Aug 10th, 12:00 AM

A Personalized Harmful Information Detection System Based on User Portraits

Yuming Li
University of Auckland, yuming.li@auckland.ac.nz

Johnny Chan
University of Auckland, jh.chan@auckland.ac.nz

Gabrielle Peko
University of Auckland, g.peko@auckland.ac.nz

David Sundaram
University of Auckland, d.sundaram@auckland.ac.nz

A Personalized Harmful Information Detection System Based on User Portraits

Completed Research

Yuming Li
University of Auckland
yuming.li@auckland.ac.nz

Johnny Chan
University of Auckland
jh.chan@auckland.ac.nz

Gabrielle Peko
University of Auckland
g.peko@auckland.ac.nz

David Sundaram
University of Auckland
d.sundaram@auckland.ac.nz

Abstract

In this information age harmful information and cyberbullying on the Internet have attracted much attention. Most of the existing research on harmful information detection only make breakthroughs in algorithmic efficiency and accuracy, while regarding Internet users as an average whole. However, there could be substantial differences among the users, which to a certain extent, account for the minorities in marginalized groups. Marginalized groups are more likely to be victims of cyberattacks with harmful information, and these cyberattacks often cause more harm to them than others, which creates a vicious cycle. In this paper, we propose a framework for a personalized harmful information detection system based on user portraits. This detection system uses different thresholds to filter harmful information according to a user's personality characteristics, so that we could maximize both the Internet experience and protection of the vulnerable minorities from harmful information.

Keywords

Harmful information detection, sentiment analysis, cyberbullying detection, user portrait construction.

Introduction

With the development of Information and Communication Technologies (ICT) and the popularization of social media, information in the form of text, image, audio and video, published by rapidly growing users and institutions on social media and other platforms is accumulating by the second. In this era of big data, information on the Internet supports everything including the economy, society, entertainment and culture. Demchenko et al. (2014) proposed the 5 V characteristics of big data: velocity, volume, value, variety, and veracity. The extreme expansion of information and the convenience of the Internet have prompted people to communicate and share information widely on the Internet. However, the convenience and ubiquity of the Internet also has many side effects, such as the spread of harmful speech and cyberbullying. There will be some users who are irrational or purely malicious when dealing with certain sensitive topics. In extreme cases, they will have extreme violent speech, cyberbullying or even doxing. When these behaviors are large-scale, they not only cause psychological and physical harm to the victims but they also destroy the free environment of the Internet (Smith et al., 2006).

Therefore, the detection and filtering of harmful speech, toxic information, and potential cyberbullying has gradually become a hot topic in sociology, psychology, and communication studies (Keipi et al., 2016). And an automatic harmful information detection system is of great significance to purify the network environment and prevent potential hazards. Existing research usually focuses on the improvement of algorithms and optimization of functions (Ptaszynski et al., 2019; Pawar et al., 2018; Iwendi et al., 2020; Perera et al., 2021), while considering Internet users as an average whole. The framework of most of the existing harmful information detection systems is shown in Figure 1.

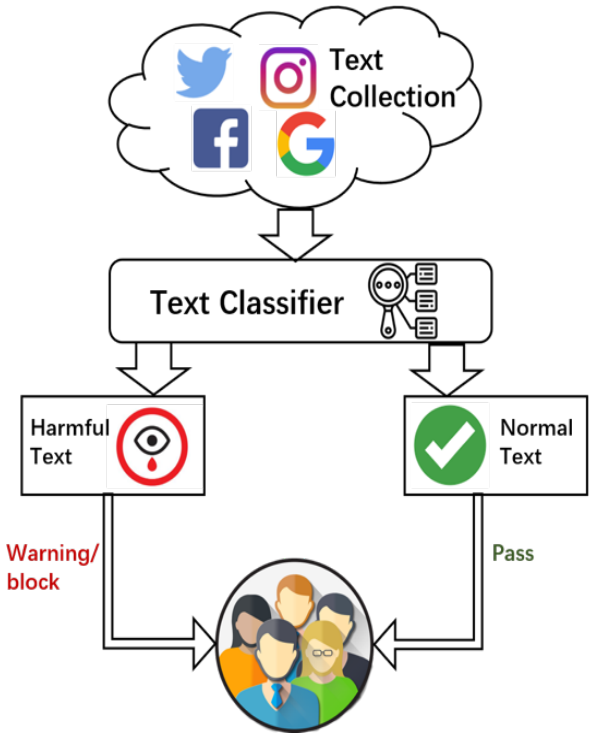

Figure 1. Framework of existing harmful information detection systems

The information collected from online or designated platforms is passed into the text classifier, and the method based on natural language processing technology is usually used to divide the text into harmful text and general text. Harmful texts are warned or blocked, while general texts are passed directly to users.

According to the data report by the Cyberbullying Research Center from 2021 (as shown in Figure 2), cyberbullying has different probabilities of occurrence among different genders, sexual orientations and races, and the degree of trauma to different groups is also different (Hinduja, 2021). Therefore, considering Internet users as an average whole can often fail to protect groups that are more vulnerable to malicious attacks. We need a relatively personalized harmful information detection system that adopts different standards for different groups, and ultimately maximizes the protection of the physical and mental health of marginalized individuals.

Using the design science research methodology (Hevner et al., 2004), we propose a personalized harmful text detection framework (shown in Figure 3) which sets a different threshold of classifier according to the traits of different users. For marginalized and pessimistic users, the framework deploys stricter filters to maximize the removal of potentially harmful information. For positive and optimistic users, more tolerant filters are used to maximize access to incoming information. This design aims to protect vulnerable groups while ensuring maximum freedom for others.

Literature Review

International attention to harmful information on the Internet can be traced back to the Safer Internet Action Plan issued by the European Union in 1999. The plan focuses on the governance of illegal and harmful Internet content, including text and video, and makes it clear the need to combat racism and spam. Over the past two decades, research on the governance of online harmful information and the control of cyberbullying has emerged, especially on the use of automated detection methods (Sugandhi et al., 2015). Among them, Sharma et al. (2018) proposed an ontology-based harmful speech classification model, which subdivided harmful speech into three levels for further detection, and they achieved an accuracy rate of over 76% on a Twitter dataset. Nurrahmi et al. (2018) added user credibility as a parameter based on traditional text classification to label and classify harmful content and potential cyberbullying from Twitter. The

innovation of this study is the credibility analysis of user, which classifies them into normal users, bullying actors, harmful bullying actors and prospective bullying actors based on the likelihood of abnormal behaviors from their tweets. Pawar et al. (2018) designed a distributed cyberbullying detection system, using the principles of natural language processing, machine learning and distributed systems to create a prototype. It guarantees performance while ensuring accuracy, but it is slightly insufficient in terms of data volume. Ptaszynski et al. (2019) proposed an Internet malicious content detection method based on brute-force search algorithms, and implemented an Android smartphone application with feasible results, but it still needs to be optimized in terms of time efficiency. Four neural network models were used by Iwendi et al. (2020) to detect potential cyberbullying texts resulting in a proposed cyberbullying detection scheme based on a deep learning framework. Most recently, Perera and Fernando (2021) proposed a system for automatic detection and prevention of cyberbullying on social media platforms, which models and categorizes the topic and sentiment attributes of posts with over 75% accuracy. In general, most of these studies focus on algorithmic accuracy and performance, and they consider Internet users as an average whole. Yet, studies in psychology and sociology state that some minority groups are more vulnerable to malicious attacks and cyberbullying than others (Llorent et al., 2016). To acknowledge that, we need to consider group differences among Internet users when we implement automated harmful information detection systems, and that highlights a gap in existing research.

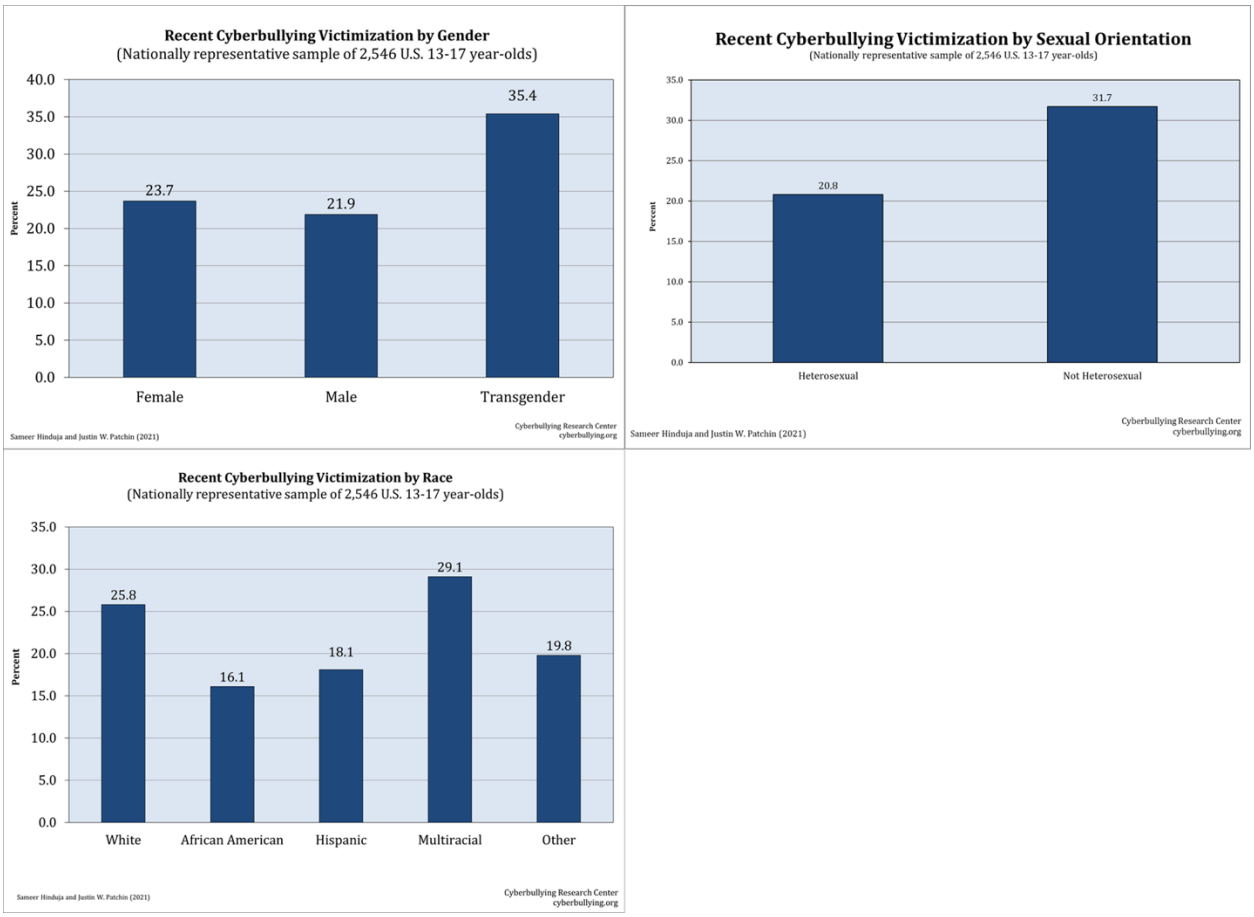

Figure 2. Recent cyberbullying victimization by gender, sexual orientation and race (Hinduja, 2021)

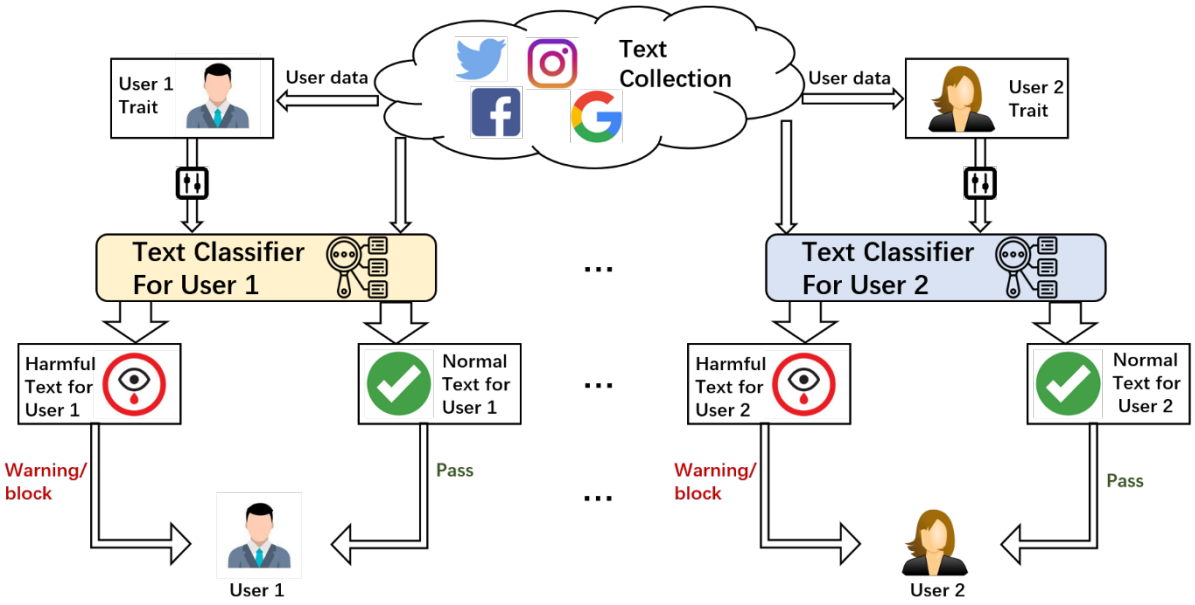

Figure 3. Personalized harmful information detection framework

Databases and datasets on harmful information are scarce. The most popular and widely used one is the Toxic Comment Classification Challenge dataset released by Jigsaw on the Kaggle platform, which contains 6 types of toxic text from Wikipedia comments. Other datasets include the Twitter Data Set provided by Waseem and Hovy (2016) in the ACL conference evaluation task, where the data is collected using Twitter API, and the Twitter Bullying Traces dataset proposed by Sui (2015).

A Personalized Harmful Information Detection System

The proposed personalized harmful information detection system is shown in Figure 4, and it is consisted of an user trait construction module on the left, and a harmful information detection module on the right. The user trait construction module determines the characteristics of a user (optimistic/neutral/pessimistic) through emotion analysis and behavioral traits classification from the user portrait construction, and it passes the calculated user trait as a parameter to the harmful text filter. In the harmful information detection module, we feed the pre-processed online text to the harmful text filter, and use the calculated user trait to adjust the threshold of the classifier for personalized filtering of harmful information.

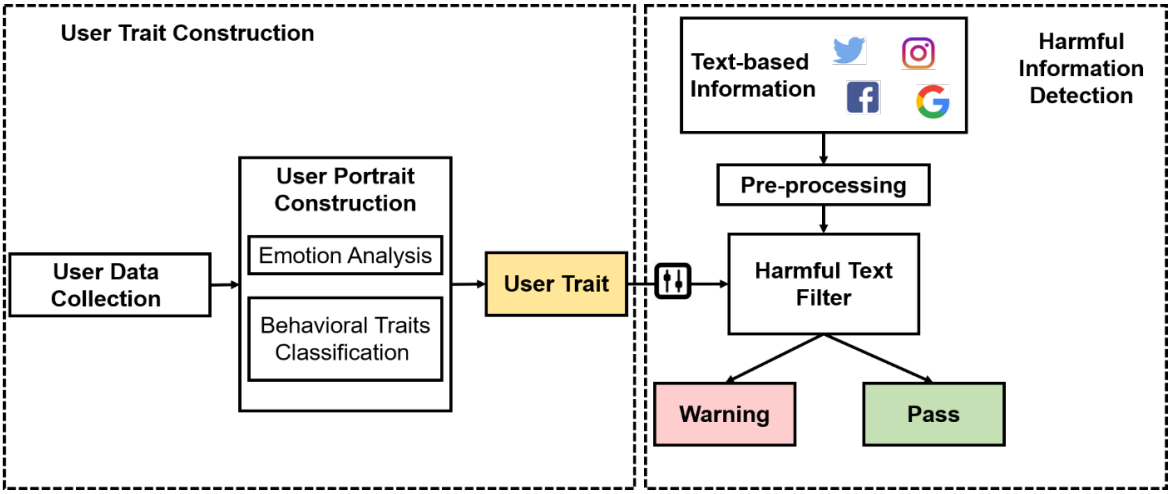

Figure 4. A personalized harmful information detection system

User Portrait Construction

The detailed process of user portrait construction is shown in Figure 5.

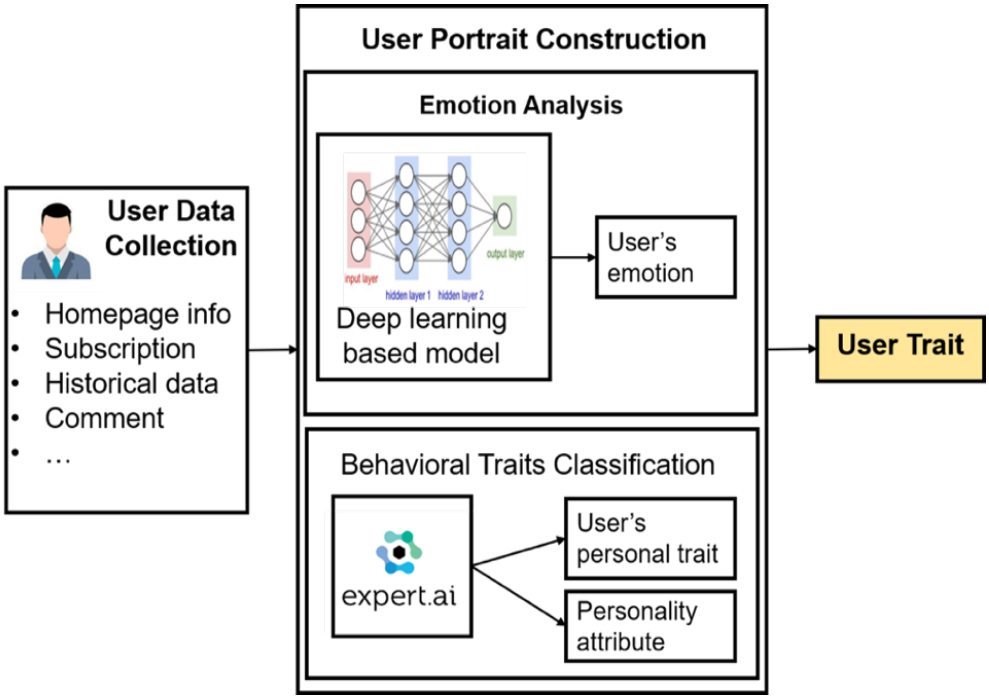

Figure 5. User portrait construction

For user portrait construction, we collect user data such as homepage information, subscription, historical data, comment, and so on. The collected data is then simultaneously used as input for emotion analysis and behavioral traits classification. For emotion analysis, we use a deep learning based method to identify the emotion attributes from the data. These attributes include three basic polarities of positive, negative and neutral, and eight basic emotions of trust, fear, surprise, sadness, disgust, anger, anticipation and joy based on Plutchik's wheel of emotions (Plutchik 2001).

For behavior traits classification, we use the API from expert.ai and their behavioral traits taxonomy, which was extended from Rothmann and Coetzer (2003), to classify the different dimensions and levels of user's personal trait and the personality attributes, as shown in Table 1. We notice that the personality attributes associating with *low* levels of all dimensions are generally pessimistic and negative, like rejection, asociality and violence. On the contrary, the personality attributes associating with *high* levels of all dimensions are often optimistic and positive, like self-confidence, inclusiveness, smartness and honesty. The personality attributes associating with *fair* levels of all dimensions are mostly neutral, like seriousness, cautiousness, rationality and calmness.

Dimension	Level	Personality Attribute
Openness	Low	Rejection; Apathy; Apprehension; Traditionalism; Conformism; Negativity; Bias
	Fair	Cautiousness
	High	Progressiveness; Acceptance; Courage; Positivity; Curiosity
Conscientiousness	Low	Superficiality; Unawareness; Disorganization; Insecurity; Ignorance; Illusion
	Fair	-
	High	Awareness; Spirituality; Concern; Knowledge; Self-confidence; Organization
Sociality	Low	Asociality; Impoliteness; Ungratefulness; Emotionality; Isolation; Disagreement
	Fair	Seriousness; Introversion; Unreservedness; Humor; Sexuality
	High	Extroversion; Pleasantness; Trustfulness; Gratefulness; Empathy
Action	Low	Sedentariness; Passivity
	Fair	Calmness
	High	Initiative; Dynamism
Ethics	Low	Violence; Extremism; Discrimination; Dishonesty; Neglect; Unlawfulness; Irresponsibility
	Fair	-
	High	Inclusiveness; Honesty; Compassion; Commitment; Lawfulness; Solidarity
Capability	Low	Lack of intelligence; Inexperience; Incompetence
	Fair	Rationality
	High	Smartness; Creativity; Competence
Moderation	Low	Dissoluteness; Gluttony; Materialism; Addiction
	Fair	Healthy lifestyle
	High	Self-restraint

Table 1. Dimension, level and personality attribute of behavioral traits from expert.ai

We use a scoring system to calculate the user trait based on polarity and emotion from emotion analysis and the personal trait level from behavior traits classification. The scoring criteria is shown in Table 2. We divide all user traits into 3 groups. A total score of 5-6 is defined as optimistic and positive; a total score of 3-4 is defined as neutral; and a total score of 2 or less is defined as potentially marginalized and pessimistic.

Emotion analysis			**Personal trait level**		
Positive	Neutral	Negative	High	Fair	Low
2	1	0	2	1	0
Trust, anticipation, joy	surprise	Fear, sadness, disgust, anger	**Overall score:** 5-6: Optimistic group 3-4: Neutral group <=2: Potential marginalized group		
2	1	0			

Table 2. Criteria of user trait calculation

Harmful Information Detection

The detailed view of the harmful information detection component is shown in Figure 6. We use the Bidirectional Encoder Representation from Transformers (BERT) based method (Devlin et al., 2018) to classify the harmfulness of online texts. BERT is a new language representation model proposed by Google's AI team in October 2018, which achieved the best results in 11 different natural language processing tests.

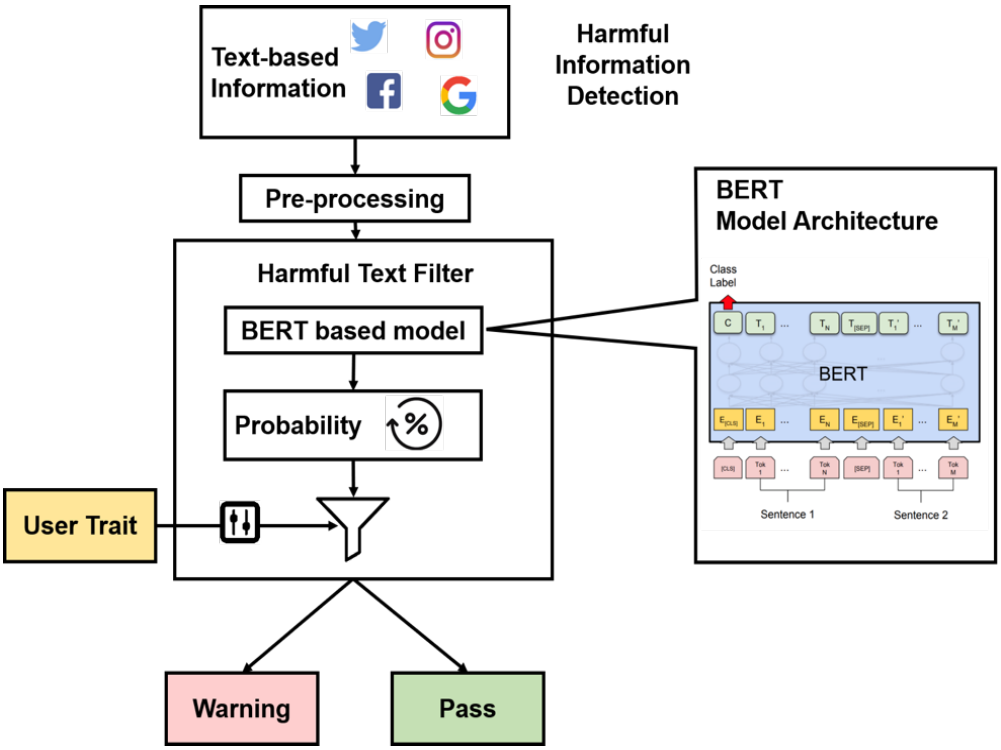

Figure 6. Harmful information detection

For the harmful text filter, we define the problem as a multi-class classification task. Based on the toxicity classification criteria from Google's Perspective API (Hosseini et al., 2017), text is categorized into toxic, severe toxic, obscene, threat, insult, identity hate and normal. After the BERT based model calculation, we get the probability that the input text belongs to each category. There are usually two ways to determine the final category. One is to define the category with the highest probability as the final category of the text according to sorting. The second is to set a threshold, for example, set the threshold to 0.5, and one (or more) categories with a probability greater than 0.5 are defined as the category to which the text belongs. Since the focus of this research is on the role of relative changes rather than absolute changes in threshold settings for filtering harmful information, we assume the thresholds for the three categories as the baseline threshold of 0.5, compared with the high threshold of 0.6 and the low threshold of 0.4.

In this paper, we innovatively use dynamic thresholds to implement personalized harmful information filtering. According to the user trait defined in the previous section, we set a higher threshold for the optimistic group. For example, if the threshold is raised to 0.6, only texts with a probability of belonging to a harmful category greater than 0.6 will be filtered or warned. This threshold setting can also ensure that this part of optimistic and exploratory groups can have a freer experience on the Internet. For the potential marginalized group, a lower threshold such as 0.4 is used, and filtering or warning will be given to those with a probability of harmful information greater than 0.4. Thereby making the judgment of harmful information stricter, and ultimately achieving better protection of vulnerable marginalized groups from cyberattacks and cyberbullying. The specific algorithm is shown below:

Algorithm 1 Harmful information detection

Input: t: text-based information; ut: user trait; $threshold\ list$ = [0.4, 0.5, 0.6]
Output: s: the solution for input information t
do text preprocessing
    precessed text pt = preprocessing (t)
do BERT based classification
    probability for each category pc = BERT (pt)
for $probability$ in pc **do**
    if ut = Optimistic group **then**
        if $probability$ > 0.6 **then**
            s = warning
        else
            s = pass
        end if
    end if
    if ut = Neutral group **then**
        if $probability$ > 0.5 **then**
            s = warning
        else
            s = pass
        end if
    end if
    if ut = Potential marginalized group **then**
        if $probability$ > 0.6 **then**
            s = warning
        else
            s = pass
        end if
    end if
end for
return s

Result and Analysis

To demonstrate and evaluate our personalized harmful information detection system, the classification result of the BERT based model is shown in Table 4. We use two evaluation indicators, Area Under the Curve (AUC) and Accuracy, to evaluate the results. AUC is an important indicator for judging classification or ranking ability, and the curve refers to the Receiver Operating Characteristic (ROC) curve. The x-axis of the ROC curve represents the false positive rate (FPR), and the y-axis represents the true positive rate (TPR). Comparing with traditional evaluation indicators such as precision, recall, and F1 value that depend on the classification threshold, the value of AUC is not affected by the threshold, and so we choose it as the evaluation indicator in this paper. Accuracy refers to the proportion of correctly predicted data among all data predicted in a category. The Accuracy is obviously higher than the AUC, which is a manifestation of uneven distribution of data.

Category	AUC	Accuracy
toxic	0.9354	93.18%
severe toxic	0.7224	98.93%
obscene	0.9386	98.13%
threat	0.7684	99.75%
insult	0.8882	97.39%
identity hate	0.8082	99.32%

Table 4. The harmful information filter classification result of BERT based model

We randomly selected 982 data samples from the Toxic Comment Classification Challenge dataset released by Jigsaw on the Kaggle platform. We simulate the optimistic group, the neutral group and the potential marginalized group, and then count the number of comments blocked or warned by the harmful text filter. The result is shown in Table 5.

Category	Neutral group (baseline)	Blocking-rate	Potential marginalized group	Blocking rate	Optimistic group	Blocking rate
toxic	180	-	203	+2.34% ↑	170	-1.02% ↓
severe toxic	2	-	8	+0.61% ↑	2	0
obscene	119	-	135	+1.63% ↑	102	-1.73% ↓
threat	0	-	0	-	0	-
insult	104	-	122	+1.83 ↑	77	2.75% ↓
identity hate	0	-	0	-	0	-
Overall	***405***	***41.24%***	***468***	***47.66%↑ (+6.42)***	***351***	***35.74%↓ (-5.5%)***

Table 5. The harmful information filter result for 3 groups

We set the neutral group with a threshold of 0.5 as the baseline in this paper. For the neutral group, 405 of the 982 comments were identified as harmful and they were blocked or warned by the filter, accounting for about 41.24% of the overall data. Since the data distribution of the dataset itself is not balanced, the harmful information identified as toxic, obscene and insult represent a large proportion. We also retain their original distribution for sampling.

For the potential marginalized group, due to the strict threshold of the classifier, the number of harmful comments blocked or warned by the filter increased by 63, and the overall blocking rate increased by 6.42%, which improved the protection of these vulnerable individuals. For the optimistic group, the number of harmful comments blocked or warned by the filter decreased by 55, and the overall blocking rate decreased by 5.5%. That means optimistic individuals have access to more information comparing to the other two groups.

Conclusion and Future Work

Most existing research on harmful information detection from text only make breakthroughs in algorithmic efficiency and accuracy, while regarding Internet users as an average whole. But we know marginalized groups are more likely to be victims of cyberattacks with harmful information, and these cyberattacks often cause more harm to them than others. In this paper, we propose a framework for a personalized harmful

information detection system based on user portraits. This detection system uses different thresholds to filter harmful information according to a user's personality characteristics, so that we could maximize both the Internet experience and protection of the vulnerable minorities from harmful text.

This design science research is at a stage with a proof-of-concept that could demonstrate its feasibility. For further iterations on implementation and algorithmic development for building user portraits, issues like privacy and ethics will become major considerations and potentially challenges. To what extent should we collect user data to balance between privacy and accuracy? How could we avoid unintended consequences caused by excessive protection? More future studies are required to answer these questions.

REFERENCES

Demchenko, Y., De Laat, C., & Membrey, P. (2014, May). Defining architecture components of the Big Data Ecosystem. In 2014 International conference on collaboration technologies and systems (CTS) (pp. 104-112). IEEE.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138.

Hinduja, S. (2021). Cyberbullying in 2021 by age, gender, sexual orientation, and Race. Cyberbullying Research Center. Retrieved April 20, 2022, from https://cyberbullying.org/cyberbullying-statistics-age-gender-sexual-orientation-race

Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2020). Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, 1-14.

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2016). Online hate and harmful content: Cross-national perspectives (p. 154). Taylor & Francis.

Llorent, V. J., Ortega-Ruiz, R., & Zych, I. (2016). Bullying and cyberbullying in minorities: Are they more vulnerable than the majority group?. Frontiers in psychology, 7, 1507.

Nurrahmi, H., & Nurjanah, D. (2018, March). Indonesian twitter cyberbullying detection using text classification and user credibility. In 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 543-548). IEEE.

Pawar, R., Agrawal, Y., Joshi, A., Gorrepati, R., & Raje, R. R. (2018, May). Cyberbullying detection system with multiple server configurations. In 2018 IEEE International Conference on Electro/Information Technology (EIT) (pp. 0090-0095). IEEE.

Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, 605-611.

Plutchik, Robert. 2001. "The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice." American Scientist 89 (4). JSTOR: 344–50.

Ptaszynski, M., Lempa, P., Masui, F., Kimura, Y., Rzepka, R., Araki, K., ... & Leliwa, G. (2019). Brute-force sentence pattern extortion from harmful messages for cyberbullying detection. Journal of the Association for Information Systems, 20(8), 1075-1127.

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. SA Journal of industrial psychology, 29(1), 68-74.

Sharma, S., Agrawal, S., & Shrivastava, M. (2018). Degree based classification of harmful speech using twitter data. arXiv preprint arXiv:1806.04197.

Smith, P. K., Mahdavi, J., Carvalho, M., & Tippett, N. (2006). An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. Research Brief No. RBX03-06. London: DfES.

Sugandhi R, Pande A, Chawla S, et al. Methods for detection of cyberbullying: A survey[C]//2015 15th International Conference on Intelligent Systems Design and Applications (ISDA). IEEE, 2015: 173-177.

Sui J. Understanding and fighting bullying with machine learning[D]. USA: The Univ. of Wisconsin-Madison,, 2015.

Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter[C]//Proceedings of the NAACL student research workshop. 2016: 88- 93.