NSE-analysis of Facebook's suicide algorithm

Word-count: 2186

Bono Lardinois 11713364,
Jan Janiszewski 10004378,
Maxim Vitalis 11303719.

October 2022

1 Introduction

Yearly, 800.000 people commit suicide. This equates to one suicide in almost every 40 seconds; hence, it is one of the leading causes of death among young people [Ritchie et al., 2015]. Suicide is a global problem that needs to be addressed.

The influence of media on suicide dates back to the first world war [Stack, 1988]. However, social media influenced suicide in recent years and has become an important topic. The pressure people experience from social media can lead to self-harm, and even suicide [Robinson et al., 2016]. At the same time, analyzing social media posts can prevent suicides. For ten years, Facebook has had suicide prevention algorithms in place. By flagging posts that indicate a suicidal ideation, users indicated to Facebook whenever a user was a high self-harm risk. A Facebook team would receive a warning for these posts, judge the situation, and decide whether to act upon the post or not. However, relying on users to flag suicidal content is not a sufficient approach to suicide prevention.

Because prevention is better than cure, Facebook started working on an algorithm that would predict if someone would potentially harm themselves or not based on recent posts in 2017. By 2018 it rolled out the algorithm in different countries across the world (due to the General Data Protection Regulation, EU countries were excluded). Shortly afterwards, it received a major backlash from different media sources, criticizing it for the algorithm [Goggin, 2019] and its use by Facebook.

This case study examines whether the negative media feedback on the algorithm was reasonable and whether Facebook's suicide surveillance algorithm satisfies ethical requirements through the Networked Systems Ethics guidelines. The report will include an analysis of risks and harms, data governance, and informed consent aspects of the guidelines. The results will be discussed first and finished with a conclusion.

2 Analysis

This section examines Facebook's algorithm within the framework of the Network System Ethics [Zevenbergen, 2016], focusing on the algorithm's risks of harm, Facebook's data governance, and the informed consent that Facebook user need to give.

2.1 Risk of harm

Even though the algorithm was designed with the best intention to support customers who show the (imminent) risk of suicide, it still can cause more harm than it helps to prevent.

First, the potential harm of wrong classification should be analyzed. The current Facebook classifier can take three different decisions, namely that a person is at imminent risk of suicide (imminent risk), that a Facebook user shows potential suicidal intent (potential risk), and that there is no suicide potential based on a user's written post (no risk). Furthermore, any decision can be correct (true) or incorrect (false), e.g., the classifier can indicate that a Facebook user is at imminent risk, while the user is actually joking. This results in a confusion matrix with four options; false positive, true positive, false negative, and true negative [Ting, 2010].

A positive decision, whether false or true, can have tremendous consequences. For instance, in case of imminent risk, Facebook immediately calls local first responders (e.g., police), who - in this case - have the right to enter a house without a warrant. Moreover, as authorities are obliged to prevent suicidal persons from self-harming themselves, they have to arrest the individual and perform a mental health check in the hospital [Singer, 2018]. This can cause tremendous mental or physical damage and stereotyping or stigmatization from their surroundings, causing individuals with suicidal thoughts to hesitate sharing their thoughts on Facebook and getting any help [Celedonia et al., 2021]. Moreover, the consequences of such interventions (e.g., feeling misunderstood by their surroundings, having self-doubt, potentially having a lower chance of employment) can cause the individuals to isolate themselves and become even more suicidal [Corrigan et al., 2014]. Additionally, it can even have further consequences, as in some countries like Singapore, an attempt to suicide will lead to imprisonment of at least one year [Haridy, 2019].

A false negative also carries bad consequences with it as individuals who know of the algorithm could rely on the algorithm too heavily, creating a bystander effect. For example, Facebook users who are worried about the suicidal thought of their friend would leave it to Facebook to figure out if the friend is suicidal as they would feel that they don't have the authority to intervene [Orlando, 2020]. This would result in that no one notices the call for help, which is the worst outcome for the algorithm, as it would mean that the wrong prediction of the algorithm lead to a suicide.

The above assumptions are in no way of only pure theoretical nature. Similar experiences have already been reported by developers of other suicide prevention

algorithms. For example, the Samaritans have launched a Twitter plug-in that sends a message to followers of a Twitter account if the Twitter account sends out suicidal messages. However, due to a strong backlash from mental health community members, it was discontinued. This backlash was mainly based on the worry that suicidal persons could censor their tweets to avoid appearing suicidal to the outside world, making it even harder for them to receive the help they need. Additionally, the concern was raised that bullies could use the Samaritan suicide prevention plug-in to bully suicidal persons and enhance the likelihood of suicide. This shows that even with the best intentions (Samaritans are an NGO to help people), suicide prediction algorithms can cause harm [Dredge, 2014].

2.2 Data Governance

As health data, the information gathered by the algorithm should be treated with the highest care. One could even argue that only medical institutions should handle data like this [Celedonia et al., 2021]. Facebook is far from taking good care of its data and was already involved in many data scandals. For example, Facebook has already reported on multiple data breaches which involved personally identifiable data [Dodds, 2020; Holmes, 2021]. To make matters even worse, Facebook sometimes even intentionally shares its user data. For example, it provided unfettered and unauthorized access to personally identifiable information (PII) of its users to Cambridge Analytica, a data analytical company, in 2015 [Isaak and Hanna, 2018]. This signals that data governance at Facebook falls short of the high standards required by medical institutions to handle medical data.

Facebook itself shows its lack of responsibility for governing its users' data by stating that in imminent danger, local resources (e.g., first responders) are contacted to prevent potential suicidal persons [Gomes de Andrade et al., 2018]. Although this may seem ethically correct (any suicide attempts should be prevented, especially if shared on social media previously), it poses a heavy burden on the privacy of Facebook users, as this means that users who write the wrong things on Facebook can expect many consequences, such as tracking of their cell phones by local authorities, unannounced visits by the local authorities, or even forced entry to their houses; all of this under the hood of suicide prevention.

On the other hand, it is worth mentioning that Facebook already had suicide prevention methods in place for more than ten years when the suicide prevention algorithm came out [Gomes de Andrade et al., 2018]. Up to now, no signs indicate that any of the data gathered from those methods surfaced in the public media. This strong argument shows that Facebook treats this data more seriously than usual user data. Furthermore, although many external instances requested access to the methods used and underlying data to the suicide prevention algorithm, they have still not been disclosed to any of those instances. This secrecy shows that Facebook does its best to protect its users' suicide data from surfacing in public.

Additionally, contrary to other suicide prevention algorithms mentioned,

Facebook's algorithm only discloses information about the suicide attempt to Facebook employees who read the post and then decide whether to call local authorities. This guarantees that only persons are informed of a suicide attempt who have no intention to harm the suicidal person and indeed only want to help.

2.3 Informed consent

How ethical the data analysis for a specific purpose is, greatly depends on whether the data owners know how their data will be used. This also applies to Facebook and how it uses the data of its users. When a user signs up for Facebook, they must share personal information and permit their information to be used in many different ways. It is not required for anyone to give these permissions, but as a consequence of not giving them, you cannot use the platform. Although it is technically true that all users have given consent, one could argue that this is not informed consent. It is not always necessary to seek informed consent, and in that case, this requirement can be waived. However, this may only happen under specific circumstances, and in all other cases, explicit informed consent is required. In this section, we analyze whether Facebook gained informed consent or is able to waive it according to the Network System Ethics guidelines [Zevenbergen, 2016].

Per Facebook's data selection and data usage sections in the privacy policy, it is clear that Facebook will collect all communications and other content users share on the platform. It is also mentioned that this data may be used to monitor the behavior and safety of individuals. Only after following the "detect when someone needs help" hyperlink a blog post generally describing the algorithm is shown. This is much further removed from the user than topics such as advertisements and location services.

Most social media users are quite used to the different kinds of permissions they need to give to join a platform. Suicide prevention and health checks, in general, are a novel introduction and could explain why Facebook users, in general, are not aware that their data is being used in this way [Burr et al., 2020]. This undermines the guidelines as presented in the NSE guidelines for "meaningful consent". This lack of knowledge about the feature's existence also implies that users can not fully understand "the scope and risk of harm" discussed earlier. Moreover, even informed users cannot know exactly how this algorithm can be used, as the description Facebook provides is only a high-level overview. With this information, one can also question how Facebook intends to develop or expand the future algorithm, which is not a topic Facebook choose to elaborate on.

Ultimately, Facebook is conducting medical research. In his paper on informed consent involving human participants in medical trials, Nijhawan lists the four cumulative requirements when waiving informed consent [Nijhawan et al., 2013]. Given the nature of the research that is taking place, it is also interesting to compare these requirements with those in the NSE guidelines. Considering both, it becomes apparent that few, if any, of the requirements are met. To illustrate, Facebook does not demonstrate how there is minimal risk to

the end user. More specifically, Facebook does not mention any risk at all to end users. Nor is the company transparent about how possible risks were eliminated through an iterative and reflective process. Taking examples from Nijhawans paper, subjects are not informed after participation, nor is it apparent why not having the waiver would impede Facebook's research.

To conclude this part of the analysis, Facebook does not have the informed consent of its users for the data analysis. Medical research is presented to the user as common practice and is hidden among items that generally are.

3 Discussion

As discussed in the first part of the analysis, there are potentially severe negative consequences when the machine learning algorithm or the Facebook employee make an error in their classification, consequences which can even cause suicide (e.g., by worsening the situation of the potentially suicidal person). This is even more of a problem considering that it is difficult to fine-tune the algorithm as there is no risk-free classification in unclear or edge cases; tremendous consequences are linked to any false prediction of the algorithm, whether it is that the individual will commit suicide or not.

Furthermore, the fact that Facebook commonly shares user data with third parties [Isaak and Hanna, 2018] makes it much harder to precisely track the risks that are involved in Facebook's data analysis. As the data that Facebook analyzes is essentially medical data, it must be handled with high levels of care and be adequately governed. There is no information about how long the data is stored, where it is stored, and by whom it is stored. There is also no information on how these companies process or analyze the users' data precisely or if there are any plans for further analysis.

Therefore, we believe that it is not possible to confidently say that there is little to no risk involved for the users. This means that a clear and explicit informed consent is necessary in order to use data of Facebook users, which is clearly not the case [Gomes de Andrade et al., 2018]. We, therefore, believe that Facebook does not have informed consent from its users to participate in the suicide prevention algorithm research and to use the data in its current form.

4 Conclusion

Based on the guidelines and our work, we conclude that the suicide algorithm could benefit society. However, we also conclude that Facebook is not the upright instance to create and govern such an algorithm. Gomes de Andrade et al. [2018] stated multiple non-profit-oriented actors who built suicide prevention methods, apps, and algorithms. Instead of building its own proprietary algorithm, Facebook should delegate this task to organizations that can better handle such delicate issues as suicide prevention [Gomes de Andrade et al., 2018]. We hope Facebook will start doing this as soon as possible, as their

current practices are not at all in line with the ethical guidelines discussed.

References

Burr, C., Morley, J., Taddeo, M., and Floridi, L. (2020). Digital psychiatry: Risks and opportunities for public health and wellbeing. *IEEE Transactions on Technology and Society*, 1(1):21–33.

Celedonia, K. L., Corrales Compagnucci, M., Minssen, T., and Lowery Wilson, M. (2021). Legal, ethical, and wider implications of suicide risk detection systems in social media platforms. *Journal of Law and the Biosciences*, 8(1):lsab021.

Corrigan, P. W., Druss, B. G., and Perlick, D. A. (2014). The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest*, 15(2):37–70.

Dodds, I. (2020). Facebook was repeatedly warned of security flaw that led to biggest data breach in its history.

Dredge, S. (2014). Samaritans radar analyses twitter to identify users at risk for suicide.

Goggin, B. (2019). Inside facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts.

Gomes de Andrade, N. N., Pawson, D., Muriello, D., Donahue, L., and Guadagno, J. (2018). Ethics and artificial intelligence: suicide prevention on facebook. *Philosophy & Technology*, 31(4):669–684.

Haridy, R. (2019). Facebook's suicide prevention algorithm raises ethical concerns.

Holmes, A. (2021). 533 million facebook users' phone numbers and personal data have been leaked online.

Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59.

Nijhawan, L. P., Janodia, M. D., Muddukrishna, B., Bhat, K. M., Bairy, K. L., Udupa, N., Musmade, P. B., et al. (2013). Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research*, 4(3):134.

Orlando, C. M. (2020). *Peer Intervention with Suicidal Disclosures on Social Media: Does the Bystander Effect Play a Role?* PhD thesis, University of South Carolina.

Ritchie, H., Roser, M., and Ortiz-Ospina, E. (2015). Suicide.

Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., and Herrman, H. (2016). Social media and suicide prevention: a systematic review. *Early intervention in psychiatry*, 10(2):103–121.

Singer, N. (2018). In screening for suicide risk, facebook takes on tricky public health role.

Stack, S. (1988). Suicide: Media impacts in war and peace, 1910–1920. *Suicide and Life-Threatening Behavior*, 18(4):342–357.

Ting, K. M. (2010). Confusion matrix. In *Encyclopedia of Machine Learning and Data Mining*.

Zevenbergen, B. (2016). Networked systems ethics.