-89, 2003.

[13] W. Medhat, A. Hassan and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal,* vol. 5, no. 4, pp. 1093-1113, 2015.

[14] P. Esposito, Director, *Inside the traumatic life of a Facebook moderator.* [Film]. North America: The Verge.

[15] *Google and YouTube moderators speak out.* [Film]. North America: The Verge.

[16] D. Sayce, “The Number of tweets per day in 2020,” David Sayce Digital Marketing, 2021. [Online]. Available: https://www.dsayce.com/social-media/tweets-day/. [Accessed 12 June 2022].

[17] Appen, “Leveraging AI and Machine Learning for Content Moderation,” 2 April 2021. [Online].

Available: https://appen.com/blog/content-moderation/. [Accessed 23 June 2022].

[18] Wishpond, “41 Up-to-Date Facebook Facts and Stats,” Wishpond, [Online]. Available: https://blog.wishpond.com/post/115675435109/40-up-to-date-facebook-facts-and-stats. [Accessed 22 June 2022].

[19] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018.

[20] Wikipedia, “BERT (language model),” Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/BERT_(language_model). [Accessed 28 June 2022].

[21] E. Kouloumpis, T. Wilson and J. Moore, “Twitter sentiment analysis: The good the bad and the omg!,” 2011.

[22] A. Yang, J. Zhang, L. Pan and Y. Xiang, “Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination,” Ang Yang, 2015.

[23] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, “Sentiment analysis of twitter data,” *Proceedings of the Workshop on Languages in Social Media,* pp. 30-38, 2011.

[24] https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/, “Twitter Sentiment Analysis,” Analytics Vidhya, 2018.

[25] A. Schofield, M. Magnusson and D. Mimno, “Pulling Out the Stops: Rethinking Stopword Removal for Topic Models”.

[26] NLTK Project, “NLTK,” NLTK Project, 2022. [Online]. Available: https://www.nltk.org/api/nltk.html. [Accessed 3 July 2022].

[27] A. Mueller, “WordCloud for Python documentation,” 2020.

[28] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient Estimation of Word Representations in,” California, 2013.

[29] R. S. C. D. M. Jeffrey Pennington, “GloVe: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[30] S. Al-Saqqa and A. Awajan, “The Use of Word2vec Model in Sentiment Analysis: A Survey”.

[31] T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.,” *Journal of Pattern Recognition Letters – Special issue in Roc analysis in pattern recognition archive,* 2006.

[32] NHS, “Mental health services,” [Online]. Available: https://www.nhs.uk/nhs-services/mental-health-services/. [Accessed 22 June 2022].

[33] R. Grosse, “Linear Classification,” [Online]. Available: https://www.cs.toronto.edu/~rgrosse/courses/csc321_2018/readings/L03%20Linear%20Classifiers.pdf. [Accessed 4 July 2022].