Whether Real Names on the Internet are Effective in Reducing Offensive Language: The Impact of Anonymity and Offensive Language on Mental Health

Lingjie Jiang[1,a,*]

[1]School of Teacher Education, Ningbo University, Ningbo, 315211, China
a. jlj186002476@163.com
**corresponding author*

Abstract: The Online real-name systems are becoming an increasingly popular policy, with the main reason for support being the perceived effectiveness of online real-name systems in reducing offensive language. Upon further exploration, this paper finds certain flaws in this view. The frequency of offensive language between real-name and anonymous accounts in previous studies may have arisen due to problems with the research methodology. Meanwhile, the impact of offensive language on an individual's mental health is not entirely negative. Offensive language should be further categorized in future studies to explore its effects on individual mental health.

Keywords: Anonymity, Offensive language, real name systems

1. Introduction

Digital communication and virtual platforms based on communication technologies are increasingly becoming an integral part of modern life. This virtual communication, free from time and space, allows global information to be more easily exchanged and disseminated. However, the increasing frequency of cyber violence, online sexual harassment and hate speech in recent years has posed a great danger to the mental health of Internet users, and can even lead to more offline violence and illegal incidents [1-3]. Offensive language means including, but are not limited to, vulgar, pornographic, taboos and hateful language. Vulgar usually language refers to rude expressions, which include offensive reference sex or bodily function or targeted family members. Pornographic language refers to descriptions that are explicitly related to sexual themes for the purpose of satisfying one's own sexual arousal or pleasure without the permission of another person. Hateful language includes any communication that demeans individuals or groups based on characteristics such as race, country, color, gender, sexual orientation, religion, and disease [4]. Because of the virtual nature of Internet communication, where users communicate with each other on online platforms without physical contact, structured offensive language - including emojis and images that convey the appropriate meaning [5] - has become an important vehicle for online forms of violence. Various countries, governments, and research institutions have paid great attention to it, trying to secure the information of users on the Internet by studying and banning offensive language [6] [7].

The idea that anonymity will lead to more offensive language originates from Zimbardo's deindividuation theory. Zimbardo's theory defines anonymity as the inability of others to identify an individual, making it harder for individuals to be evaluated, judged, or punished, and this allows individuals to have a reduced level of internal control, making them more likely to engage in normally inhibited behaviors, including antisocial or illegal behavior [8]. This is the main argument in favor of real names in today's cybersecurity debate, but does anonymity really make individuals use more offensive language in their use of the Internet? Previous research has tended to focus more on the criminal behavior that has occurred in anonymous platforms and its psychological processes, such as the fact that cyber violence carried out in anonymous platforms can give individuals a sense of feeling that they can get away with [9], but there is a lack of research on the relationship between anonymity and the frequency of actual aggressive language. Among the available studies, some point out that there is no significant correlation between anonymity and the frequency of aggressive language [10]. Some studies support that anonymity will lead to more aggressive language, but in the category of aggressive language it is clear that anonymity will only increase target-less aggressive language [11].

In contrast, the above studies often suffer from the research shortcomings of different sources of data bases, different levels of anonymity of survey platforms, and the use of negative keywords as a measurement feature of aggressive language in the language detection methods used in previous studies. However, previous studies have shown that utterances with negative words do not necessarily have negative connotations such as insult or harassment, but may also occur because of the degree of expression of emotion or reinforcement. Therefore, the actual criteria for aggressive language and the relationship that exists between the online anonymity regime and the frequency of aggressive language should perhaps be remeasured in the course of future research, so as to better focus on and protect users' online safety as well as psychological health.

2. Offensive language

Current research on offensive speech often cites or expands on Jay 1992's definition of it, which Jay says offensive language often includes vulgar, pornographic, taboos and hateful language [12]. And in a 2006 study, Jay noted the nature of aggressive speech in making victims feel negative psychological feelings such as anxiety and punishment [13]. In subsequent extensions of the definition, offensive speech is defined as discourteous speech [14]. Because of these characteristics, offensive speech is often closely associated with negative behaviors such as cyberbullying, sexual harassment and hate speech when communicating online. Nonetheless, the composition of offensive speech is more complex than simply negative communication texts. Based on the social norms theory, offensive speech is often included in criticism of social subjects who violate social norms or when engaging in conflicting social situations. However, this part of offensive speech is more often generated by altruistic motives [15].

Because of the relevance of offensive speech to people's online communication behavior - both negative and positive - governments and platforms are attempting to protect individuals' good Internet communication by detecting and banning offensive speech [16] [17]. In practice, however, the operational definition of offensive speech is more complex. Early detection of offensive speech was often based on negative terms, such as insulting words or words with obvious links to sexual organs [11]. In practice, however, humorous remarks, jokes with friends, or expressions of strong emotions can contain such negative words, which were also identified as offensive in earlier tests. Currently, text mining programs based on natural grammar processing (NLP) are used to detect offensive speech, and advanced neural networks and artificial intelligence are used to learn and analyze text content, but due to grammatical errors, ambiguous expressions, and complex text content such as emoji, the interpretation of offensive language [17].

2.1. Offensive Language with Mental Health

The current mainstream view that offensive speech often has a negative impact on mental health is due to the cyberbullying, sexual harassment, and other behaviors included in offensive speech. There has been considerable empirical research showing that cyberbullying is closely related to adolescent mental health, with higher levels of negative emotions such as depression and anxiety in adolescents who are subjected to cyberbullying [7]. A study of subjects at a California health clinic showed that women were more often sexually harassed online than men, and that the experience of sexual harassment was positively associated with self-reported levels of depression and anxiety over a 30-day period [18]. Hate speech is already considered illegal, and studies have demonstrated the psychological harm caused by hate speech during online communication [19]. However, fewer studies have directly examined the relationship between aggressive speech and mental health levels. Bucur et al. noted that people with self-reported depression diagnoses used more aggressive language, but did not confirm a causal relationship between the two [20]. Also, in terms of specific content, subjects who reported a depression diagnosis used more non-targeted aggressive language and focused more on negative self-exposure and expression. Other related studies have focused on the relationship between negative behaviors and individual depression levels, such as the effect of negative comparisons on depression levels, but did not include aggressive speech as a key factor.

In contrast, Jay, in a follow-up study, pointed out the methodological limitations of the existing studies. During previous studies, self-reported questionnaires were often used for aggressive speech and consequent mental health levels, which could not give a precise definition of aggressive speech, while self-reported mental health levels were prone to higher bias. Meanwhile, offensive language detection programs created with neural network learning are often used in practice nowadays to achieve more accurate screening and analysis, but offensive language detection programs that use vocabulary as the main screening mechanism for text are still widely used in the process of previous studies [21].

Another group of scholars is also skeptical about the negative impact of aggressive language on mental health. In Ferguson's article exploring the impact of social media, he mentions that despite the increase in aggressive speech on social media, overall violence in the world has been declining despite two world wars [22]. In social norms theory, aggressive speech against immoral behavior is often altruistically motivated, so does this intrinsic incentive for aggressive speech also lead to negative effects on mental health levels? No studies have yet explored this aspect.

3. Anonymity

With the popular application of the Internet in daily life, the influence of online speech on human life is increasing, and the regulation of online speech has become a hot topic in recent years. Online anonymity is generally regarded as one of the main causes of non-regulated speech as well as online violence [23]. Therefore, the debate calling for a real-name system on the Internet to enhance regulation and protect the right to online anonymity has also gained widespread attention. A growing number of platforms require individuals to authenticate with their real names, such as Facebook, which requires users to be identified by their real names, and gaming platforms such as Blizzard and Ubisoft, which require tying to users' identities.

Traditionally, anonymity is defined as the absence of information and conditions that can identify an individual in a social environment, and the inability to accurately correspond an identity to a specific individual [24]. Hayne and Rice proposed in 1997-pointed anonymity can be distinguished as technical anonymity and social anonymity, with the former referring to the absence of information, traces, and other elements that can be traced to an individual's identity. The latter, on the other hand, is when an individual believes that he or she cannot be identified in a given environment [25].

However, this definition of anonymity still treats anonymity and non-anonymity as dichotomous states. In recent years, with the changing privacy norms of platforms for individuals, the transition from anonymity to non-anonymity is seen as a continuum [26]. For example, in some platforms individuals are required to upload a real photo, but the platform account is not tied to the individual's real name and the user can operate under a pseudonym.

The use of cross-platform accounts further complicates the definition of anonymity, for example, the Huffington Post requires users to be tied to their Facebook accounts and will display their comments on Facebook [27]. In Alice and Danah's paper, this is called "context collapse", where an individual no longer has a multifaceted, separate anonymous identity across online platforms, even if the user is anonymous on one platform, but this anonymity does not give the user a separate, confidential anonymous identity due to the real-name nature of the cross-platform account to which they are tied [28]. Many studies currently assess anonymity through the level of three metrics: traceability, durability and connectedness [29] [30]. Differences in these three levels will affect the level of anonymity of online identities and also have an impact on the behavior of individuals. For example, connectedness indicates how likely it is that an individual will post information that is known in the real world or by other important social relationships, and the higher the connectedness of an online identity, the less likely it is that an individual will post anti-regulatory statements [27].

Current research has focused more on dichotomous anonymity states and lacks research on continuous anonymity states.

3.1. Anonymity and Offensive Language

Zimbardo's theory of "deindividuation theory" was one of the first theories of how anonymity is intrinsically linked to offensive speech. He pointed out that anonymous speech leads to deindividuation and that this phenomenon leads to a weakening of the individual's internal self-control, making it easier for the individual to engage in behaviors that are normally unacceptable and in violation of social norms. However, this theory has recently been shown to lack actual evidence that individuals are in a state of deindividuation when they are anonymous [8].

The SIDE (Social identity model of deindividuation) theory is a further extension of Zimbardo's theory. The theory suggests that the regulation of individual behavior by anonymity depends on two factors, namely, anonymity and the degree to which the individual identifies with the group. When individuals have a high level of group identification and a low level of self-identification, they are more likely to conform to social norms. And when an individual's identification with the self is high, anonymity is more likely to weaken the influence of social norms. Anonymity when no group can identify individuals creates socialized goal orientation, i.e., everyone is more willing to work for group goals. And when an individual can be identified, individuals are more willing to work for their own goals [31].

At the same time, the strategy theory in SIDE theory states that individuals will use the influence of anonymity to achieve their own goals. For example, people in minority sexual orientation communities are more willing to express their views in anonymity to counteract stronger opposing forces, which may be unpopular in real name situations. However, such views need to conform to norms within the minority community, otherwise the publication of such views also lacks the strength of support in anonymous situations. This de-suppression is not entirely positive, however, and some unjust minorities are also more likely to express their views in anonymous situations [31] [32]. Sia et al. in 2002 made a experiment which supported group polarization is more likely to occur in anonymous situations was also demonstrated in an experiment [33]. Undoubtedly, when this polarization occurs among negative groups, it will bring about even more offensive views and rhetoric.

While most of the findings based on actual studies support that anonymity will increase aggressive speech, such results are subtle when more details are considered. Daegon and Alessandro's study noted that individuals using accounts with non-real names were more inclined to use aggressive language [11]. However, this study, along with other studies conducted when real-name policies were implemented in South Korea, showed that while real-name systems reduced the frequency of aggressive speech by individuals, they also reduced the enthusiasm of individuals to interact [34]. Interactions become more cautious under real names, so this may also be one of the factors that led to the data showing a decrease in offensive speech. Not coincidentally, when Moore et al was studying the comments of Huffington Post in 2019, it showed individuals used fewer offensive words when user accounts were linked to Facebook. However, overall comment quality tended to decrease when they had a lower level of anonymity due to higher connectedness [27].

Some studies also exhibit the opposite view that anonymity does not directly increase offensive speech. In a study of Russian-language forums, anonymous users were not shown to significantly increase the use of offensive words [10]. In studies based on social norms theory, when non-anonymity serves to increase authenticity and reliability as a means of discouraging and denigrating non-normative behavior, non-anonymous users are more likely to use offensive speech because it is altruistically motivated [15].

4. Conclusion

Since the rise of online platforms, the debate between online anonymity and online real-name system has never stopped. The mainstream view that "anonymity will lead to more offensive speech, thus endangering the mental health and even physical health of individuals and society" has been one of the important points supporting the real-name system [24]. However, according to the above, this view seems to lack strong evidence to justify it. In some of the studies anonymity did lead to more offensive comments, but it also led to a decrease in the overall number and frequency of comments [11] [27] [34]. Could the fact that individuals are more cautious about posting comments under real names lead to another type of social pressure on users to use social networks under their real names? Other studies have also indicated that anonymity or not did not affect the number of offensive comments made by individuals [10]. Secondly, does aggressive speech cause more mental problems? Although negative behaviors consisting of aggressive speech have often been shown to lead to more mental problems such as depression and anxiety [7] [18] [19]. However, in the current research, more studies still use word feature recognition programs or neural learning-based aggressive speech recognition programs, which are still deficient in recognizing aggressive speech in ambiguous contexts, and some humorized speech or untargeted aggressive language are generalized [7] [11] [14] , and do these speech still cause negative psychiatric problems in individuals?

As for offensive language that has a target, there are distinctions among them. Social norms theory suggests that some offensive speech is altruistically motivated to criticize anti-normative behavior, often targeting social concepts such as business and government, but does this offensive speech also lead to negative experiences for other individuals who view the content [15] ? There is no further research on the psychological impact of the different components of offensive speech. Jay and Ferguson's articles suggest that online aggressive behavior may be a proxy for offline aggressive behavior [21] [22]. It ventilates the negative emotions of the initiator and may even reduce the aggressive behavior in reality. And viewed from this perspective, aggressive language may even reduce the mental burden on society in general.

Thus, does a more real-name system lead to a more civilized and harmonious online communication atmosphere and online community? The answer does not seem to be the case, and the answer to this question will be even more complicated when some of the benefits of anonymity are taken into account, such as an individual's right to privacy, or the idea of wanting to guarantee one's independence between different platform identities. Fortunately when anonymity is considered as a continuum rather than a dichotomy [28], there is more room for trade-offs - for example, regulating persistence at the anonymity level, where long-term pseudonyms lead to less offensive speech than short-term pseudonyms [27] - but unfortunately there is still less research on the different properties of such pairs of anonymity.

In the future, we can consider how different traceability, durability and connectedness will affect individual speech. Second, in future studies, we can use language recognition programs that incorporate artificial intelligence to make the identification of offensive speech more contextual, and humorous language that is not hostile but contains vulgar words can be excluded from further identification, thus enhancing the reliability of the study. At the same time, on this basis, we may be able to make a more careful classification and judgment of offensive speech as a way to distinguish the relationship between offensive speech that is untargeted or made for altruistic purposes and the mental health of individuals. This may help us develop a more realistic policy in the anonymity vs. real name debate.

References

[1] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In Proceedings of the 24th international conference on world wide web (pp. 29-30). DOI: https://doi.org/10.1145/2740908.2742760

[2] Reed, E., Salazar, M., Behar, A. I., Agah, N., Silverman, J. G., Minnis, A. M., ... & Raj, A. (2019). Cyber sexual harassment: Prevalence and association with substance use, poor mental health, and STI history among sexually active adolescent girls. Journal of adolescence, 75, 53-62. DOI: https://doi.org/10.1016/j.adolescence.2019.07.005

[3] Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. Social Identities, 22(3), 324-341. DOI: https://doi.org/10.1080/13504630.2015.1128810

[4] Jay, T. & Janschewitz, K. (2008). The pragmatics of swearing., 4(2), 267-288. DOI: https://doi.org/10.1515/JPLR.2008.013

[5] Althobaiti, M. J. (2022). BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 13(5). DOI: 10.14569/IJACSA.2022.01305109

[6] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, 13825-13835. DOI: 10.1109/ACCESS.2018.2806394

[7] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012, September). Detecting offensive language in social media to protect adolescent online safety. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (pp. 71-80). IEEE.DOI: 10.1109/SocialCom-PASSAT.2012.55

[8] Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order vs. deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.). Nebraska symposium on motivation (Vol. 17, pp. 237–307). Lincoln: University of Nebraska Press.

[9] Wright, M. F. (2013). The relationship between young adults' beliefs about anonymity and subsequent cyber aggression. Cyberpsychology, Behavior, and Social Networking, 16(12), 858-862.DOI: http://doi.org/10.1089/cyber.2013.0009

[10] Potapova, R., & Gordeev, D. (2015). Determination of the internet anonymity influence on the level of aggression and usage of obscene lexis. arXiv preprint arXiv:1510.00240. DOI: https://doi.org/10.48550/arXiv.1510.00240

[11] Cho, D., & Acquisti, A. (2013, June). The more social cues, the less trolling? An empirical study of online commenting behavior. In Proc. WEIS. DOI: https://doi.org/10.1184/R1/6472058.v1

[12] Jay, T. B. (1992). Cursing in America. Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/z.57

[13] Jay, T. B., King, K., & Duncan, D. (2006). Memories of punishment for cursing. Sex Roles, 32, 123–133.DOI: https://doi.org/10.1007/s11199-006-9064-5

[14] Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. DOI: https://doi.org/10.48550/arXiv.1705.09899

[15] Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. PLoS one, 11(6), e0155923. DOI: http://doi.org/10.3886/E72764V1

[16] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, 13825-13835.DOI: 10.1109/ACCESS.2018.2806394

[17] Pal, S., Chakraborty, A., Chakraborty, R., & Mitra, I. Detection Of Hate Speech And Offensive Language Using Machine Learning And Neural Networks With AI Explanation. https://link.springer.com/chapter/10.1007/978-981-15-2740-1_17

[18] Reed, E., Salazar, M., Behar, A. I., Agah, N., Silverman, J. G., Minnis, A. M., ... & Raj, A. (2019). Cyber sexual harassment: Prevalence and association with substance use, poor mental health, and STI history among sexually active adolescent girls. Journal of adolescence, 75, 53-62.DOI: 10.1016/j.adolescence.2019.07.005

[19] O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. Pediatrics, 127(4), 800-804.DOI:https://doi.org/10.1542/peds.2011-0054

[20] Bucur, A. M., Zampieri, M., & Dinu, L. P. (2021). An exploratory analysis of the relation between offensive language and mental health. arXiv preprint arXiv:2105.14888.DOI:https://doi.org/10.48550/arXiv.2105.14888

[21] Jay, T. (2009). Do offensive words harm people?. Psychology, public policy, and law, 15(2), 81.DOI: https://doi.org/10.1037/a0015646

[22] Ferguson, C. J. (2021). Does the Internet Make the World Worse? Depression, Aggression and Polarization in the Social Media Age. Bulletin of Science, Technology & Society, 41(4), 116-135.DOI:https://doi.org/10.1177/02704676211064567

[23] Lea M, Spears R and de Groot D (2001) Knowing Me, Knowing You: Anonymity Effects on Social Identity Processes within Groups. Personality and Social Psychology Bulletin 27 (5): 526–537 DOI: https://doi.org/10.1177/0146167201275002

[24] Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:“On the Internet, Nobody Knows You’re a Dog”. Computers in Human Behavior, 23(6), 3038-3056. DOI: https://doi.org/10.1016/j.chb.2006.09.001

[25] Hayne, S. C., & Rice, R. E. (1997). Attribution accuracy when using anonymity in group support systems. International Journal of Human–Computer Studies, 47, 429–452. DOI: https://doi.org/10.1006/ijhc.1997.0134

[26] Beyer, J. L. (2012). What does anonymity mean? Reddit, activism, and ‘creepshots’. Jessica L. Beyer, 6.

[27] Moore, A., Fredheim, R., Wyss, D., & Beste, S. (2021). Deliberation and identity rules: The effect of anonymity, pseudonyms and real-name requirements on the cognitive complexity of online news comments. Political Studies, 69(1), 45-65. DOI:https://doi.org/10.1177/0032321719891385

[28] Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New media & society, 13(1), 114-133. DOI: https://doi.org/10.1177/1461444810365313

[29] Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The “nasty effect:” Online incivility and risk perceptions of emerging technologies. Journal of computer-mediated communication, 19(3), 373-387. DOI: https://doi.org/10.1111/jcc4.12009

[30] Fredheim, R., Moore, A., & Naughton, J. (2015, June). Anonymity and online commenting: The broken windows effect and the end of drive-by commenting. In Proceedings of the ACM web science conference (pp. 1-8). DOI: https://doi.org/10.1145/2786451.2786459

[31] Postmes, T., & Spears, R. (2002). Behavior online: does anonymous computer communication reduce gender inequality? Personality and Social Psychology Bulletin, 28(8), 1073–1083. DOI: https://doi.org/10.1177/01461672022811006

[32] Spears, R., & Lea, M. (1992). Social influence and the influence of the ‘‘social’’ in computer-mediated communication. In M. Lea (Ed.), Contexts of computer-mediated communication (pp. 30–65). London: Harvester-Wheatsheaf.

[33] Sia, C., Tan, B. C. Y., & Wei, K. (2002). Group polarization and computer-mediated communications: effects of communication cues, social presence, and anonymity. Information Systems Research, 13(1), 70–90.DOI:https://doi.org/10.1287/isre.13.1.70.92

[34] Cho, D., Kim, S., & Acquisti, A. (2012, January). Empirical analysis of online anonymity and user behaviors: the impact of real name policy. In 2012 45th Hawaii international conference on system sciences (pp. 3041-3050). IEEE.DOI: https://doi.org/10.1109/HICSS.2012.241