Detecting Self-injurious Content and Assessing Sources of Online Support on YouTube and Twitter Social Networks

A Thesis Presented to the

Department of Computer And Information Sciences

by

Muhammad Abubakar Alhassan

In Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy in
Computer and Information Sciences

December 19, 2022

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Dedication

This doctoral thesis is dedicated to;

- My late Father, Alh. Abubakar Alhassan and my mother, Aisha Muhammad, for their endless love and support.
- My country, ”Nigeria.”
- My sponsor, ”Petroleum Technology Development Fund - PTDF.”
- The family of Alh. Hassan Danbappa, for their vital support, love and prayers.
- My wife for her continued love, care and support.
- My siblings, especially my late brother, Aliyu Abubakar Alhassan, and friends for their informal support, prayers and best wishes during my academic research.

Acknowledgment

In the name of Allah the most gracious and the most merciful

Firstly, I would like to express my gratitude to my supervisors, Dr Diane Pennington and Professor Ian Ruthven, for their professional academic guidance and support; without them, completing this PhD study would not have been possible. Their constructive feedback and encouragement contribute immensely to the successful completion of this thesis. Undoubtedly, it was a remarkable experience working under their supervision. Moreover, I would like to thank members of our iSchool group for their feedback, especially during my seminar presentations. Lastly, the outstanding administrative support received from Julie Sobocinski, Nicole Hannigan, Zakia Majid, and other non-academic staff members, including the International Student Support Team (ISST), was acknowledged; I thank you all.

Abstract

Self-harm is a significant public health issue typically common in young individuals. The behaviour is associated with significant mental health problems like anxiety, depression and eating disorder. The issue of self-harming behaviour has been viewed as the 'tip of the iceberg;' only a few self-harming individuals seek clinical support. Although people who self-harm tend to be isolated and remain less socialised in offline settings, it was found that most self-injurers are socially active in online spaces, especially on social media sites. Self-harming individuals voice their behaviours and seek help on social media by creating and sharing content with online users. Hence, the need to investigate such content is critical. Little is known about the nature of self-harm content on online social spaces and what distinguishes such content from non-harmful content. Also, our knowledge of digital sources of information for self-injurers and their operational strategy on online social networks is insufficient. While the literature reported that members of society constantly misjudge self-harming individuals, there is an inadequate understanding of the public perceptions and attitudes on digital networks concerning self-harm behaviour. The objective of this research was to gain a better knowledge of (1) how YouTube and Twitter users discuss self-harm behaviour, (2) the views and opinions of online members regarding self-harm, and (3) the strategy of support organisations disseminating self-harm related information on social networks.

Additionally, this doctoral study aimed to propose and evaluate an automatic technique for detecting self-harm content in digital social spaces. The research investigation was performed to fill a knowledge gap, as past empirical studies analysing digital self-harm content on social media have primarily used qualitative techniques. While surveys and interviews are sources of data collection for many researchers in this field, this doctoral study sourced data from two popular social media sites (YouTube and Twitter). These platforms allow self-harming people to convey information that is hard to disclose in surveys or interviews. The study employed a mixed-methods approach and used *state-of-the-art* machine learning techniques to analyse the retrieved data. The analyses of self-harm content from the platforms revealed essential themes such as pro-self-harm, anti-self-harm and clean commentators.

Additionally, this doctoral study uncovered the different opinions of users concerning self-harm across the examined platforms. A model was proposed using supervised machine learning techniques that automatically classify comments showing self-harm signs. The classification tasks were performed in binary and multi-class settings. The model based on the binary classification achieves higher performance (precision and recall) accuracy. Its performance outweighs that of the model built in a multi-class scenario. On the other hand, support organisations engaging with self-harming people on Twitter social networks exhibit different strategies while disseminating information to support positive well-being. Although the study recognises the limitations of utilising YouTube and Twitter data, the analysis illustrated how the platforms were used to communicate self-harm behaviours.

Previous Publications

Parts of this doctoral research have been published in the peer-reviewed conferences listed below;

1. Alhassan, M. A., & Pennington, D. R. (2022, February). YouTube as a Helpful and Dangerous Information Source for Deliberate self-harming Behaviours. *In International Conference on Information (pp. 347-362). Springer, Cham.*
2. Alhassan, M. A., Inuwa-Dutse, I., Bello, B. S., & Pennington, D. (2021, July). Self-harm: Detection and Support on Twitter. *In ECSM 2021 8th European Conference on Social Media (p. 255). Academic Conferences Inter.*
3. Alhassan, M. A., & Pennington, D. (2021, June). Investigating Non-suicidal Self-injury Discussions on Twitter. *In International Conference on Social Media and Data Mining-ICSMDM. (p. 62-71)*

Previous Short-paper

1. Alhassan, M. A., & Pennington, D. (2020, March). Detecting critical Responses from Deliberate Self-harm Videos on YouTube. *In Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (pp. 383-386).*

Contents

List of Figures

List of Tables

List of Tables

Chapter 1

Research Introduction

1.1 Introduction

Deliberate self-harm (DSH), self-harm (SH), and self-injury (SI) are non-suicidal self-injury (NSSI) terms used to describe intentional harm to oneself with no intention to die but to cope with emotional difficulties associated with mental health problems and traumatic experiences [1]. However, the past few decades reveal a significant controversy among researchers regarding the definition of deliberate self-harm (DSH) [10]. The disagreement started when experts from different fields used the term differently in various studies [10].

Several studies defined DSH as the intentional harm to a body through self-injurious behaviours such as cutting, burning, or drug abuse with no intention to die [11–14]. It is a behaviour that some people (especially young individuals) consider as a form of coping with difficult emotions associated with various mental health problems like depression, anxiety, and traumatic experiences resulting from a child or sexual abuse [15]. While some scholars believe that intentional self-harm does not involve killing oneself [14, 16, 17], other researchers argue that the behaviour is correlated with the intention of suicide [18].

However, it has been reported that studies concerning self-harm conducted in the United Kingdom frequently use DSH, while research investigations completed in the United States or Canada commonly use NSSI [8, 12, 15]. Regardless of the intent, this doctoral study used both terms to represent any form of self-harm behaviour. Evidence from the literature confirmed that self-harm is prevalent in young people, especially females [3, 12, 15]. On the other hand, it was found that intentional self-injurious behaviour is among the common leading of death in young people between fifteen to nineteen years [19] and an existing study confirmed that this class of population is socially active on social media [20].

Meanwhile, there is an intense debate on the impact of social media on self-harming people [21,22]. As an online space, social media facilitates social connections worldwide and supports creating and sharing of digital multimedia content. Twitter, YouTube, Facebook, and Instagram are the leading social connection platforms many people use to post various information about mental health, such as depression [23] and pro-anorexia or eating disorders [24]. While adverse reports have been stressed in conventional media [25], social media is described as a place where vulnerable people are exposed to an atmosphere of triggering content that encourages or discourages self-harm [26].

According to many positive reports from young individuals, social media can also provide a space where they can come out of hiding, discuss otherwise private anxieties with peers, and receive online support [27]. It has been confirmed that self-harming people tend to use online spaces more often as opposed to people that are not self-harming. Young individuals with mental health issues, particularly those suffering from self-harm, exhibit different use of social media [28], with positive and negative reports [25,29].

Consequently, with the increase in online social interactions emerging through modern social media, the risk of exposure to self-injurious content has increased [30]. Among the critical areas of computer science and mental health research that require urgent attention are:

- Automatic detection of self-harm posts on social media [31]
- Identifying the critical sources of self-harm content on social media [32]
- Investigating what online members communicate (including their opinions) about self-harm [21,33]

- Examining the impact of online social support target for individuals that are self-harming [33, 34]

However, this chapter introduces the focus of this doctoral research, including the critical research questions, motivations, aims and objectives. The chapter then closes by outlining the overall thesis structure as shown in Table 1.1.

1.2 Background

Social media platforms support various *face-to-face* interactions digitally, permitting accessible ways of communication among people worldwide. While the number of users is increasing globally, most of these platforms enable individuals to share information, express their opinions and disclose experiences, individual concerns, and mental health issues. Research evidence confirmed that self-harming people use online spaces more often than those who are not self-harming [28]. Even though it is difficult for self-harming people to speak to parents or family members about their behaviours due to stigma and stereotypes [35], this challenging behaviour drives them to shift online to interact with the online community and seek support [25].

Therefore, online social settings are continually used for discussions regarding intentional self-harm. Considerable research examined self-harm information posted by online users of social media. In the last few decades, a notable amount of mental-health research, mainly related to self-harm, has been done using social media data. For example, the study of [36] thematically analysed images tagged as self-harm from Instagram, Twitter and Tumblr. While the authors found only a few posts pointing out that self-harm is attractive, none of the analysed posts actively promote online users to self-injure. Instead, the platforms were used to voice out complicated emotions creatively, conveying motivations to online users via text messages regarding self-harm prevention and recovery. Similarly, another study reported that young self-harming individuals could use social media as a venue to speak out and stop hiding, disclose their self-harming experiences, and exchange mutual support [37].

On the other hand, mental health issues associated with DSH, such as depression

[38], eating disorders [39], pro-anorexia and suicide have been extensively studied using real-world data from social networks like Twitter [40]. Although self-harm is not linked with suicidal intent, it has been reported that intentional self-injury is usually common in young populations [41]. While suicide and self-harm share some common risk factors, both behaviours are more prevalent among young individuals between 15 to 19 years old [3, 33].

Studies revealed that accessing medical support by self-harming people is challenging and viewed this critical problem as the *'tip of the iceberg'* issue [14]. Meanwhile, social media is one of the online resources that connect self-harming individuals from different places. It gives them a sense of community formation through social interactions. Therefore, people that are self-harming may turn to online mediums, especially social media to call for help. For instance, the research study of [42] discovered that specialised forums accessed by self-harming individuals help them maintain stable emotions to fight self-harm.

However, current studies reported that the use of social networks among self-harmers could positively or negatively impact their well-being [25]. The influences of social media on vulnerable young people, especially those that are self-harming, are critical to explore. On the positive side, social media platforms like YouTube, Twitter, and Instagram could be a pleasant online atmosphere for lonely and stigmatised individuals to interact with their peers. On the other hand, social networks could negatively affect their mental health [25]. For example, excessive use of social media could be linked to depression and anxiety, which are all associated with self-harm [21]. Sharing harmful content could promote dangerous behaviour among online members and normalise or reinforce intentional self-injury [37].

Consequently, what young people post and view on social media are more likely to be associated with their risk of self-harm. According to co-construction theory [43], young people's post on digital social networking platforms reflects their offline views and behaviours. This implies that young individuals who often share self-harm content on social media are more likely to participate in self-harm activities offline. Social media posts related to self-harm could reliably indicate that the poster has mental health problems, and regular viewing of self-injurious posts could strengthen the existing offline self-harm behaviour. Vulnerable young people, especially those struggling with self-harm, may experience more severe reactions.

Some researchers presented an excellent illustration of a grounded, descriptive study examining the prevalence and distribution of online posts associated with the risk of depression and suicide in adolescents [44]. The investigators analysed online posts regarding a severe self-injurious act, popularly known as the Blue Whale Challenge (BWC). Thus, the BWC is a *"suicide game"* in which young individuals are persuaded to engage in self-injurious behaviour that results in ending their lives. It was acknowledged that the BWC is linked to several deaths of teenagers worldwide.[1] Consequently, exposure to such games on social media could significantly impact vulnerable young people.

However, identifying self-harm-related social media posts is a challenge when creating digital interventions. As observed from the literature, only a few self-harming individuals seek clinical support [14]. Therefore, it is crucial to (1) identify vulnerable users in need of urgent support [25, 45], (2) gain adequate knowledge of self-harm discussions online [21], and (3) examine the impact of informal support sources available to self-injurers on social networks [46, 47].

This doctoral study analysed self-harm content from YouTube and Twitter social networks and answered the research questions explained in Section 1.3. The idea of utilising self-harm posts from both social media platforms was to understand how self-injurious content proliferated from one platform to another while exploring the views and opinions of online users. For instance, video channels uploading videos discussing self-harm on YouTube could be linked to Twitter and accessible to general users of the online networks. In addition to other social media applications, current studies demonstrate the importance of YouTube and Twitter online platforms to self-harming individuals [33, 37]. Even though related research studies examined these platforms, further research is needed to add to the body of knowledge in this area.

[1] https://www.bbc.com/news/blogs-trending-46505722

1.2.1 YouTube and self-harm

As a free social media platform, YouTube is one of the leading popular online search engines that allow individuals or organisations to share videos with online members worldwide.[2] It has been reported that YouTube is among the leading online resources that facilitate social connections through searching, watching and commenting on videos.[3] In addition to searching for videos, YouTube offers an attractive interface and features for people to search and play videos, like or dislike a video, and respond to the video uploader via comments. To self-harming individuals, the literature shows that YouTube could be a helpful or harmful platform [37] and videos presenting DSH information attract many views and comments from the target audiences [48].

While few studies examined self-harm content on YouTube, the study of [49] analysed first aid videos available on YouTube and discovered that not all self-harming people look for first aid advice on the platform. Overall, videos containing first-aid tips regarding self-injury could reinforce self-injurious behaviour and support the idea that clinical support is not required to treat or recover from self-injury. People who view first-aid information videos regarding NSSI tend to rate the videos favourably. In another similar study, comments on videos presenting self-harm information were qualitatively explored [37]. The authors used coding rubrics to examine the responses regarding videos related to self-harm. The authors found that most comments made by the viewers conveyed personal self-injurious experiences and did not include recovery information. Also, fewer commentators mentioned they had recovered or reported the urge to recover. The most repeatedly occurring video comment was self-disclosure, in which commentators discussed their self-injurious experiences. However, the researchers' approach could not be applied to an extensive collection of video responses. Further effort is needed to examine self-harm discussion on a larger scale.

Although existing research indicated that YouTube is among the prefered choice of social connections for self-harming users [48], it is not clear how the online audiences react to those videos and how they respond to the videos [37]. Meanwhile, in a spe-

[2] https://www.alexa.com/

[3] https://www.statista.com/topics/2019/youtube/

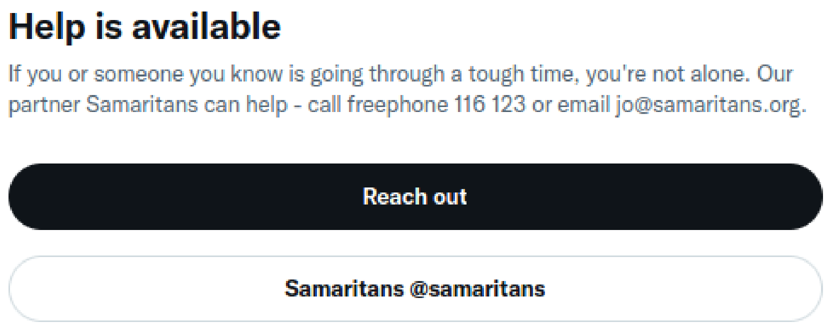
cialised forum dedicated to self-harming individuals, a team of professional moderators must monitor users' responses to identify at-risk online members who require urgent support [50]. For free and general social media sites like YouTube, tracking video responses signalling that a commentator needs urgent attention is challenging. This is due to the rapid growth of online content that could bury people battling self-harm's voices.

Additionally, from the available existing studies, it is observed that YouTube is one of the preferred online social platforms for self-harming individuals. Despite the current studies, the complete scope of self-harm content and the nature of self-harm activities, especially on YouTube, remains unknown. Also, how people accessed self-harm videos and their perceptions of the videos are unclear. Accurately interpreting the context and meaning of self-harm-related discussions across some online mediums, especially YouTube and Twitter, can give insight into the roles and influences of self-harm activities.

1.2.2 Twitter and self-harm

On Twitter, people create and share content related to their health condition and social and political activities. The amount of data created on Twitter is rapidly increasing. Research evidence confirmed that self-harming people use the Internet (social networks) more often than non-self-harming individuals [28]. The particular purpose of using the social networks concerning self-harm among self-injurers is both *internal* and *external* validation of *'self'*, a desire to be regarded as a legitimate individual of value, and to be recognised as more than simply a self-harming person [51]. However, following the literature, this study has seen the *positive* and *negative* reports of how individuals suffering from mental health issues interact with social media platforms, particularly self-harming people [25].

In a recent comparative study, it was reported that Twitter is unique in two ways from Tumblr and Instagram [36]. The images on Twitter were more interactive and were not seen tagged as self-harm on Instagram or Tumblr. As examined by the authors, one of the most prevalent reasons for posting self-harm images was the expression of



Help is available

If you or someone you know is going through a tough time, you're not alone. Our partner Samaritans can help - call freephone 116 123 or email jo@samaritans.org.

Figure 1.1: Twitter help notification

one's feelings — a most popular Tumblr theme [36]. Also, posters frequently included quotations to frame their pain. Those quotations reflected low self-esteem and harmful connections with oneself and shared devastating emotions. Furthermore, interpersonal distress was prevalent. In other words, the researchers found frequent notes discussing how some people had treated the poster—however, those posts highlighted loneliness and the difficulty of trusting others [36].

Another research study found celebrity influence regarding self-harm on Twitter [33]. From the qualitative study, the author identified twenty-one celebrities mentioned in posts associated with DSH. The narrative made a solid case for and against acquiring support from others. The tweets examined by the researcher were not related to a conversation or pointed to another person; instead, they were made to educate or support online members [33]. The researcher found that most tweets indicated self-harm-related issues such as scars, seeking attention, or crying for help.

In many instances, some users directly reply to tweets where the tweeter user posts a message to seek online support. Although the researcher demonstrates the significance of Twitter in creating a medium for self-harming people to seek online assistance, the study was limited to a few instances of self-harm tweets of not more than 140 characters in length. While this number has been doubled (up to 280 characters) recently[4], more effort is needed to investigate self-harm related discussions in a larger scale and the views of the tweeters using current *state-of-the-art* approaches.

[4] https://developer.twitter.com/en/docs/counting-characters

Meanwhile, when users search Twitter for self-harm-related information, the platform will display a notification and offer a hotline and organisation to contact for online support, see Figure 1.1. While the platform delivers essential support through helpful organisations to self-harming users, forming an online community, knowledge and awareness constructed through the support process could normalise self-harm [33]. The traditional media has stressed this issue in various reports and described online social media as a space where users are hauled into an enveloping environment of harmful content, such as images that can trigger or invite online users to self-harm.

For example, the story of **Molly Russell**, that committed suicide in 2017, shows the link between the cause of her death and the content she accessed on social media[5]. Consequently, the government of the United Kingdom advised social media organisations to take greater accountability for dangerous online content that could promote self-harm, suicide or any dangerous behaviour. Even though there are several reasons underpinning DSH, in this doctoral study, it is argued that online support on social networks could facilitate proactive measures to mitigate the negative effect of self-injurious behaviour.

A research investigation reported that young people from seven European countries agreed with at least one of the eight possible reasons causing or underpinning self-harm [10]). While the examined participants were free to report numerous reasons for self-harm, it was uncommon for a distinct motive to be linked to a particular self-harm behaviour [10]. Consequently, the research study found that nearly 71% of the participants agreed with the following reason: *"I wanted to get relief from a terrible state of mind"* Accordingly, this observation was validated in a similar study [52]. Interpersonal motives like *"I wanted to frighten someone"* and *I wanted to get some attention"* were found as motivations for self-harm behaviour [52]. Participants' responses in the study of [10] and [52] are associated with self-injurer's psychological conditions.

Therefore, it is deduced that if the purpose of self-harming is to get attention, which is harmful to one's physical and psychological state [10], then social networks like Twit-

[5] https://www.bbc.co.uk/news/newsbeat-47131178

ter provide a space for self-harming users to tweet about self-harm and seek help. While people looking for online support were shown to be at higher risk of dangerous behaviour than those pursuing offline help [28], the effect of posting self-injurious related content on social media platforms could be considered a *double-edged sword* issue. Firstly, it allows the investigation of self-harm content through data analytics techniques. Secondly, the influence of such content expresses the voice of isolated self-injurers seeking support. Inspired by these effects, this study aims to evaluate injurious-related content on Twitter and answer the following research questions presented in Section 1.3.2.

1.3 Research questions

1.3.1 YouTube

In the phase one study, this dissertation formed a number of research questions to achieve the aim and objectives of the study.

1. *Who uploads videos on YouTube concerning self-harm, and how are the videos rated?*

   Addressing this question will increase our understanding of the different groups of people involved in sharing videos about self-harm on YouTube. While analysing the characteristics of the videos, the study aimed to understand the video's *ratings* by the audiences. Hence, in an attempt to answer this question, the below silent question pondered over.

   - *To what extent are deliberate self-harm videos accessible to young audiences?*

     Various videos discussing self-harm are freely available on YouTube. While the video presenters tend to share their self-harm experiences and encourage seeking support, some videos could trigger and promote self-harm. Additionally, some videos may display graphic content and other information that violates YouTube's terms and conditions. In other words, YouTube restricts access to triggering content, and the platform's policy prohibits viewers from uploading video content that promotes self-harm. Therefore, this question aimed to explore the accessibility of such videos to young audiences (those below 18 years). While addressing the research question, the study raised the following question.

2. *What are the discussions surrounding YouTube videos presenting information about self-harm?*

   In this question, the idea was to examine the user-generated content discussions (viewer's comments) on videos about self-harm on YouTube. Identifying the common topics surrounding videos about intentional self-injury could inform better decision-making among health experts and social media organisations. Also, parents or families offering support to self-harming individuals will understand the role of social media in promoting or preventing self-harm. Nevertheless, this dissertation addressed the above research question using a current state-of-the-art topic analysis technique and uncovered themes associated with those videos.

3. *What sentiments do users express concerning videos about self-harm on YouTube?*

   YouTube provides an essential feature (commenting) for users to interact with one another. Although the platform did not have an interactive feature like emojis for users to express their feelings, the comment feature allowed the audiences to represent their opinions or interest concerning a video. Therefore, in this question, our study investigated the sentiments expressed by the people viewing deliberate self-harm videos on YouTube. Intuitively, commentators' opinions could be used as a yardstick to measure the degree of helpful or harmful videos that encourage or discourage intentional self-harm.

4. *How can we detect critical comments from DSH videos on YouTube?*

   The responses from a video contained information that could provide our research community and the YouTube media teams with opportunities to reach vulnerable self-harming users. Our research study viewed this question as a *binary* and *multi-class* problem and propose a model to detect critical comments (based on severity - see C.1) from the video responses.

1.3.2 Twitter

This doctorate research addressed the following research questions in order to achieve the aim and objectives of the doctoral thesis.

1. *What are people discussing regarding intentional self-harm on Twitter?*

   People from diverse backgrounds around the world are tweeting about self-harm on Twitter. There is inadequate knowledge of what the general public is discussing about self-harm. In this question, the goal was to understand the nature of self-harm discussions and how it evolves on Twitter. Similar to the existing research [38], the idea was to examine the behavioural attributes associated with the tweeter's post by studying the sentiments and linguistic tones.

2. *What strategies support organisations operating on Twitter to facilitate and encourage self-harm recovery*

   Sources of support for self-injurers on Twitter use distinct strategies to promote self-harm prevention recovery. There is no sufficient evidence to showcase the impact of support organisations on social networks towards self-harming individuals. While utilising longitudinal data from a set of organisations supporting self-injurers on Twitter, the objective of addressing this question was to uncover the distinct strategies used by those organisations and inform standard best practices.

3. *How do followers of the examined organisations opinionate on the information shared by the support accounts over time?*

   The Twitter social network's subscribing (follow) feature allows users (followers) to receive tweet updates from the account handle. This attribute promotes a sense of community between the followers and the support handles. This question aimed to understand the relationship between followers seeking help about self-harm and online support organisations. This enables the analysis of the frequency with which support organisations are socially engaging with online members.

1.4 Aims and objectives of the study

This doctoral thesis aimed to understand the nature of deliberate self-harm content on social media and what online users communicated around the behaviour, particularly on YouTube and Twitter social networking sites (SNS). More precisely, the objectives are as follows:

- To explore the current existing studies in the area and identify critical areas of improvement.
- To attain an in-depth qualitative understanding of the different online users posting self-harm content on social media.
- To understand the attitudes and views of social media users towards self-harming individuals.
- To propose and evaluate a technique to detect critical self-harm responses automatically requiring urgent attention.
- To examine the impact and strategies of organisations reaching out to self-harm users on social media.
- To understand users' responses to online support organisations fighting self-harm on social networks.

1.5 Why YouTube?

The motives behind choosing YouTube in our research study are twofold. Firstly, YouTube is one of the leading popular social media platforms with more than two billion users worldwide.[6] Following its commencement in 2005, YouTube has become popular and considered the most prominent online video platform globally. The giant video social media site is now a notable online space for numerous audiences globally, with a broad selection of user-generated content covering different categories such as

[6] https://www.statista.com/topics/2019/youtube/

entertainment, educational, and health-related videos. Evidence confirmed that self-injurers share videos on YouTube and disclose and share personal stories to support target audiences.

Secondly, the literature inspired us to focus on YouTube as the platform provides an online avenue for self-harming people to view NSSI videos and share experiences through comments. YouTube's interactive features such as comments, likes and dislikes facilitate interactions between users and the video uploaders or presenters. These features represent unique ways viewers favourably or unfavourably viewed NSSI videos. For example, while viewers favourably (like) some videos discussing self-harm on YouTube, their responses to those videos could be harmful or helpful to vulnerable audiences [37].

1.6 Why Twitter?

Twitter is ranked by the statistica[7] as one of the leading social networking platforms in terms of active users. Although the platform was launched in 2006, it is still considered among the universally leading and popular SNS. However, this dissertation considered sourcing data from Twitter because of a number of reasons. Firstly, the platform's distinct features (listed below) make it a suitable data source for many researchers around the globe.

- Twitter is among the leading social networks, with 290.5 million monthly users in 2019 worldwide and was predicted to be 340 million by 2024. Users include political leaders, government and non-governmental organisations, academics, health professionals and many more.

- Regardless of the categories of users, people tend to produce nearly 500 million posts per day on Twitter [53].

- It serves as a data source among researchers through the Representational State Transfer (REST) or streaming APIs that any user could access.

The number of active users plays a significant role in choosing the appropriate SNS

[7] https://www.statista.com/topics/737/twitter/#topicOverview

data. For example, in November 2016, around 24% of US adults, including men and women, were using Twitter. Most of the surveyed users were between 18 to 29 years old, and many hold college degree certificates [54]. Meanwhile, in April 2021, 38.5% of the global Twitter users were between 25 to 35 years, while only 24% represent users under 24 years.[8] However, this indicates that the diverse population group could be on Twitter, making it a suitable space for sourcing research data. In addition to the above characteristics, this study considered Twitter due to supporting evidence discovered from the existing research that highlights the relevance of Twitter in studying self-harm content or users experiencing self-harm [40]. An investigation conducted by another researcher informed us that Twitter could be an online space for self-injurers to seek *"support for and from others"* [33].

1.7 Why this research is important?

The importance of this doctoral study cannot be understated. The significance of this research is threefold and discussed as follows.

1.7.1 Online safety

Although people of different ages use social media, it affects individuals' lives differently. It is discovered that most social media users are young individuals of eighteen to twenty-nine years [20]. Evidence revealed that nearly 18% of this population engaged in deliberate self-harming [14], a behaviour associated with early signs of suicide. Meanwhile, there are concerns about the safety of digital online social media platforms. Users could be exposed to triggering content that could normalise dangerous behaviours or increase the likelihood of suicide.

Exploring self-harm content to provide a safe online environment and fight against portraying harmful content is essential. Therefore, this research study will investigate the influence of self-harm content, particularly on YouTube and Twitter platforms. In addition to identifying the underlying sources of self-injurious content, it is crucial to

[8] https://www.statista.com/topics/737/twitter/#topicOverview

understand (through topic analysis) the actors involved in sharing dangerous and helpful information on those platforms. Hence, the study will provide evidence regarding online users' sentiments relating to self-harm and add to the current research on the general nature of self-harm content and social media.

1.7.2 Online support

Social media connects people of different attitudes from diverse locations. It allows people to communicate their feelings and activities in different ways. Different people are participating in sharing information related to self-harm on online social spaces. The anonymity of social networks encourages self-harming people to speak about their feelings and experiences that are difficult to express in offline settings. Therefore, it has become necessary to distinguish between online users seeking support versus those promoting self-harm behaviour.

Consequently, another significance of this research study is to assess online social support for self-harming people accessible through charity organisations such as the Samaritans[9] and Minds[10]. These are organisations based in the united kingdom and recommended by the National Health Service (NHS).[11] Although they operate on social media to fight against self-harming behaviours, the knowledge of the strategies they use in providing support is limited. Hence, it is essential to examine the behaviour of those organisations to increase our knowledge and inform best practices.

1.7.3 Inform best practices and support decision making

Social media companies have terms and conditions for people to accept before joining and using their services. Posting and sharing self-injurious content is against the policies of those platforms. Even though these companies are taking measures against users posting triggering content, some individuals continue to post self-injurious content to promote the behaviour. Therefore, another critical aspect of this research is to present

[9] https://www.samaritans.org/

[10] https://www.mind.org.uk/information-support/types-of-mental-health-problems/self-harm/about-self-harm/

[11] https://www.nhs.uk/mental-health/feelings-symptoms-behaviours/behaviours/self-harm/getting-help/

a suitable approach that automatically detects content (comments) showing signs of self-harm to facilitate content moderation effectively.

Similarly, government policies change over time to improve the citizens' health, safety, and protection. With significant concerns about the proliferation of harmful content on online social connections, it is essential to provide research evidence to support decision-making and policies governing social media operations. This study is crucial as the findings could practically support policies to protect vulnerable children or young people in online spaces.

1.8 Organisation of the thesis

The present chapter introduces the doctoral study and explains the background of the PhD research. While explaining the significance of the study, the chapter explained essential questions considered in the study, including the research aims and objectives. Therefore, starting with chapter two, this section describes the general structure of this thesis, as shown in Table 1.1.

Table 1.1: Thesis organisations

Chapter	**Content**
Chapter 2	This chapter reported the definitions of fundamental terminologies used in the research study. While the chapter explained the procedure followed to explore the literature and identify relevant research studies (2.1.2), it reports the current knowledge in this growing field. Accordingly, the chapter provides the review findings of the examined research works and outlines the current approaches (2.4.2) utilised by several researchers from different fields. Similarly, the chapter discussed critical areas (2.4.5) needing further investigation in this research area and how this doctoral study plans to fill the research gaps found in the literature (2.5).

Chapter 3	This chapter describes the research methodologies. It then discusses the research design (3.3) and the Application Programming Interfaces (APIs) used in both research phases - YouTube (3.4.2) and Twitter (3.4.3) studies. While the chapter explains the research methodology adopted in this dissertation, it also reports the data analysis procedure and current *state-of-the-art* techniques and research tools used (3.7).
Chapter 4	This chapter reported the findings of the phase one study. It discusses the different sources of videos concerning self-harm on YouTube. Correspondingly, the fundamental themes surrounding viewers' responses to deliberate self-harm videos on YouTube were discussed in this chapter. Also, the chapter explained the commentator's opinion on the examined videos and proposed a machine learning model to detect critical responses needing urgent attention from the YouTube content moderators.
Chapter 5	The findings of the phase two study are explained in Chapter 5. The chapter outlines the diverse group of tweeters who participated in Twitter self-harm-related discussions. It also presents the nature of the discussions concerning DSH and its associated topics. Also, the chapter highlights the strategies operated by different support handles on Twitter to combat self-harm. It then discussed the nature of the online activity (social engagements) between online users and support organisations helping self-harming individuals on Twitter.
Chapter 6	This chapter presents the doctoral research discussions, including the research implications and limitations. In other words, the chapter reports the value and importance of the research outcome. While providing the doctoral research implications, the chapter outlines the study's limitations.

Chapter 7	This chapter summarises the overall study and provides conclusions regarding its outcome. It also recommends a set of future works to be considered in this area of research.

1.9 Summary

This chapter presented the entire doctoral thesis. It established the background and justified the significance of the study. It also explained the research questions, aims and objective of the study. While deliberate self-harm has a considerable risk of potential suicide, YouTube and Twitter social media platforms remain among the leading innovative mediums of social interactions that need further investigation to understand the nature of self-injurious content on the platforms. However, the chapter described a summary of the thesis's structure. In the next chapter, this doctoral thesis examines the research topic's existing studies and summarises current knowledge gaps that need to be filled.

Chapter 2

A Scoping Review of Deliberate Self-Harm Contents on Social Media

2.1 Introduction

As the name implies, the scoping review was conducted to determine the scope of the current studies performed by other researchers on the influence of self-injurious content on social media platforms. Thus, the study reviewed the existing works, identified the current methods and analysis techniques, and discovered the challenges and future directions requiring more attention. Moreover, this chapter presents the *step-by-step* procedure followed in analysing the existing studies. Searching the relevant articles from reliable sources is one of the significant steps in surveying the literature.

Consequently, this study started by establishing the literature search criteria. The boundary indicating the inclusion and exclusion of the critical phrases used in searching the related studies was defined. While stating the goal of our review, the chapter also discusses the basic concepts, attributes, and theoretical models of self-harm behaviour. In the subsequent sections, this chapter presents the relationship between self-harm and suicide and how researchers described the duo in mental health research. Furthermore, this chapter explains the approaches used by the existing relevant research and highlights the strengths and weaknesses of the examined studies. Similarly, the chapter presents evidence related to the benefits and risks associated with using social media among self-harming users. The review focused on the central topic concerning self-harm and social media and presented the underlying arguments and gaps that need to be filled.

2.1.1 Aim of the scoping review

In general, the literature review of this dissertation aimed to investigate the current existing studies describing the impact of social media concerning deliberate self-harm. Also, the review was aimed to:

- Identify, analyse and describe the current research findings relating to the role of social media and self-harm.
- Analyse existing studies on the effect of self-harm content on social media on vulnerable users and the general public.

Therefore, the review process began with establishing the inclusion and exclusion criteria for selecting relevant research studies to achieve the aim mentioned above.

2.1.2 Searching the literature

This study iteratively searched and reviewed relevant literature from different online databases. The literature on self-harm and social media is crucial to this research. Thus, the survey and analysis of the existing studies were carried out simultaneously with the data collection process. In other words, the research reviewed similar works while collecting and analysing real-world data from Twitter and YouTube social media platforms. This study investigated the current findings to identify issues associated with the existing techniques and suggest a novel approach that can be employed to enhance future work. When searching the literature, this study considered different online databases as shown in Table 2.2.

Literature was searched using terms related to NSSI and logical operators. For example; *Self-harm AND social media, Self-injury AND social media, Self-harm AND* *social networking sites, Deliberate self-harm AND social media, Deliberate self-harm AND social networks.* Most online database provides researchers with options for retrieving content in different formats. In our research, the search results of each online source were extracted in Microsoft excel formats (.xlsx/CSV).

2.1.3 Inclusion and exclusion criteria

Table 2.1: Criteria for inclusion and exclusion of related works

Code	Assessment criteria	Category
001	Studies examined social media use or contents concerning self-harm and used qualitative or quantitative data.	Included
002	Research papers were excluded if they do not contain information about self-harm and social media.	Excluded
003	Studies are added if they report relevant information related to self-harm and social media.	Included
004	Similar research was included if they investigate social networks and discusses their impact on people that are self-harming.	Included
005	A research that focused on digital intervention such as mobile applications to prevent self-harm were excluded.	Excluded
006	Research papers published in a different language other than English were removed.	Excluded

In this research, the review process considers selecting relevant existing research according to the criteria explained in Table 2.1. Thus, the first step considered in assessing the existing studies was choosing research papers written in English. Titles and abstracts of the existing studies were checked and examined for eligibility. The search strategy used different self-harm and social media phrases due to Boolean operations used in searching the related works. Example of such terms includes *'self-injury'*, *'self-injury'*, *'non-suicidal self-injury'*, *'deliberate self-harm'*, *'social networks'*, *'social* *media'*, *'YouTube'* and *'Twitter'*. Relevant papers were extracted and analysed using the Excel software. Studies reporting the effect of social media content and its impact on self-harming users were added.

Table 2.2: List of online sources

Online database	**Number of papers**
Scopus	215
PubMED	511
ProQest	472
PsycINFO	179
Web of Science	294
Total	**1,671**

Furthermore, the review excluded papers on using a mobile application for self-harming individuals. This is because our research focused on the contents of social media associated with NSSI. Research papers on social media's effect on mental health issues such as depression and anorexia were also removed. Articles published outside the defined period were dropped. Similarly, the review process also removed studies highlighting issues related to cyberbullying on social media.

However, the review analysis retrieved 1,671 research papers from different online databases - see Table 2.2. As shown from the figure, 879 were duplicates and therefore removed and obtained a list of 792 relevant articles. The review process screened these studies and excluded research papers that failed to meet the inclusion criteria set for this study. Thus, 71 papers not explicitly discussing DSH and social media were removed. Additionally, 25 papers were pulled out of the review as they focused on digital interventions - such as mobile applications for self-harming people.

Moreover, the review removed 97 irrelevant articles and 403 papers focused mainly on DSH or NSSI. Similarly, a few non-English papers (only 8) and 159 studies that failed to focus on social media's direct impact on self-harming individuals were excluded from the review. This results in 29 studies that meet our inclusion criteria. While performing an in-depth review, this study found and included 4 relevant research papers. The related studies included were published between 2011 to 2021 years. Accordingly, the searching, screening and inclusion process was repeated to include recent papers published in 2022. Consequently, 3 papers were found and added to the review. Overall, a total of **36** relevant research papers were included in the study.

2.2 Deliberate Self-harm (DSH)

2.2.1 Self-harm Definition

Various definitions of self-harm have significantly affected our research community due to the arguments surrounding the intent and form of self-harming behaviour [8]. Another contributing factor to the diverse concept of self-harm is that researchers use different terminologies to convey the meaning of self-harm without considering the direct or indirect forms of the behaviour. For example, self-harm, self-injury, deliberate/intentional self-harm (DSH) and non-suicidal self-injury (NSSI) are all terminologies referring to self-injurious behaviour. It is vital to look at the meanings of these terms and how they have been used in different investigations. Nock defined *self-injury* as any behaviour (performed deliberately) that could lead to psychological or physical harm to oneself without suicidal intent [55].

Moreover, Nock further describes NSSI as the intentional destruction of one's body tissue through (1) cutting of the skin, (2) burning, or (3) bruising with no intention to end life. This behaviour is critical to public health and notably common in young people, especially girls or females. However, some of the existing literature view the nature of DSH as either *direct* or *indirect* self-harming behaviour [39, 55].

Direct self-harm

On one hand, a form of *direct* and intentionally damaging of body tissue by cutting or burning the skin without suicidal intent is considered to be a *direct* SIB [55]. For example, a person can directly cut his/her skin with a sharp object and this requires no steps or process that leads to self-injury. Direct self-harming involves doing any act of self-injurious behaviour physically. Regardless of the suicidal intent, it is characterized by the deliberate destruction of body tissues [15]. On the other hand, a person may

Table 2.3: DSH definition as suggested by [8]

> *“An act with a non-fatal outcome in which an individual deliberately did one or more of the following;*
>
> - *Initiated behaviour (e.g., self-cutting, jumping from a height) which they intended to cause self-harm;-*
> - *Ingested a substance in excess of the prescribed or generally recognized therapeutic dose;-*
> - *Ingested a recreational or illicit drug that was an act the person regarded as self-harm;-*
> - *Ingested a non-ingestible substance or object.”*

excessively drink alcohol or take overdose medication which *indirectly* will take some stages (due to chemical changes in the human body) that would lead to *indirect* self-harm [55].

Indirect self-harm

Indirect Self-harm is a situation whereby an individual will maltreat or abuse himself or herself physically without directly cutting the skin. One common form of this behaviour is the eating disorder [56]. As opposed to direct self-harm, indirect self-harm tend to show different presentations. As pointed out earlier by [57], it is considered a feeling of expression or an act of behaviour punishing oneself for regaining self-control. Indirect self-injury can be understood as a *self-destructive* action but does not involve an intentional physical injury. Some researchers emphasized that *indirect* self-injury is practically significant in the hospital, constant or persistent, renders a source of severe discomfort to medical practitioners or family members and possesses the potential to create signs of visible damage over time [58]. Misuse of substances, abusive behaviour, eating disorders, continued involvement in abusive relationships, and engaging in immoral behaviours all fall into a category of *indirect* self-harm.

However, even though DSH is often replaced or interchanged with NSSI, DSH is usually used as an umbrella term for SIB. The term is commonly used in research investigations conducted in European countries Australia [56]. Table 2.3 presents a standard definition of DSH behaviour suggested by [8]. This definition was used in a comparative study of seven European countries (Australia, Belgium, England, Ireland, Netherlands, and Norway) that reported the reasons and prevalence rates of self-harm in young people from those countries. As the first of its kind, their study is considered to be pioneer research that reports Child and Adolescent Self-harm in Europe (CASE) [8]. From their investigations, DSH is twofold common in females compared to males. Many self-harming young individuals do not seek support from clinical services or speak to anyone else about their behaviour, and cutting was the most common form of SIB reported by the participants [17].

Meanwhile, the NSSI terminology appears in most of the research studies carried out between Canada and the United States [41, 55]. More recently, according to the *International Society for the Study of Self-injury - ISSS* [1], the NSSI is defined as *"the deliberate, self-inflicted damage of body tissue without suicidal intent and for purposes not socially or culturally sanctioned"*. This definition consists of 4 essential parts. In the first part, any harm resulting from self-injury is regarded as deliberate and is associated with the expected result of the behaviour itself. The second part deals with the actions that result in direct physical injury, such as cuts or marks on the skin. The third part considered self-injury behaviour as different from suicidal behaviour. Lastly, the fourth part of the definition viewed behaviours that resulted in physical damage to the skin and are recognised in society as a form of cultural or religious beliefs as non-self-injurious behaviours. Moreover, this definition is consistent with the concept of NSSI described by Nock [55]. Self-harming people tend to use different behaviours to injure themselves. Many researchers have explored the nature and frequency of self-injurious behaviours people use to cope with the terrible state of mind, as discussed in the next section.

[1] https://itriples.org/what-is-self-injury/

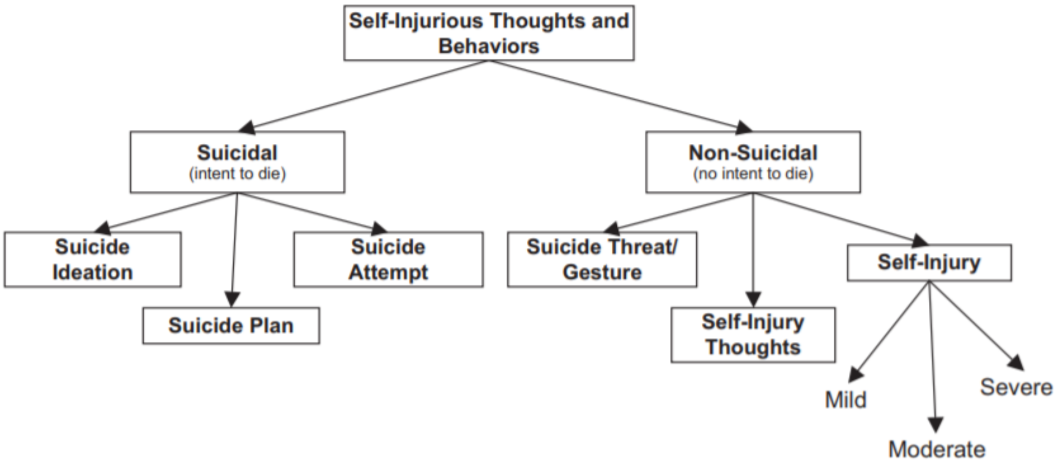


Figure 2.1: Self-harm and suicide classification, Source: [1]

2.2.2 Self-harm versus suicide

It is imperative to highlight the pieces of evidence demonstrating the link between self-harm and suicide. Suicide and self-injury are serious public health issues among teens, with a high prevalence of self-injury among young people and suicide becoming the world's second most prominent cause of death for young persons worldwide. Hereditary and behavioural, psychological, family issues and social and cultural influences are significant contributors to self-harm and suicide [3]. There is an increasing debate among researchers about the relationship between suicide and self-harm [59]. Most arguments are centred around the intent of the person who self-harms. For example, a study found that individuals engaged in self-harming can have suicidal thoughts, leading to the risk of suicide [3]. Looking at Figure 2.1, it is observed that the thoughts associated with self-injurious behaviour are broadly classified into two (suicidal and non-suicidal) by [1].

The author suggests that any direct self-harming behaviour performed by a self-injurer could either be suicidal or non-suicidal. Self-injury performed with an intent to die fundamentally consist of (1) suicide ideation in which a person has thought of committing suicide. (2) a suicide plan in which the self-harming individual consider a way to end their life; and (3) a suicide attempt by engaging in any harmful behaviour with an intent to die. Moreover, the non-suicidal category is divided into three subcategories. The suicide gesture is a behaviour that influences the self-harming individual to believe they perform the act to end their life while they do not have such intention. They engage in suicide gesture attitudes to communicate their feelings and ask or look for support [60]. While self-injury refers to direct, intentional harm to oneself with no intention to die, having thoughts on engaging in a behaviour that could result in self-injury is called self-injury thoughts.

2.2.3 Self-harm behaviours

In many of the reported studies, evidence discovered that people engage in NSSI using various methods [8,61]. Among these methods are burning the skin and scratching the skin until it bleeds. Several investigations confirmed that the most popular method used by self-injurers in hurting themselves (commonly on the stomach, arms or legs) is cutting using a razor blade or knife [14,55]. Meanwhile, other self-harm techniques such as pulling hair, picking wounds, and biting or hitting oneself are less frequent among self-injurers [61]. Undoubtedly, some of the less frequently self-harming techniques (like wound picking and biting oneself) are regularised and normal in our communities. Therefore, adding those methods to research investigations could be why a high prevalence of self-harm was discovered in some of the existing studies.

The diverse range of self-harming behaviour and the intent or motive underpinning the behaviour contribute to self-harm's different meanings [8,62]. This factor also affects the prevalence rate of SIB in our communities. However, it is essential to measure SIB as this will give more insights into understanding the intent of the self-harming person and the appropriate way to provide medical support and advice for sufficient recovery. In line with this, researchers proposed different scales of measuring NSSI. Some of these scales are suitable for quantitative or qualitative research as they provide questions to help researchers understand the self-harm behaviour of the participant under investigation. For example, some researchers proposed the self-injury thoughts and behaviour interview SITBI that examine the presence and frequency of different self-injurious behaviours [63]. While SITBI evaluates different self-injurious behaviour comprehensively and consistently, the SIBTI also assess critical elements associated with suicide such as suicide plans, suicide ideation and attempts.

Who self-harms?

Self-harm is one of the public health issues affecting children and young people worldwide. It is a behaviour of punishing or harming oneself through cutting, burning, and other related techniques of injuring oneself to cope with stress or intricate feelings. Although adults engaged in NSSI, evidence indicates that only 1% of adult people are self-harming as opposed to young individuals with 23% [64]. While most psychologists and medical experts believed this behaviour is typical in young people [3, 61], most studies found differences in gender as females have a high proportion of self-harmers than males [8]. Moreover, the age of young self-harming people was considered differently by many researchers. Several studies in this field considered young people to be between 14 to 25 years [52, 65, 66]. Although there is an age difference and many other factors (such as self-harm definition and study approach) that affects the prevalence of self-harm, over 30,000 young self-harming individuals between the age of 15 to 16 participated in a large comparative study conducted in seven European countries [8].

Moreover, one in five girls and approximately one out of ten boys participants in the child and adolescent self-harm in Europe (CASE) investigations had thought of injuring themselves but decided not to. Overall, the findings of [8] show that many young individuals between the age of 15 to 16 years from the surveyed European countries experienced more rates of DSH and self-harm related thoughts. In other words, the CASE study discovered a high repetition of self-harm among the participants. However, even though reports indicate that nearly 23% of young people are self-harming, studies discovered that the reasons they engaged in this behaviour could be interpersonal or intrapersonal, as discussed in the next section.

Why self-harm?

Several studies reported reasons underpinning self-injury behaviour. Researchers from different disciplines, including medical professionals, have wondered why people self-

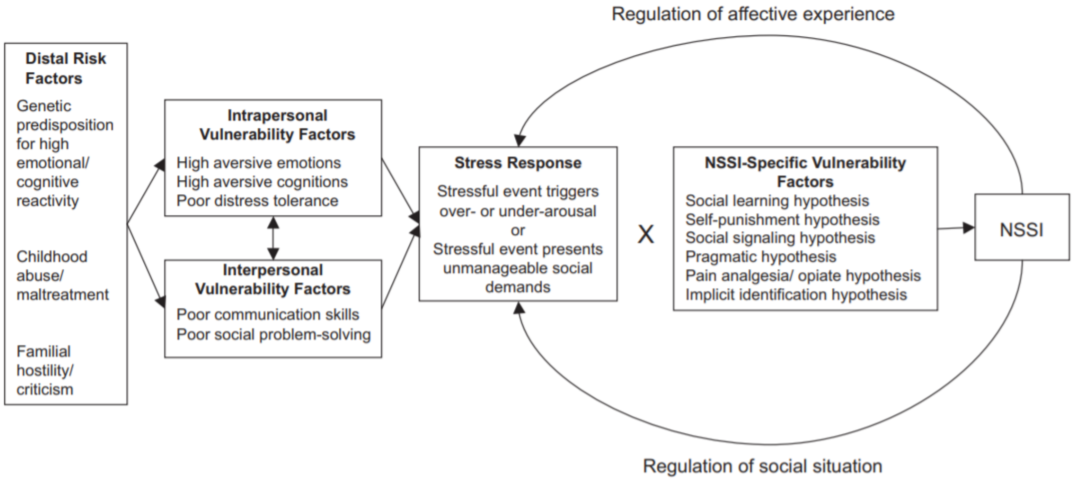

Regulation of affective experience
Distal Risk Factors
Genetic predisposition for high emotional/ cognitive reactivity
Childhood abuse/ maltreatment
Familial hostility/ criticism
Intrapersonal Vulnerability Factors
High aversive emotions
High aversive cognitions
Poor distress tolerance
Interpersonal Vulnerability Factors
Poor communication skills
Poor social problem-solving
Stress Response
Stressful event triggers over- or under-arousal
or
Stressful event presents unmanageable social demands
X
NSSI-Specific Vulnerability Factors
Social learning hypothesis
Self-punishment hypothesis
Social signaling hypothesis
Pragmatic hypothesis
Pain analgesia/ opiate hypothesis
Implicit identification hypothesis
NSSI
Regulation of social situation

Figure 2.2: Theoretical model of self-injury. Source: [2]

harm [2,52]. Over many years, many theoretical models have been improved, suggesting that self-harm is carried out to protect self-harming individuals from rage. Usually, medical descriptions commonly highlight the effect of low esteem and the desire to seek others' attention. Meanwhile, these theoretical statements require adequate empirical backing. Empirical investigations attempting to understand the relationships and risk factors associated with self-injury through examining self-harming and non-self-harming people have become atheoretical.

The study of [55] produced a list of possible factors that are related to self-injury. However, it is not clear why and how those factors lead to NSSI. An example of this is a history of domestic or child abuse resulting in self-injury [67, 68]. Nock suggests a theoretical model that incorporates various research findings and explains how they might lead to the development of self-injury [2]. The model contained three critical propositions, as shown in Figure 2.3. Firstly, self-injury is frequently carried out due to its functions-instant way of controlling an individual's affective experience and changing one's social setting.

Secondly, there is an increase in risk associated with self-injury due to factors that constitute issues regulating an individual's cognitive state and changes in one's physical environment. Those common risk factors also increase the possibility of different maladaptive behaviours such as eating disorders and drug abuse. Meanwhile, such behaviours are associated with self-injury. Thirdly, the possible self-injurious risks increases due to certain factors specific to the self-harming behaviour that influences the individual's choice of behaviour. Although self-injury is a significant health issue, the author conceptualised the motives underpinning this behaviour as *intrapersonal* such as "affect regulation" or *interpersonal* "help-seeking" purposes.

Furthermore, the investigation carried out by [69] reported reasons for young people's intentional self-harm. The researchers conducted an anonymous survey (cross-sectional) from seven European countries; England, Netherlands, Australia, Ireland, Belgium, Norway, and Hungary. They analysed data from over thirty thousand school pupils of fourteen to seventeen years. Their findings indicated that (1) 'wanted to get relief from a terrible state of mind' and (2) 'wanted to die' were the most reported self-harm motives by the participants. By using principal component analysis, the authors discovered two critical motives (*"a cry of pain"* and *"a cry for help"*) as the underlying dimensions associated with the purpose of intentional self-harm. Most self-harming participants reported engaging in self-injury due to one or more reasons associated with those motives. As opposed to males, females reported multiple reasons as to why they are self-harming. Consequently, from the examined countries, the researchers discovered consistency in the choice of self-harming motives among both males and females.

However, most young people struggling with psychiatric disorders or mental health issues access social media and participate in various online activities that could result in negative and positive effects. For example, a previous study confirmed that self-injurers utilise the internet more often than non-self-injure individuals [28]. Similar research examined the types and rates of self-injurious activities carried out by young people in an online environment [70]. The researchers believed that young self-harming individuals engaged in such activities to learn functions associated with the activities. Even though it has been confirmed that nearly 23% of young people are self-harming [64], it is believed that this group of population are the most active users of social media. Because self-harming individuals tend to be isolated and often do not speak out about their behaviour and seek help, they frequently access the internet more than people that are not self-harming [28].

2.3 Self-harm and the Internet

One example of a valuable information source that is unusually common in our life is the Internet. In January 2021, it was estimated that approximately 4.66 billion are active users that engaged on the Internet globally.[2] This figure represents about 59.5% of the worldwide population. Presently, it is almost impossible to find a nation without the Internet. At the same time, it connects billions of users around the world and facilitates information sharing in our contemporary communities. Children and young people represent a large portion of internet users globally. While this group of population are the most active users of the Internet (social media platforms [20]), it is clear that nearly 16% to 23% of them engage in DSH behaviour [15].

However, the rapid growth in Internet usage among this group and the consequent freedom of creating and sharing content could strengthen the Internet's ability to influence users' psychological health [29]. The online environment could negatively normalise intentional self-harm and facilitate access to harmful content that will likely promote suicide and NSSI [37]. In other words, the potential consequences of the Internet on self-harming users is a serious concern. The risk of intentional self-harm due to increased use of the Internet and its association with other public health issues like suicide has been explored. For example, a recent systematic review on the effect of using the Internet regarding self-harm discovered both positive and negative influences [29].

Even though the Internet contributes to offering support through social networks to people struggling with mental health problems, the online space may provide access to self-harm content and normalize the behaviour [29]. Individuals that are battling with self-harm use the Internet due to low self-esteem and isolation [29]. This makes them prone to engage in online social interactions by connecting with their peers, seeking help online, balancing emotions and lowering loneliness [71]. As opposed to offline relationships, their online social connection may create a strong friendship between them and reduce the barriers to their offline interactions, such as anxiety, low esteem,

[2] https://www.statista.com/statistics/617136/digital-population-worldwide/

and isolation [72]. Moreover, a further study showed how the broad usage of the internet is associated with a reduction in family relationships and the size of the peer group. [73].

Popular social networking platforms and online discussion boards have provided young self-harming people with opportunities to gain validation and a feeling of self-injurious behaviour. It will be difficult for parents and medical professionals to ascertain sources of possible NSSI content online. A recent study explored self-injury on Instagram [45]. Users' hidden NSSI contents with ambiguous hashtags make it hard to discover and inform or alert members of the online community about the graphic [45]. A similar study examined the use of self-injurious hashtags (over twelve months) on Instagram and demonstrated how users formulated numerous variants of self-harm hashtags to avoid being reported by other users [74]

Furthermore, a novel systematic report of internet use and self-harm in young people showed that information was exchanged on self-harm and hiding techniques, a minor feeling of isolation and building-up of good attitudes, such as looking or asking for help [29]. The researchers inferred that social media has both good and bad consequences for people struggling with self-injury.

2.3.1 Websites case studies

Although little is known about NSSI, websites presenting information about the behaviour were strongly condemned in literature and the media. The presence of websites devoted to the dangers and consequences of self-harm has drawn the attention of many researchers around the world. Although these sites provide information, most of them facilitate interactions among users through messaging and live chat. From the existing studies, there has been a significant assessment of the sites that promote communicating beliefs toward self-harm and providing methods of self-injury. It has been claimed that websites presenting self-harm information "encourages" NSSI and could increase the spread of the behaviour among users [75].

The authors explored accounts of self-harming users who access NSSI websites and conducted an in-depth interview with a set of young participants through the internet by exchanging several emails [75]. This offered the participants a room to express themselves freely and speak about the things that are hard to discuss in a face-to-face setting. Thus, the researchers suggest that self-harm websites should provide means of reaching people who could support and build a sense of community for self-harming people [75].

In other words, the three major themes, "coping with psychological distress", "feeling part of a community", and "feeling understood", found by the authors, could positively benefit users of NSSI websites. Meanwhile, there are negative implications of using these sites, such as feeling marginalised from the wider community. Also, users may be discouraged from seeking support offline (medical or professional experts). Although online mental health services cannot replace conventional offline support, the examined user's account questioned the impact of mental health support and services due to the effectiveness and accessibility of the NSSI websites.

Analysing NSSI contents over different websites assists in the evaluation of whether those sites possess unique content or could likely be dangerous to the online community. Another research evaluates and quantifies the content of NSSI on non-clinically (professional) approved websites created by non-professional individuals [76]. The investigators hypothesised that such websites contained helpful information, such as advice that self-injurers should avoid being isolated and speak to family members or clinical experts for help. Meanwhile, those websites may contain harmful content, like tips on hurting oneself or posting and sharing triggering pictures [76]. In their analysis, the researchers excluded professional websites and non-personal NSSI websites. They found several NSSI online groups, including individuals (that account for up to 18) NSSI sites, from the retrieved top 100 search results. Consequently, this produced 53 websites and 71 when added to the 18 websites mentioned earlier [76].

The researchers considered this set of websites and developed a coding scheme (1) relating to some essential variables recognised from previous studies that examined the scope of NSSI on YouTube [48] and (2) through assessing and randomly choosing sites to uncover missing variables from the previous investigation. In other words, codes were formed by reviewing the websites to find hidden variables. The study of [76] achieved an inter-rater reliability score of .976 across all variables, indicating a high agreement between the coders. It was discovered that most personal websites were developed by female adults that described NSSI experiences, disclosed cutting behaviour and highlighted NSSI to deal with negative feelings [76]. While websites often provide helpful information for self-harming people, several sites address non-self-harming users with defensive messages.

Moreover, the investigators indicated that many of the examined NSSI websites tend to share ways of NSSI concealment, such as using long sleeves to cover scars [76]. This practice could strengthen the idea of hiding NSSI behaviour and discourage self-harming people from speaking out about the behaviour and seeking help. The authors discovered that certain websites shared first aid information, and this may normalise the behaviour and encourage self-harming site visitors to self-injure cautiously and treat injuries [76]. While the findings of these researchers are similar to a study that investigated NSSI videos on YouTube [48], they show that websites presenting NSSI information are artistic and melancholic [76]. These features could glamorise NSSI among vulnerable site visitors. At the same time, they found that most of the analysed websites contained images portraying NSSI contents [76], and such images are available on social media, especially Instagram [77].

2.3.2 Special forums for self-injurers

Several online discussion forums dedicated to self-harming people and health specialists have repeatedly raised issues regarding the kind of advice provided among self-harming peers. Some researchers investigated how young self-harming individuals disclosed their concerns in an online forum and what response or advice they received from other self-harming members of the support forum [78]. Their knowledge of presenting self-harm related concerns and responses were focused on Conversation Analysis (CA) and Discourse Analysis (DA). Using these techniques, the investigators uncovered and highlighted valuable insights into the issue of self-harm help-seeking and support interactions. The data they utilised were collected from the *SharpTalk* forum, a system designed by the authors specifically for the research purpose. Although it was a secured system that acts as social media, it only operated for three months.

However, their findings indicated that participants primarily created site norms and expectations. They also highlighted the possibility of giving advice even if it was not requested. A request for emotional help is frequently accompanied by advice seeking. Intentional self-harm continued to be normalized, and users or participants combined perceptions of the behaviour as a regular activity with advice on being safe for effective recovery.

2.3.3 Social media definitions and operations

The concept behind social media started more than 30 years ago when Jim Ellis and Tom Truscott (both graduates of Duke University) designed and developed a system called *'UseNet'* that allows people to post messages online [79]. These messages often referred to as posts or articles altogether called news, are usually sent to one or more newsgroups in the system. Moreover, another system called open diary emerged in the late nineties. The open diary system was built and launched online in October 1998 by Susan and Bruce Abelson [80]. The system brings together numerous different diary writers to efficiently share information about a diary. In this manner, people who use the system can comment on a diary; this is considered the pioneered innovation of modern blogs and comments in our social media.

Furthermore, like the world-leading social media application (Facebook), the system provides functionality through which users can make their diaries private or publicly available to other system users. However, there is ambiguity in the meaning of the term 'social media for over two decades. According to [81], social media is an easy-to-use service that people can use to interact with others online. The author believed that social media is not all about the technology behind it but rather about the online social connections between people. [82] defined social media as *"the collection of websites and web-based systems that allow for mass interaction, conversation, and sharing among members of a network".*

However, looking at social media as a combination of two terms, one can interpret social as how one or two individuals meet and interact. These can be as one-to-one or one-to-many interactions. On the other hand, media can be referred to as a channel or tool through which people create, store and share information. These include the internet, magazine or news media, and television. Thus, this study adopts the definition of social media used by [82].

Nowadays, there are several social media applications that people around the world frequently access. Among these systems are Facebook, Twitter, YouTube, and others. Even though some research findings suggest that Facebook is the most popular in terms of active users, a sample of US adolescents reported that Instagram, YouTube and Snapchat are their most preferred social media platforms. A growing body of literature discovered that young people who self-harm frequently access or use the internet [25,83], and social media is one of their preferred social connection tools [29]. More recently, a study highlights the benefits of social media through which people who self-harm can access or use in obtaining support [33]. The author uncovered the attitudes and beliefs of people on Twitter towards self-harm users. The researcher demonstrates that on social media like Twitter, self-harm is not a joke and several critical pieces of information that escalate celebrities may influence the behaviour [33].

2.4 Self-harm and social media

The increase of online (social media) NSSI activities has elicited societal attention concerning their perceived negative influences on self-injured people [29,83]. Conventional media highlighted the possible dangers of online activities, referring to the Internet as a *"toxic digital world."*[3] A thorough assessment of social media activities related to NSSI may contribute to an overall understanding of NSSI contents on social media and encourage further research. Although there is a similar study in this domain conducted by [84], it appears to be a notable rise in research volume on this subject since then.

In the past couple of decades, the relationship between social networks and self-harm behaviour is a topic of interest to many researchers [25,26,77]. Social media has a mixed impact on the well-being of young people, with reports of higher self-esteem and social support and triggering contents such as exposure to graphic content [25,26,85]. Over the last decades, the rapid instances of self-injury and the allegations of social

[3] https://www.bbc.co.uk/news/uk-england-london-26828124

media's role in promoting harmful behaviour have led scholars to focus increasingly on understanding the uses and influences of social networking involving NSSI [77]. The impact of social media on people that are self-harming has been studied significantly.

In a research study that investigated the role of social networking and DSH, it was discovered that frequent use of social media is independently related to poor mental health [85]. Meanwhile, the authors found gender differences regarding self-harm thoughts and behaviours. Young females have more thoughts of self-harm than young males on Instagram [85]. Thus, significant exposure to self-harm content on Instagram correlates with increasing self-harming behaviours [85]. This is similar to recent findings claiming that passive use of Instagram could result in loneliness, depression and self-injurious behaviours [86].

Moreover, existing evidence found that the Internet and social networking are double-edged swords, offering benefits and problems to self-harming users [25]. The advantages of using online networks are evident. In addition to preventing professional help and portraying methods of self-injury behaviour, there are severe risks to online behaviour such as increasing stigmatisation, reinforcement and normalisation [25, 26, 87].

Similarly, a study confirmed that the SNS connect young people who are uncomfortable with face-to-face clinical support due to depression or suicide which are basic self-harm risk factors [88]. Nearly 75% of young individuals reported using the SNS to seek health-related information [88]. For instance, an early report on the usage of specialised sites offering online social support for people struggling with self-harm indicated exceptional levels of support coupled with normalisation of self-harm behaviours [84]. Remarkably, many investigators have found these platforms helpful because they provided the expected support needed by self-injurers [25, 77]. In contrast to that, others are worried that the SNS could promote self-harm through exposure to NSSI contents and self-harming techniques [86, 87].

Furthermore, demonstrated by a significant number of SH groups and memberships on Facebook and MySpace, from hundreds to thousands, the analysis of [89] shows that NSSI contents grow on social networking, which is similar to an earlier study [42] concerning self-harm and online discussion forums. While assessing the scope and nature of DSH content on SNS and informational websites, the authors highlight that those online spaces could be innovative for sharing reliable NSSI information [40,89]. In those platforms, the nature or state of the content moderator needs to be appraised, as this could considerably impact the reliability and quality of the NSSI online information.

Meanwhile, the researchers' analysis unveiled that online groups for self-injurers operating on social media are usually supportive and educational [89]. They recognise freedom of expression and attempt to increase awareness [89]. While audiences worldwide can easily access the groups, they lack trigger warning messages and are often moderated by non-clinical experts [89]. Usually, NSSI groups reach out to people on Facebook and are open and accessible to general members of the online community. The findings of [89] show that those groups allow users to post multimedia or non-multimedia information concerning harmful behaviours. For instance, while MySpace showed a more comprehensive level of content moderation, NSSI images could still be accessible in the platform [89].

Nevertheless, the authors' work were broadly consistent with prior research that examined NSSI videos on YouTube [48]. Videos commonly included young adult women communicating neutral or anti-NSSI information, including hopelessness, discouragement and encouragement, and several overlapping elements related to different mental sicknesses [48]. Furthermore, the visual depiction of self-injury on videos with character and non-character was observed by the researchers [48]. While non-character videos showed a wide degree of self-injury, one of the five examined character videos presented high graphic content, such as attempting to end life [48]. This was the only video without a trigger warning message. Considering the graphic nature of such videos and their representation of NSSI, it is evident that they have become highly concerned [48].

2.4.1 Descriptions of the included studies

This section reports a detailed description of the relevant studies considered in this research. Related works were reviewed, critiqued and analysed. The descriptions of the existing studies were summarised based on (1) the reported positive and negative influences of social media, (2) the approaches used by similar studies, and (3) the social media platforms they investigated.

Positive views

In the examined studies, the benefits of building online social connections, thereby reducing self-isolation, were considerably discovered [87,90]. One of the common notable positive effects of social media on self-injurers is the decrease in social isolation [25]. Many authors believed that by engaging with their peers online and sharing personal information regarding NSSI experiences, the self-harming users could reduce social isolation and gain positive well-being [26,87].

The analysis of [91] is one of the primary investigations showing the effects of exposure to self-injurious content on Instagram. The researchers carried out two-wave panel research among young individuals between 18 to 29 years. In addition to other hypotheses, they examined a general hypothesis that exposure to NSSI contents on Instagram could be positively associated with the behaviour itself and suicide-related consequences [91]. In other words, they worked to assess if the current suicide and self-injury risks factors are associated with exposure to NSSI contents on Instagram and confirmed if the exposure could possibly expand or sustain the likely outcomes. Their findings indicated that among the young people (N = 313, 43%) they examined who were exposed to NSSI contents on Instagram, only 20.1% showed to have deliberately sought this content [91].

Therefore, unintentional exposure to NSSI contents on Instagram was the fundamental link for many subjects exposed to the behaviour. Consequently, the vast majority of users encountered Instagram's NSSI contents accidentally [91]. Although Instagram provides a feature to report users sharing harmful content that violates their terms of use, many of the platform's users are not aware of the feature [92]. Meanwhile, the researchers found many users (around 59%) that experienced emotional difficulty due to exposure to NSSI photos [91]. Only 33% of the examined users were not emotionally disturbed by NSSI graphic contents. Consequently, the authors found no correlation between the NSSI behaviour and changes in emotion [91].

In contrast to non-copycat users that reached up to 66%, nearly 32% of the users who took part in their research indicated that exposure to NSSI contents significantly influenced them to imitate others [91], and this has been reported by similar studies on Instagram [87, 92]. Thus, demonstrating that a substantial proportion of the participants did not perform similar self-injurious behaviour depicted in the NSSI images [91]. Their findings discovered that self-harming individuals who viewed images of NSSI on Instagram tend to demonstrate NSSI related consequences [91]. Their study found two critical themes: helpful and not-helpful images. Looking at both themes, they discovered sub-themes that denote reactions of people commenting on NSSI images [91].

The vast majority of the examined users believed that images were helpful as they reported that such content helps to increase support among previously or presently self-harming individuals [91]. This is similar to a recent study on Instagram that reported online activities of individuals attending clinical support due to mental health issues [93]. Additional research examined NSSI groups operating on Facebook and reported similar findings that found group members promoting NSSI prevention and advising self-injurers to seek clinical support [94].

Some members were clean from self-injury and remained active in the group to help others and suggest ways of recovery [91]. Moreover, many users described the examined pictures as evidence that some people were struggling with self-harm behaviour. Meanwhile, several users noted that images depicting severe injuries due to NSSI illustrates how severe some people self-injured [91], and sharing those images online could allow users to compare images of their injuries and compete among themselves [86].

Participants of a study conducted by [87] who engaged in self-harming reported posting images of their injury on Instagram. Wanting to raise awareness, imitate and help others were the common reasons the participants posted those images. The researchers found the participant's reactions to viewing self-harm content on Instagram as triggering, desensitisation and inspiration to stop harming themselves [87].

Diverse self-harm contents are shared online (mostly on social media) and the platforms provide space for young individuals to seek validation [90]. Concerning self-harm, social media creates contagion [86]. Those platforms allow users to compare images of their injuries and compete among themselves [86]. Users tend to post contents that

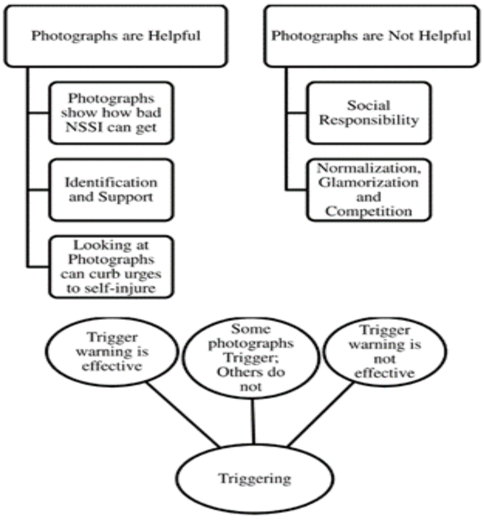

Photographs are Helpful
Photographs show how bad NSSI can get
Identification and Support
Looking at Photographs can curb urges to self-injure
Photographs are Not Helpful
Social Responsibility
Normalization, Glamorization and Competition
Trigger warning is effective
Some photographs Trigger; Others do not
Trigger warning is not effective
Triggering

Figure 2.3: Self-harm and suicide risk factors, Source: [3]

could be helpful – promote recovery or harmful – encourage self-harm [37, 70]. People's reactions to online self-injurious photos were found to be positive or negative [87].

The work of [95] thematically analysed users' reactions to NSSI images online. The researchers utilised open data from a database containing people's opinions/comments on NSSI photos. These images portrayed mild to severe wounds resulting from self-injurious behaviour [95]. Their investigations explored themes connected with reactions to NSSI images posted by the members of the online community. It is essential to recognise the likely influences of such photographs on users who viewed them because (1) there is speculation that NSSI graphic contents could trigger NSSI urges and (2) normalised and reinforced online [30, 87, 95].

Recent evidence revealed that online (social media) activities strength to decrease self-harm behaviour through promoting recovery [77]. For example, most participants (nearly 73%) studied by [42] reported their association in a self-harm discussion board as helpful or beneficial. Online members of the NSSI communities can boldly support their peers' decision to stop self-injury, and this has been reiterated by recent studies [70, 85, 90]. Individual stories of self-harm recovery were observed to suggest support for those self-harming users and proof that total recovery is feasible [42].

Negative views

Research evidence reports that young people tend to seek health information from online social platforms, and numerous of the information found on those platforms were inaccurate, dangerous and misleading [96]. There is a serious concern regarding the negative impact of social media use, especially on young individuals' mental health [97]. Such platforms pose risks such as desensitisation and normalisation to self-injurers [98]. Although contents about self-harming behaviours are banned on Instagram, such contents are still visible on the platform due to the inability or lack of an effective intelligent detection mechanism [96].

Researchers recognised the impacts of using the Internet or social media to view and discuss information associated with NSSI [25]. There is an argument among researchers regarding social networking sites' role in people who are harming themselves [25, 77]. The use of SNS among self-injurers has been recognised with various positive influences, such as community formation and peer support [25,26,77]. Meanwhile, some researchers believed that people suffering from NSSI behaviour could inappropriately use the SNS, which would increase the risk of potential NSSI [25, 26, 77].

Young individuals using social media and receiving clinical support concerning self-harm and suicide were studied recently [99]. Results of the surveyed study indicated that those using social media had a significant rate of self-harm as opposed to those who denied using social media. Similarly, as opposed to those that are not using social media, participants using social media reported a high risk of suicide, which is a significant factor that is associated with self-harm [99].

In another study conducted by [37], it is clear that people who participate in NSSI discussions online tend to share their personal stories. The researchers found that viewers' comments on NSSI videos on YouTube primarily disclose their self-harm experiences. In some cases, the authors believed that these experiences, including the video's content information, could maintain NSSI behaviour. However, their findings support the claim that NSSI videos on YouTube social media could maintain self-harming behaviour [48].

Furthermore, another investigation examined four online groups for self-injurers that are socially active on Facebook [94]. The precise manner in which the NSSI group could be dangerous to its followers is through exposure to triggering content [94]. While the online platforms provide a sense of anonymity among self-harming users when disclosing their behaviour and personal experiences [100], it is clear that some users take this advantage and abuse, harass or bully other people online [94]. Moreover, several posts generated by online users contained triggering and graphic content [94]. Although the researchers were not expecting such content, they added a category to their coding scheme and captured the triggering content [94]. Because some group members are vulnerable, such dangerous content could be critical or negatively impact their mental health. Even though Facebook moderators will remove any content encouraging NSSI, the authors found posts presenting self-injurious behaviour and suicide techniques [94].

Another investigation found a negative impact of online resources (websites) for self-harming people [76]. The authors confirmed that certain websites often present techniques related to NSSI concealment [76]. For example, they were advising self-harming individuals to cover their body scars with long sleeves. While this is a way of encouraging self-injurers to hide their behaviour, it can also discourage them from speaking out and looking for support. Furthermore, the investigators found a number of websites sharing first-aid related information [76]. Site visitors that are self-harming could learn to harm themselves cautiously without seeking clinical help for serious injuries. In other words, the potential dangers associated with visiting such websites include normalising self-harm and exposure to new ways of harming oneself [76]. Recent research investigations restated these issues [85, 92]

Moreover, the findings of [49] indicated some possible negative consequences of accessing NSSI first aid videos. While some of the video presenters highlighted *"safe"* guidance for self-injury, the researchers acknowledged that self-harming viewers might not stop injuring themselves [49]. While only one of the examined videos promotes help-seeking from experts, the investigators believed that viewers might be discouraged from stopping NSSI and looking for medical treatments from health specialists [49]. Suppose that the viewers could mirror themselves to the information presented in the examined videos. In that case, the count of views and number of likes associated with the videos could increase NSSI reinforcement [49]. Because most of the analysed NSSI first aid videos did not recommend seeking medical support to treat injuries, the authors believed this could increase the risk of infections and serious injuries [49].

Furthermore, while Twitter and YouTube remain among the leading social media platforms, there is a concern about how people use these platforms to promote a blue whale challenge [101]. Recent investigations discovered how people portray the challenge and the likely factors that could put individuals at risk [101]. The researchers found critical themes surrounding the challenge and how the examined videos from YouTube violated the Suicide Prevention Resource Center (SPRC) safety and protection guidelines [101]. The authors claimed that the analysed posts could have a problematic impact and may significantly normalise the challenge through exposure to self-harm and NSSI reinforcement [101]. Several studies examining NSSI contents on social media such as [37, 92, 95, 98] and [47] reported the risks of NSSI reinforcement.

Positive and negative views

Although some studies showed that self-harming people tend to look for support online [102], it is evident that the online environment reduces their isolation and lower NSSI urges [26,90,98]. Individuals who engage in self-harm behaviour benefit from peer support (such as recovery encouragement) while exchanging or sharing emotional disclosure [103,104]. On the other hand, such individuals are likely to encounter triggering content and other related critical online activities that could normalise or influence self-injury maintenance [26,98]. However, while some research studies demonstrate the positive and negative influences of social media concerning intentional self-harm [85,87,92], other studies remained neutral about the impact of these platforms in promoting or preventing the behaviour [40].

For example, The experiment conducted by [95] on online NSSI images revealed a dichotomy outcome. The researchers indicated that the participants reported positive views and declared that those images decrease *loneliness* and *self-injury enactment.* Meanwhile, other individuals argued that those photographs *promote self-injury* and *reinforce* the behaviour [98]. While the work of [94] showed that NSSI groups on Facebook facilitate and promote help-seeking from experts and encourage peer support, the researchers revealed that some of the group messages carried verbal harassment against self-harming members. Again, some of the posts they examined contained graphic content and were also triggering [94].

In another study conducted by [105], the researchers thematically examined responses of self-injurious people to understand the motives encouraging them to communicate or participate in online NSSI communications. Most of the participants indicated positive motives like *helping others*, and getting *NSSI support* were the underpinning reasons they engaged in NSSI online communications [105]. More recently, another line of research investigated the motives of posting and sharing self-harm content on Instagram [70]. While there are numerous reasons underpinning posting self-harm content on social media platforms, the authors discovered that participants share harmful content on Instagram because they want to; "(1) feel a sense of belonging, (2) speak to other people who feel the same way as themselves, and (3) offer help and support to other people" [70].

Meanwhile, from the study of [105], few participants reported negative interactions, such as viewing graphic content that is triggering. However, various investigations conducted by researchers from different fields on the role of social media concerning DSH were based on different approaches [25, 26, 106]. The following section reviewed and analysed different techniques used by the examined studies.

2.4.2 Existing approaches and analysis

Research in computer and information sciences largely depends on quantitative, qualitative, or mixed methods, including various techniques and theories from different fields. Recently, the global health crisis due to the coronavirus (COVID-19) has changed many ways researchers conduct investigations in this field [107]. For example, in qualitative research, COVID-19 restrictions created a critical challenge for researchers to recruit participants for a face-to-face interview. Despite all the challenges, researchers in this field continue making significant contributions using those methods to achieve aims

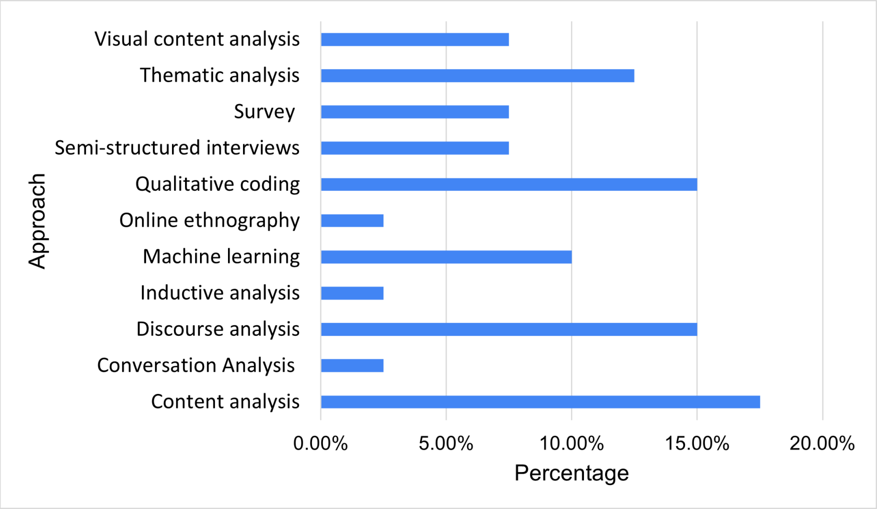


Figure 2.4: Research approaches used by the included studies.

and objectives. However, this thesis reviewed the included studies and examined various research approaches. The idea was to understand how researchers have used these methods, assess the predominant approaches, and suggest a unique technique that could be applied to investigate NSSI content on social media.

Figure 2.4 presents a horizontal bar graph that represents the percentage of different techniques adopted from the examined studies. It is evident that most of the analysed studies (nearly 20%) used a descriptive content analysis approach. Research studies that applied qualitative coding and thematic analysis methods account for up to 14% and 12%. On the other hand, visual content analysis, surveys, and semi-structured interviews were utilised at almost the same rate. The remaining methods were applied in a few studies and only 10% of the examined studies employed machine-learning techniques.

However, in the sub-sections below, our study discussed the various techniques used by the current studies in our research community. This is to understand the widely accepted techniques and standard best practices for addressing self-harm and social media research issues.

Qualitative related studies

This section report the qualitative techniques used by the included studies. The idea was to understand the most prevalent methods and how they have been used or applied to different self-harm and social media research.

Thematic analysis (TA): some of the included studies thematically assessed the impact of social media on people that are self-harming. The work of [95] thematically assessed online pictures depicting NSSI. The authors used thematic analysis to preserve the detailed description of the dataset. This approach is a qualitative technique that is sensitive to people's differences that could be overlooked while using other methods such as content analysis. Even though thematic analysis improves theory-driven review of several phenomena, the investigators employed the approach for three reasons. Firstly, they believed that the inadequate knowledge of NSSI photo descriptions needs an exploratory procedure. Secondly, the method is fine-tuned to combine the nature of the dataset into more general theoretical information. Thirdly, the proof implicated in their investigation may be diversified, for example, people may be for or against NSSI images.

Like the approach adopted in their previous work [48], the authors produced two coding rubrics. The first rubric explored the general nature of responses extracted from the videos discussing NSSI on YouTube. In assessing the comments coding guidelines, the researchers (individually) reviewed the comments thoroughly and improved the rubric. On the other hand, the second coding rubric centred on the most occurring responses ascertained from the first analysis. However, the researchers applied the coding rubrics in the analysis and found themes surrounding responses to NSSI videos on YouTube [37]. The outcome of their findings showed several comments that implied self-disclosure. Around 38.39% of the commentators shared their personal NSSI experiences. Viewers' feedback to the user who uploaded the video and appreciation of video quality received up to 21.95%. While the messages shared accounted for 17.01% and admiration for the source uploader reached 15.40%, about 11.15% of the audiences' responses commended and encouraged the video uploader. Assessing self-disclosure responses for recovery-oriented information confirmed that 42.89% of the comments did not suggest recovery and revealed that they were still self-harming 34.00%.

On the other hand, some investigators [36] analysed NSSI posts from the Instagram, Twitter, and Tumblr social media platforms. The researchers thematically examined posts tagged as self-harm irrespective of whether the pictures clearly show self-injury or the text surrounding the analysed NSSI image describes self-harm. However, their findings revealed four themes: *(1) communicating distress, (2) addiction and recovery, gender and the female body, (3) identity and belonging (4).* Furthermore, their study confirmed that none of the examined photos expressly supported self-harm or suicide, and no image could be interpreted as sensationalising NSSI.

The study of [33] analysed public data concerning self-harm on Twitter social networks using a qualitative inductive thematic method. Typical narratives generated by the tweets were analysed using inductive thematic analysis (ITA) [108]. Within the textual tweets, user-added images are presented. The author evaluated those images similar to the text to understand the context and information the image communicated to the target viewers. However, the researcher sought to understand how SH behaviour is formed on the Twitter social network. The author's findings revealed the following themes using the Inductive Thematic Analysis (ITA) method on a collection of 362 tweets.

- Self-harm is not a joke
- Celebrity influence
- Support for and from others
- Eating disorders and self-harm
- Self-harm videos and personal stories

In order to uncover the hidden trends in publicly released YouTube videos, video responses, and tweets from Twitter that specifically mention the blue whale challenge, the study of [101] employed a TA approach and coded 60 videos uploaded on YouTube, including 1112 video responses and one 150 tweets from Twitter that are mainly related to blue whale challenge. To assess the safety of content messages, the authors used a metric provided by the Suicide Prevention Resource Center (SPRC) and deductively coded the examined videos.

Furthermore, the study of [109] performed a detailed thematic analysis focused on exploring the impacts of online self-harm pictures. The authors' inductive analysis of textual data was conducted several times, and open coding was also applied in the process. A variety of coding techniques were subsequently used for the entire corpus of the data. The authors collected the data and analysed it iteratively. The themes that emerged during the process were used to construct and improve the coding scheme [109].

Content analysis (CA): the review of this dissertation found similar studies using visual and textual content analysis approaches. The investigation performed by [36] used visual content analysis and analysed NSSI pictures from the Instagram, Twitter, and Tumblr social media platforms. While the authors considered sampled pictures classified as self-harm including their associated text comments, they discovered that more than half of the pictures marked as self-harm did not show self-harm explicitly. In cases where there was graphic representation, self-injury was the most prevalent among the depicted images of self-harm. Another similar research [32] assessed 1,293 pictures of self-injury on Flickr using video content analysis including distinct visual approaches to identify and analyse the visual structure of self-injurious pictures. The author explored specific attributes associated with the images such as the focused distance, camera position and point of view.

Moreover, Instagram is one of the social networking platforms that attract many teenage users around the world. Due to the increasing nature of unstructured data generated through social media applications like Instagram, some contents could be vague and ambiguous. Consequently, hard to understand not only the self-harm research community and other critical stakeholders like systems designers that can develop a new innovative way to prevent self-harm contents. By using *'#selfharmmm'* as a hashtag [45], retrieved (in twelve days) Instagram's posts or contents associated with the tag. In addition to the structured evaluation and content analysis approach, the investigators used triangulation methods and analyzed a sample of two hundred and twenty-five posts [45]. The goal of their study was to assess ambiguous terms associated with NSSI on Instagram.

While the authors observed the content's meaning and consistency, they aimed at investigating NSSI content's warning labels on the platform. Their findings revealed ten ambiguous hashtags on Instagram and discovered a trending NSSI image labelled as *"#Mysecretfamily"*. This hashtag is linked to several hashtags representing and describing the wider group's mental health issues and NSSI communications. For example, the researchers believed that #cat and #deb are hashtags associated with the *"#Mysecretfamily"* hashtag to communicate or share contents related to NSSI and depression. Although the researchers found no guarantee in the content's advisory warnings, only a small portion (one-third) of the examined hashtags produced contents with advice warning labels.

Similarly, a comparative analysis of Tumblr, Twitter, and Instagram was carried out by [40]. In six months, the authors retrieved DSH contents from the three social networking platforms using the *"#cutting"* hashtag. The researchers' work extracted a total of seven hundred and seventy posts with seventy-eight from Twitter, three hundred and thirty-three from Tumblr and three hundred and fifty-nine from Instagram. The researchers used a content analysis approach and coded posts using a pre-defined list describing the study's codes and definitions. Intuitively, the authors aimed at investigating and comparing NSSI contents labelled with the *"#cutting"* hashtag across the three social networks. However, five different codes emerged from their investigations: graphic content, negative self-evaluations, references to mental health, discouragement of deliberate self-injury, and recovery-oriented resources. From the entire sample, the researchers found 60% portraying self-harm graphic content.

Furthermore, the study of [98] utilised *#ritzen* Hashtag, a German hashtag representing *#cutting* and retrieved images that directly depicted wounds on Instagram. Initially, this produced 1135 images linked with the 30 most common German NSSI hashtags. Secondly, the authors extracted images, including user accounts connected with the identified hashtags. In addition, the researchers adopted a quantitative approach and independently assessed NSSI images, associated comments, and user accounts for content analysis. In other words, they defined instructions (with examples) and guidelines to support the coding of the sample NSSI images into different categories.

Moreover, upon completing the first round of coding by two psychologists (undergraduate students), the inter-rater reliability score of their agreement was low. This lead to the second or final coding from two professionals working in the field of NSSI. While considering posting behaviour based on weekly and daily trends, the authors assessed the relationships among image characteristics and comments. Furthermore, they analyzed and classified a total of 8,154 comments connected with the NSSI images. Content of those comments was allocated to one of the following categories: (1) complimenting on the wound, (2) empathetic reaction, (3) offering help, (4) warning/asking the user to stop the behaviour, (5) abuse, and (6) discussions that not referring to the user.

Facilitating the evaluation of NSSI online-related activities while working with young self-harming individuals who have earlier or are presently self-harming is crucial. The study of [89] used *self-injury* and *self-harm* keywords and searched informational/interactive websites, social networking, and video sharing (YouTube) websites. For example, the study [49] searched YouTube to discover videos presenting NSSI first aid procedures available on the video-sharing site. The authors retrieved NSSI first aid videos using different phrases like; self-injury wound and self-injury first aid. The purpose of their study was to determine the nature and extent of NSSI first aid information on YouTube by conducting a content analysis of 40 NSSI first aid videos.

In order to examine the type of interaction and frequency of themes surrounding SH discussion on Facebook, a multi-valued conceptual CA approach was adopted by [94] and assessed NSSI groups' interactions on Facebook in three months time period. The most common topics found by the researchers are replies to verbal abuse of the self-harming users (16.8%), discussion of personal problems without support seeking (11.2%) and suggestions to seek support directly from other group users (11.0%).

In another research [23], Tumblr was sought for online content relating to depression, self-harm, and suicide. The authors investigated Tumblr posts associated with selfmutilation and cutting keywords, as many young individuals experiencing self-harm-related behaviours also suffer from depression. To better understand the search terms, the researchers looked at popular postings from each of those terms on Tumblr.

While the researchers considered user accounts that had their postings on the top five search results, they chose five user accounts for each specific search term. Thus, the investigators found a substantial proportion of depression-related content on Tumblr posts showing self-harming behaviours. Considering the prevalence of depressive and harmful content on Tumblr that could lead to suicide, the authors suggested the need for suicide prevention efforts to act on Tumblr strategically.

A research study conducted by [9] used a synchronic corpus-linguistic and examined NSSI posts on Tumblr [110]. In examining the NSSI textual content, the researchers applied a text analysis tool called Linguistic Inquiry and Word Count (LIWC). Using the LIWC software, the authors considered the following linguistic and affective variables presented in Table 2.4. In addition to those, they also used the tool in assessing another set of NSSI distinct categories such as NSSI methods, terms related explicitly to cutting and NSSI terminologies, and NSSI motives.

Table 2.4: LIWC Linguistic and affective categories used by source [9]

Linguistic	**Affective**
first-person singular pronouns third-person singular pronouns first-person plural pronouns third-person plural pronouns	emotional tone, negative emotion anxiety, anger, and sadness.

However, the researchers demonstrated that the predominant emotional tone (negative emotion) of the examined NSSI posts provided insight into individuals or users experiencing mental illness [9].

Semi-structured interviews (SSI): recently, the investigation of [87] conducted a qualitative analysis of young self-injurers' motivations for sharing pictures of their NSSI injuries on Instagram. At the same time, the researchers aimed to get a deeper understanding of self-injurers' reactions to viewing those images and determine how they react to comments on their individual NSSI photos. The authors conducted their study using a semi-structured interview and recruited participants from a more comprehensive dataset analysing the prevalence of NSSI on Instagram [98]. The investigators downloaded (hourly) all images and user profiles connected with the sixteen German hashtags most often associated with images of NSSI wounds (such as #ritzen, "#cutting").

Consequently, following the four weeks of their data extraction on Instagram, they randomly chose a total of hundred individuals from the dataset, contacted them through Instagram messaging, and invited them to participate in the research. Additionally, participants were questioned whether the wounds or scars depicted in the posted photographs were the results of their personal NSSI. Participants were included in the research if they agreed. A semi-structured interview (using Instagram messaging) was conducted in which interviewees were asked individually about their experiences of NSSI [98].

A sample of 21 young individuals aged 16 to 24 in Wales who had previously experienced self-harm was interviewed using semi-structural interviews by [109]. While Facebook was used to recruit the participants, the researchers aimed to gather the opinion of different online members that experienced DSH, including those who are not actively interacting with the existing online social networks for self-harmers. The authors set up and deployed adverts on Facebook pages supporting online members experiencing DSH. The researchers aimed to reach users who liked pages promoting the positive well-being and mental health of people in Wales, particularly charities and adolescent groups [109].

Therefore, the study interviewed participants at different locations such as the cafe and university library [109]. The topic of the interview guide and examined the participant's self-harm experiences, self-harm motives, obtaining professional and unprofessional support, utilising the Internet, and operating the Internet before, during, and following self-harm [109]. The investigators used digital audio recording equipment to record the data, and a professional transcriptionist performed the data transcription. The researchers employed a constant comparative method to study the similarities and differences between different user accounts [109]. Similarly, the initial interview adopted in their investigation identified the critical knowledge gaps concerning different internet use levels.

Nevertheless, the examined participants reported the position of the Internet in normalising DSH, and recent investigations have reported the same issue [86, 87, 98]. The most considerable purpose of using the Internet regarding self-harm is the photographs as opposed to textual interactions [109]. Pictures draw out physical reactions from individuals and encourage behavioural representation. This is especially true with Tumblr, which allows self-harming people to post images without revealing their identity.

Discourse analysis (DA): of the investigated studies, only a few utilises the discourse analysis method. Similar to a study conducted by [78], the work of [46] examined interactions between self-injurers and SharpTalk forum moderators. The authors aimed to assess the nature of problem expression and answers or advice-giving in the online forum. Even though the researchers utilised *facework's* framework to understand social interactions, the overall objective of their study was to examine the online interactions among young self-harming individuals, moderators and health professionals in a SharpTalk forum. While participants exhibited endearment, encouragement and unity, the study shows several mitigation devices are employed and imply that young people in their supporting contacts are oriented to a *'protective'* line.

In the same vein, [100] concentrates on what defines membership in an online forum (SharpTalk) for self-harming individuals. The researchers used a discursive method to comprehend and understand how forum members utilise the forum and interact with others. In their study, the authors applied the discursive method, including membership categorisation analysis and conversational analysis, to comprehend forum users' activities. Notwithstanding, there is a considerable study on how individuals use phone lines or face-to-face meetings with mental health practitioners and other health needs using conversational and discourse analysis has been done. In line with this, a similar study applied conversational analysis methods to understand forum users and moderators' interactions in terms of problems presented concerning self-harm and the answers provided on the SharpTalk forum [78].

Moreover, the work of [32] examined self-injurious images on Flickr using discourse analysis and described those photographs as *"photographic images capturing intentionally injured and/or scarred bodies or body parts."* Even though the author retrieved the sample pictures using the keyword *"self-injury,"*, he coded and categorised Images according to essential self-portraits techniques. However, his analysis found some image uploaders raising awareness about self-injury and supporting online users using the *"To Write Love On Her Arms"* (TWLOHA) movement campaign.

Online ethnography (OE): the review investigations discovered that only a few studies adopted ethnographic methods. A more recent investigation [111] adopted an online ethnography supported by a qualitative scheme for interpretation in medical anthropology [112], highlighting its social, cultural and structural elements. This conventional approach applies to textual or graphic data in online interactions [112, 113]. The method presented a suitable means of exploring the communications of self-harm as contextual in social media users' offline lives and social settings.

Quantitative related studies

Survey studies: recent research hypothesised that SH exposure on Instagram could be connected to the behaviour itself and suicide-related consequences [91]. The researchers examined the relationships connecting exposure to NSSI contents on Instagram with (1) self-injurious behaviour, (2) suicide ideation, (3) specific suicide plan, (4) suicide risk, (5) suicide predictors such as hopelessness, and (6) the purpose of living. Intuitively, they centred on this broad category of NSSI outcomes to decrease the possibility of omitting essential relationships. However, in their analysis, they conducted a two-wave survey study of young adults 18 to 29 years old residing in the United

States [91]. Instagram exposure to SH content was evaluated in the first wave, and all constructs related to SH and suicide were surveyed in both waves.

The authors focused on cross-sectional connections between Instagram exposure to SH content and many outcomes (such as measures of wave one). Also, they evaluated panels that examined changes in Instagram exposure to SH content and outcomes from wave one to wave two measures. They found that exposure to SH contents on Instagram was connected with the behaviour of itself and suicidal ideation. They discovered that disclosing such contents was also associated with emotional distress. Their investigations showed that many of those exposed to self-harm on Instagram encountered such contents by accident. Consequently, their findings show that studied users who saw self-harm on Instagram during their lifetime managed to show higher NSSI related issues. More importantly, their study confirms that young individuals who get intentionally or unintentionally exposed to self-harm on social media like Instagram are at greater risk of self-injurious behaviour.

Another study [92] surveyed young undergraduates and adult Instagram users between the age of 18 to 49 years. Around four hundred and seventeen people participated in the study, and 60% were white and females represent 77%. Even though only 3% of surveyed individuals were using the platform for less than a year at that time, many of them (nearly 88%) indicated that they have been using Instagram for more than three years. While Instagram banned some critical trending hashtags promoting self-harm, their findings confirmed that a small number of participants - 13.5% heard of *#selfharm, #selfharmm* or *#selfharmmm.* The researchers collected responses from participants regarding Instagram use, self-harm awareness, theoretical measures (knowledge, attitudes, subjective norm, behavioural control, behavioural intentions) and demographic information to understand the level of awareness and intent of using Instagram's self-harm reporting feature.

Nevertheless, even six months following the tool's implementation, their findings imply that the reporting tool could not reach people successfully. Moreover, they observed that around 20% of the surveyed users were informed that the tool is available. Although regular use of Instagram was not associated with the intent of using the reporting tool, the researchers discovered that all the variables of the theory of planned behaviour were significantly correlated to using the self-harm reporting tool.

Recent experimental pilot research examined the effect of peer messages communicating hopelessness and hope thoughts regarding NSSI recovery [30]. The investigators hypothesized that exposure to negative comments would increase hopeless views of NSSI recovery while sharing and presenting hopeful messages would increase positive perceptions. Their research designed fictional peer responses and randomly allocated participants to either helpless or positive recovery-oriented responses placed in a screenshot of videos discussing NSSI on YouTube. An online survey was used to assess participants' attitudes concerning NSSI recovery. While advertisements for the study were distributed on online social platforms dedicated to NSSI, such as subreddits with the consent of the platform's moderators, individuals with a history of NSSI were recruited from those NSSI online spaces.

Machine learning (ML): some of the reviewed existing studies utilised machine learning techniques. This approach involves using either supervised learning, unsupervised learning, or both methods in building an automatic model to detect SH content on social media [114, 115]. For instance, a recent study presented the initial automatic image-recognition algorithm that automatically detects images depicting cutting on Instagram [116]. The authors incorporated two computational methods (1) web scraping and (2) computer vision. The first technique automatically retrieved NSSI contents shared on Instagram using specific hashtags representing various languages - *#cutting*, #suicide, *#ritzen* (cutting), and *#selbstmord* (suicide) [116].

Firstly, The researchers created an algorithm training database that includes choosing images from Instagram relating to NSSI. Secondly, they utilised a total of 600 images related to cutting on Instagram. While downloading the pictures using *4K Stogram* software, the investigators manually picked photographs depicting new injuries or bruises [116]. In the training phase of the classifier, the authors considered images that did not depict or show self-injury ("negative" images) information. Thirdly, the researchers developed and trained an automatic image-recognition algorithm based on a convolutional neural network classifier to automatically detect images representing cutting-related contents [116].

However, in the initial cycle of the training phase, their algorithm was able to predict three classes of pictures; “NSSI-cutting”, “NSSI “, “NSSI-cutting”, and “no NSSI” [116]. The first two classes implied NSSI-related contents, and the “NSSI-cutting” class included only pictures representing only cutting. Moreover, the first cycle considered 1,384 pictures, while the second cycle of their experiment used 1,200 images. While the training was improved, their algorithm got an error of *19% and *13% during the first and second training cycles and achieved *81% and *87% classification accuracy, respectively [116]. Both training cycles produced false-positive/false-negative rates of .24/.08 and .23/.07 [116]. The accuracy of the first training cycle converged almost to one constant after 100 periods of training. Notably, the investigators believed that a better level of performance could be achieved in the second training cycle due to fewer predicted classes [116].

To better understand self-harm posts and develop automated methods for detecting such content, the experiment conducted by [31] utilised data from Flickr. As owned by Yahoo Inc., the platform is one of the most prominent websites hosting pictures. However, the researchers combed through a considerable number of public Flickr posts and chose those that included the keywords *"self-harm"* and *"self-injury"* [31]. While the authors conducted an in-depth examination of self-harm content on Flickr by utilising various features, their investigation indicates that those features (linguistic, owner, temporal, and visual characteristics) of self-harm content are distinct from normal content [31]. Consequently, they utilised these prominent features and built a framework that automatically detects (with high accuracy) self-harm content from normal content using supervised and unsupervised machine learning methods [31].

Another similar research has recently proposed a Bag of Sub-Emotions (BoSE), a technique that uses emotions to detect Reddit users struggling with self-harm [117]. By utilising users' posts, the authors extended the BoSE to understand users' emotional changes. However, when compared with other methods, their approach demonstrated good performance results [117].

Due to the significant correlation between depression and self-harm, the investigation by [50] aimed to create a new dataset (the Reddit Self-reported Depression Diagnosis (RSDD)) to identify forum users with self-reported depression diagnoses. In order to construct the RSDD, the authors utilised data publicly accessible on Reddit and annotated users [50]. Nevertheless, to assess self-harm content and classify vulnerable users at risk of potential harm, the researchers retrieved data from ReachOut.com social platform, a site specifically for people struggling with mental issues like self-harm [50].

To perform text classification across many input texts, the investigators designed generic neural network architecture, and proposed two models that use this architecture to accomplish two tasks: (1) self-harm risk classification and (2) identifying depression [50]. Therefore, assessing risk for self-harm, also known as self-harm risk classification, requires measuring the current level of user's risks using the person's post on a mental health support forum and the comments and replies they have posted in the thread [50].

Even though both tasks have a similar goal of predicting the mental health state of a user, they differ in two significant ways. Firstly, the experiment being conducted by the researchers is different, as the prediction is being made using a four-point scale versus Boolean (binary classification) [50]. Their study employed a *two-step* procedure in developing the general models: (1) finding essential features in every training data input, and (2) utilise the discovered features to improve the model classify users [50].

Meanwhile, in another related study, some authors designed a content crawler and retrieved (using NSSI hashtags) self-injurious contents such as texts, images, and videos on Instagram [118]. By utilising the crawler, the researchers identified many NSSI contents posted on Instagram between January 2016 to December 2018. Even though Instagram users have developed many variants of self-harm hashtags to avoid being reported by other users [74], the authors derived the commonly hash-tagged words that are linked to NSSI, grouped them into six different classes, and analysed the data to understand its trends [118]. Nevertheless, NSSI images are difficult to classify because of the certain constraints which includes (1) capturing images posted by users and (2) the time and costs of obtaining vast amounts of annotated data [118].

Meanwhile, only image labels are available. Typically, they include various objects at distinct scales dominating self-injurious content [118]. In order to address these constraints, the author's approach relied on techniques for weakly supervised object localisation and applied the Adversarial Complementary Learning (ACoL) technique [119]. The findings show that the amount of NSSI-related posts increased drastically, indicating that potential NSSI content is posted on Instagram at an increasing rate [118]. To classify NSSI photos, the researchers built a binary image classifier using a deep learning technique. Furthermore, even though their technique and results are crucial for various reasons, their study introduced a unique method of classifying NSSI content showing the risk of potential DSH [118].

2.4.3 Platforms, fields and published dates of the reviewed studies

Several studies have demonstrated the role of social media concerning NSSI and how it impacts our mental health [26, 87]. Researchers have mixed views on the effects of social media on people experiencing self-injurious behaviour [40, 77, 106]. However, looking at Figure 2.5, it is observed that many of the examined studies focused on the Instagram social media platform. Among the reasons why researchers considered Instagram is that the online social platform promotes ease of data sharing with academic researchers [120]. Undoubtedly, this allows many scholars to assess the platform's NSSI contents to support decision-making among social media teams, clinical professionals and other stakeholders [77]. Furthermore, researchers pay attention to Instagram due to the growing concern about exposure to the graphic content of NSSI on social media [77]. The platform is among the leading SNS that promotes the sharing of user-generated content, especially images.[45]

Although there are fewer studies on Flickr, SharpTalk and Reddit, our analysis indicates that contents of special discussion forums for self-harming people were less investigated. In addition to Instagram, the second and third most common examined social media platforms were YouTube and Twitter. Twitter has recently increased the

[4] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[5] https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

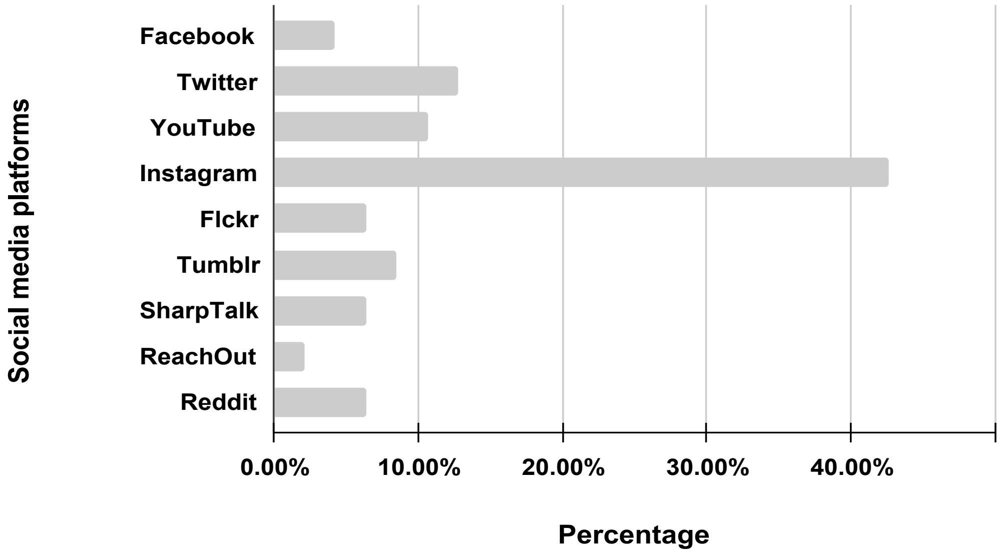


Figure 2.5: Investigated social media platforms

limits of tweet character size to 280,[6] and most of the relevant research analysed Twitter posts using the initial size limit of 140 characters [33]. This signifies that further study is needed to investigate NSSI posts on Twitter. Meanwhile, YouTube social media facilitate viewer's interactions through comments. While some of the existing studies analysed NSSI video content [37, 49], more research is needed to understand (in large scale data) what the commentators are discussing regarding the NSSI videos [30, 37]. Because the rate user generated contents on social media is on the increase, self-injurious content could be hidden. Hence, it is essential to develop effective intelligent detection techniques that could automatically discover harmful content on social media platforms [96].

However, as discussed above, researchers (from different fields) participated in conducting various investigations concerning self-harm and the impacts of social media in promoting or preventing DSH behaviour. Accordingly, the review conducted in this study explored the existing researcher's area of specialisation as displayed in Figure 2.6. Most of the reviewed studies (around 38%) were authored by researchers from psychology, such as [30, 49]. Meanwhile, a few studies were completed by educationists

[6] https://developer.twitter.com/en/docs/counting-characters

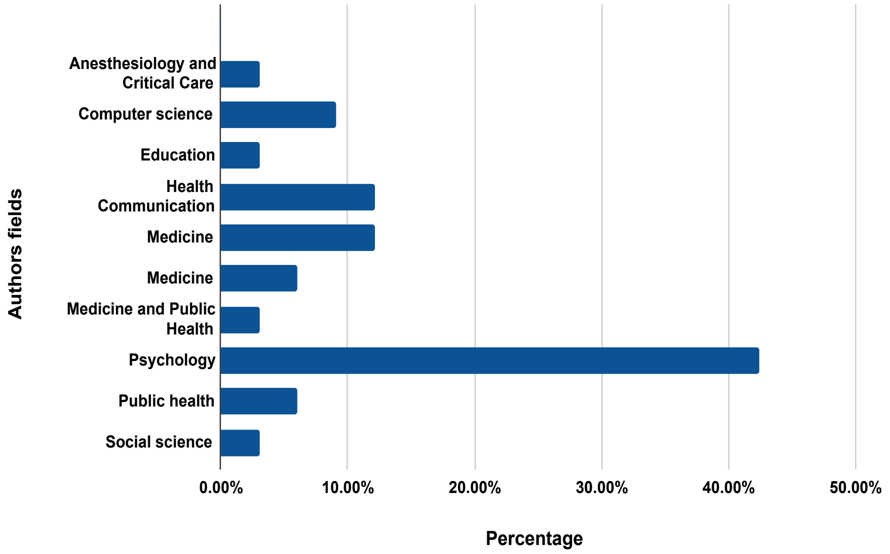


Figure 2.6: Researcher's field from the examined studies

and social scientists [9, 109]. While computer scientists and researchers from health communication account for up to 11%, professionals from medicine were the second dominant group of experts [45].

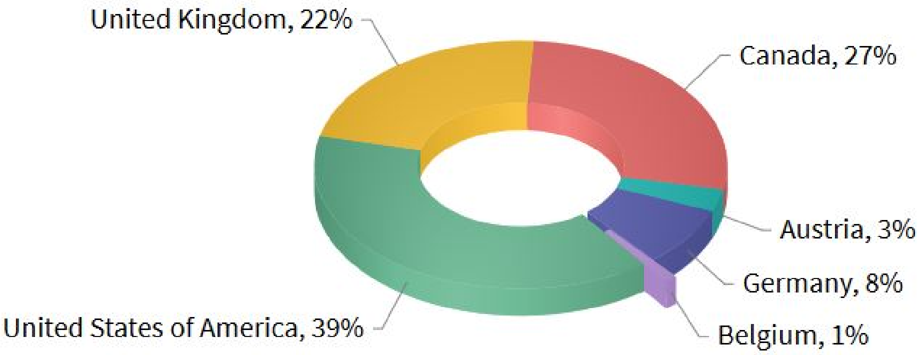


Figure 2.7: Locations of the examined studies

Furthermore, in our review analysis, this research studied the places or countries where the examined studies have been conducted. The pie chart in Figure 2.7 illustrates the percentage of locations that conducted the reviewed studies. While the United

States of America is the leading country with 39%, it was discovered that Canada and the United Kingdom are the second and third leading countries with about 27% and 22%. Similarly, the review analysis found a few studies from Austria, Belgium and Germany.

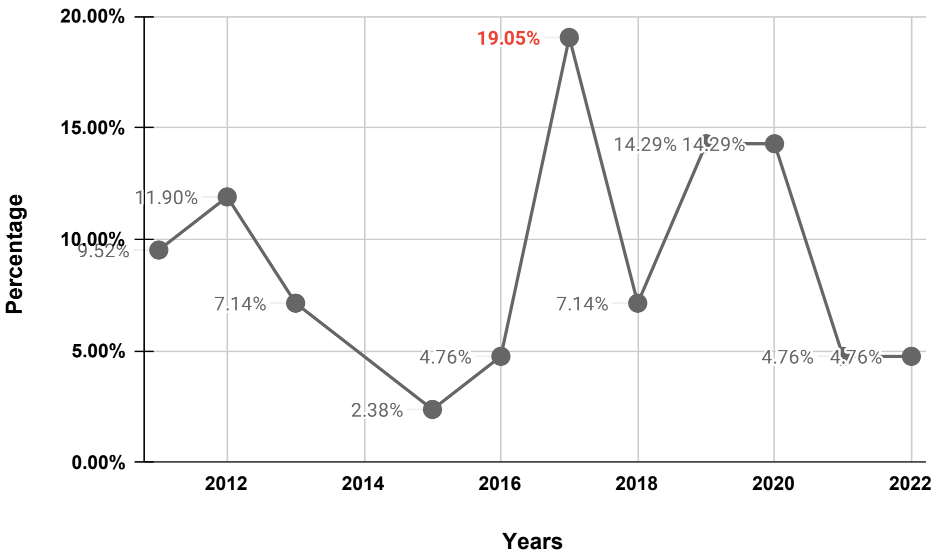


Figure 2.8: Years of the reviewed studies

This study examined the published dates of the included research papers. Because the research community is constantly changing, it is essential to know the most recent findings. The spike from Figure 2.8 indicates that most of the studies were published in 2017 and subsequently in 2019 and 2020. Although few studies were found in 2015, it was discovered that an equal rate of relevant articles was published in recent years (2021 and 2022).

2.4.4 Strengths and weaknesses of the existing studies

Our scoping review uncovered the key strengths and weaknesses of the included studies. Many studies reported the impacts of social media concerning self-harm and what clinical professionals need to know about the role of social media in self-harm prevention [90, 97]. For example, the work of [37] demonstrates that responses to NSSI videos on YouTube could be helpful or harmful to the target audiences. Viewers interact and exchange peer support and suggest ways of recovery to stop people from self-harming. Even though their research revealed that some of the examined content could be harmful, it is not clear if such content could be easily identified from a large amount of unstructured data generated by YouTube. More effort is needed to detect commentators needing urgent support to prevent future self-harm automatically.

As seen from Figure 2.5, it is evident that most of the investigated studies focused on Instagram as opposed to other social networking platforms. The study conducted by [116] introduced the first image-recognition model that automatically detects photographs with or without NSSI from pictures on Instagram. Moreover, the study examined the prevalence of pictures depicting or not portraying NSSI from Instagram's images posted using NSSI multi-language related hashtags. Meanwhile, there is a need to improve the NSSI image detection algorithm using large scale datasets encompassing diverse NSSI images.

The investigations carried out by [40] offered a systemic evaluation of NSSI content publicly accessible on three major social media platforms (Twitter, Instagram and Tumblr) over a long period. Their research adds to the growing collection of empirical investigations alarming the nature of DSH content on social media [37, 45, 48, 76]. Using the content analysis approach, the authors uncovered posts containing negative self-evaluations and graphic information. However, despite their significant research contributions, there are a number of drawbacks associated with their investigations:

- The hashtag (#cutting) they used in retrieving NSSI contents and the random sampling approach they considered could not be generalised to NSSI content on social media platforms. For instance, our recent study discovered that, in some cases, the #cutting hashtag returned irrelevant tweets discussing hairstyle cutting instead of NSSI posts.

- Due to ethical issues and data protection policies, the authors analysed public data. Therefore, their findings could not be generalised as it does not include private data.

- The researchers did not assess the effect of user's interactions with the examined contents such as comments, likes and shares.

Therefore, it is not clear whether the online community accepts or rejects the analysed posts. Additional effort is needed to characterise users who create and share NSSI content on SNS and assess the impacts and mental health consequences of engaging with the content. Moreover, the study conducted by [109] improved on the current empirical evidence on the role of social media on people experiencing self-harm with a focus on NSSI images. While participants reported the impact of the Internet in normalising self-harm, the authors discovered that the analysed images were the primary purpose interviewees reported for using the Internet for DSH purposes. Meanwhile, there is a drawback in the sampling procedure considered by [109]. The participants examined by the investigators may not be a representative cohort of self-harming people who are using the SNS, as the study was limited to only participants with a Facebook account.

Moreover, the findings of [45] uncovered self-harm ambiguous hashtags (such as *#Blithe*, *#cat*, and *#selfinjuryy* on Instagram. While these tags are popular on Instagram, the authors discovered that contents associated with the hashtags have no advisory warning information for the respective audiences. Due to the popularity of the *#selfharmmm* hashtag, the researchers limited their analysis based on the hashtag. However, their investigation does not represent the entire NSSI content publicly available on Instagram or other social media platforms.

Furthermore, the work of [32] adds to the existing knowledge on the impacts of NSSI pictures on self-injure users. The researcher indicated that the dominant discourses of self-injurious photos on *Flickr* greatly determine the shape of the pictures [32]. The author's work confirmed that NSSI photograph uploaders expressed their injury and scars in the form of narrative resilience [32]. However, the study of [32] is accompanied by some shortcomings. Firstly, investigation focused on user-generated content (self-injurious images) instead of the people that uploaded the images. Therefore, this could not ascertain why self-injurers post their NSSI images online and their experiences from sharing those content. Secondly, the approach adopted by the researcher could be challenging to apply in a considerable amount of NSSI photos.

Similarly, another research that examined NSSI images on Instagram using a computational approach made a substantial contribution by proposing an algorithm that can automatically detect the presence or absence of self-injury from digital photographs [116]. While the algorithm was applied to examine the prevalence of self-injurious pictures on Instagram, the study increases awareness of the possible impacts of self-injury photos on the platform [116]. Among the drawbacks of the study is that it require more training using large samples containing all the possible patterns of digital images representing NSSI. Additionally, the data collection process was done in reversed direction. Therefore, the proposed algorithm may have removed or missed some aspects of the NSSI content.

In a related investigation, it was observed that people are intentionally or unintentionally exposed to NSSI images on Instagram [91]. Exposure to NSSI pictures evoked emotional difficulties among online members and is statistically correlated to self-harm related consequences [91], and this has been further emphasised by recent studies [85, 86, 92]. Notwithstanding the significant benefits of their research investigations, it is clear that the researchers' data was part of a big project on the effect of DSH on new media [91]. Thus, the larger research project limits its study to recruiting only Netflix participants. While they restricted the examined Netflix users to only 18 years and above, it is essential to acknowledge that their study excludes younger participants who may be self-harming and affected by the harmful effects of NSSI exposure on Instagram.

While crawling and retrieving self-harm related images using NSSI-related hashtags, the research investigation of [118] confirmed the prevalence rate of NSSI photographs on Instagram and proposed a classifier that detects NSSI photos that achieved up to 94% performance accuracy. Despite the significant contributions of the study done by [118], the author's proposed classifier could be enhanced by training and testing on a large amount of data to increase performance accuracy.

2.4.5 Key areas of improvement

The review conducted in this study identified several research gaps that require further investigation. Despite the vital interest of many researchers in addressing various challenges surrounding this area, there is a growing concern about how social media platforms influence people with mental health issues like self-harm as the behaviour is related to suicide [22, 77]. Although most studies confirmed that self-injurers have no intention to end their lives, it was evident that some social media content could trigger vulnerable users and increase the risk of self-harm [26, 85, 92].

Although several areas of improvement were discovered from the existing studies [21, 26, 77], this section reports the critical research gaps addressed by this dissertation. The investigations of [91] demonstrate the evidence that young self-injurers who intentionally or unintentionally viewed NSSI contents are at high risk of expressing the negative impacts of those contents. Their study reveals that NSSI content on social media (Instagram) is critical and needs further investigation. Similarly, another study suggested that a video-search system like YouTube could be an online space for people to present NSSI-related information to a wide range of viewers [30, 37]. YouTube could expose triggering videos to vulnerable young viewers whose responses or interactions require more investigation [89].

However, the study of [37] did not clearly show how the NSSI videos could have short and long-term effects on viewers. Examining NSSI video responses signifies a unique way of evaluating the effect of the videos. The author's coding scheme (coding rubric) can not be applicable to large-scale data representing comments of NSSI videos as it will be difficult and time-consuming. Even though the coding rubric is reliable, it only reflects the investigator's study perspective [37]. People communicating DSH on YouTube may or may not agree with some of the themes discovered by the researchers [37]. Thus, not all viewers of NSSI videos may comment on NSSI videos on YouTube. Extra effort is needed to investigate and understand what commentators are communicating as there is insufficient knowledge of what self-harmers are discussing online and why they participate in the discussions [32, 36].

Another study distinguished between self-harm and non-self-harm content on Flickr social media [31]. Using the features from those contents, the authors proposed a framework that automatically detects self-harm content using supervised and unsupervised machine learning techniques [31]. Meanwhile, the study of [31] was limited to identifying self-harm content rather than actual users at risk of self-harm. More effort is required to study online users and identify vulnerable at risk of possible self-harm. While the novel work of [33] was the first to explore publicly available real-world data concerning self-harm on Twitter, the author's contribution increases our understanding of Twitter's role as a source of online support for self-harming users. The researcher [33] qualitatively studied the attitudes and beliefs of online members towards individuals experiencing self-harm. Hence, the approach used by the investigator could not be suitable for a large scale of Twitter posts (tweets) relating to self-harm.

On the other hand, Twitter serves as a source of support for self-injurers [33]. Understanding the behaviour among members of the community and the influence of information they received could be contributing factors that could prevent or normalise the behaviour. Therefore, more effort is needed to investigate Twitter's sources of help for people with mental problems like self-harm. Understanding the impact of those sources and their strategies on Twitter could broaden our understanding of the influences of online social support in self-harm prevention and recovery.

2.5 Summary

This chapter reviewed the existing research regarding the role of social media in supporting or preventing self-harm. The evidence from the examined studies presents social media's positive and negative effects on DSH. Although the review discussed the basic concepts of NSSI and social media, the literature suggests the role and differences between the diverse online sources (such as special forums and websites) of help for self-injurers. The review aimed to assess the current evidence about the possible impact of social media on self-harm.

Due to the changes in social networking sites and their increasing nature, research studies will always lag behind fundamental improvements. Social media could positively or negatively influence vulnerable young users. For example, evidence from the existing studies demonstrates the differences between offline and online behaviours among self-harming individuals. People that are self-harming tend to be isolated and lack the courage or motivation to speak about their behaviours in face-to-face interaction. Meanwhile, online resources, mainly social media, have become a convenient place for those people to disclose their information, express their opinion, share experiences, and exchange support.

Additionally, the stigma and stereotypes associated with intentional self-harm can make online social platforms like YouTube and Twitter exciting places for self-injurers to interact anonymously. The anonymous nature of these platforms could provide self-harming people with higher freedom to express opinions that are hard to describe [15]. In contrast to physical interactions, online social media platforms increase social interactions among users, especially between those with common interests [121]. Hence, self-harming adolescents are more likely to engage in online activities than non-self-harming people [28].

Meanwhile, this doctoral study discovered some research gaps that need to be filled. While existing research confirmed that YouTube is one of the sources of information for self-injurers, there is a critical concern regarding the responses to videos presenting self-harm information. Even though the sources of those videos are unknown, discussions surrounding the videos and the viewer's opinion about the videos are crucial to understanding the audience's role in preventing or promoting self-harm.

Furthermore, existing research confirmed that individuals who recovered from self-harming find it challenging to control the urges to injure during recovery due to strong addiction associated with the behaviour. Therefore, social media have become a space for them to share their experiences and seek support. While self-injurers exchange or share content that could be helpful or harmful, this doctoral study considers the effect of exposure to such content a double-edged sword issue. It allows the investigation of such content through data analysis (including machine learning techniques) and detection of the individuals impacted by the content. Therefore, this study was interested in identifying such content, detecting various groups of users who engage in self-harm, and mitigating non-self-harming users that could be influenced.

However, the scoping review presents the existing approaches used by different researchers and their fields of study. In light of this, the literature increases our knowledge and understanding of the relevant approaches to fill the research gaps mentioned above. Thus, in the context of this research, social media analysis is crucial as it could help to reveal hidden information about self-harm behaviour from unstructured data. While using natural language processing, the analysis will add to the existing knowledge about the nature of self-harm discussions and users' sentiments towards the behaviour.

Following the review of the current studies in this area, several research approaches were used by many researchers - see Figure 2.4.2. In addition to other approaches like online ethnography, a small proportion of the examined studies use ML techniques. Due to the effectiveness of the ML approaches in classifications and predictions, it is necessary to utilise them in understanding and detecting harmful content in online spaces to improve safety and reduces the risk of self-harming behaviours.

Chapter 3

Methods

3.1 Introduction

As seen in the previous chapter, this study conducted a scoping review of the available research on the influence of self-harm content on social media. Thus, some of the existing gaps found from the review were (1) a lack of self-harm detection mechanisms, (2) insufficient knowledge of self-harm discussions on digital social platforms, (3) little understanding of user's sentiments towards self-harming individuals and (4) limited or no knowledge of the strategy of organisations providing social support for self-harming people. Following the diagram in Figure 3.1, the present chapter discusses how this study approached these challenges and addresses the research questions explained in Section 1.3.

Meanwhile, the research questions and objectives informed the study's design process. This research employed a mixed-method approach due to (1) the nature of the research questions under investigation and (2) the data used in the study. The following section discusses the method's concept and the rationale for its use. The design of this study and the data collection (including the analysis) were also discussed in Sub-section 3.1 and 3.4.

3.2 Research methodology

There are three generally acceptable research methods; qualitative, quantitative, and mixed methods. Words, images and videos are utilised to denote qualitative investigation. This method allows researchers to gain in-depth insight into unrecognised or poorly understood research issues [122]. Basic qualitative research procedures involve interviews, recording observations in plain language, and analysing existing research to examine theories [122].

On the other hand, using numbers or figures characterises quantitative research. While it is used to verify or disprove theories and hypotheses, this method can be utilised to build and generalise facts regarding a particular area of research [122]. Experiments, and questionnaires or surveys, are a few instances of using a quantitative approach in a research study. A quantitative study systematically gathers and analyses data from various sources that utilise statistical or computational techniques.

Meanwhile, mixed-methods research is popularly used to describe a study incorporating qualitative and quantitative data into a particular study. A mixed-methods study can incorporate qualitative and quantitative features throughout the investigation process, from the initial stage (data collection) to the analysis and interpretation stages. However, as seen from the literature, research studies employing machine learning techniques are primarily quantitative. They entail data modelling and statistical procedures to gain insight or knowledge of the data.

Consequently, in any research investigation, the researcher's methodological approach informs the characteristics of the data used in the research and vice versa [122]. For instance, quantitative research mainly reports the attributes of the data as numeric. Similarly, in social research using content analysis, the investigators could interview people or ask them to complete surveys. Hence, the methodological approach (quantitative or qualitative) used by the researchers could guide the type of participants required for the study due to time and cost constraints. Accordingly, considering the research questions of any study, the data gathered from the study through surveys or interviews could guide the selection and use of the appropriate methodological process.

Thus, this research uses the mixed-methods approach.

3.2.1 Why mixed-method study?

Mixed methods analysis started in social sciences and recently extended to the field of computing [123]. Its processes have been established and improved over the few decades to address several research issues [124]. The approach has been widely used to address design issues and improve Human-Computer Interactions [125]. For instance, individuals working with a system could be invited to complete a survey study or recruited to participate in an interview to enhance or build a new interface or system.

While considering the background of our study and the research questions under investigation, mixed-method is the suitable approach to address the research questions. The reasons for choosing the mixed method are twofold. Firstly, the literature review conducted in this study informed choosing both quantitative and qualitative approaches in our investigations. This is because, from the related works, it has been observed that these methods are very effective in analysing self-harm content from social media. For instance, the work of [33] qualitatively explored tweets concerning self-harm on Twitter and a similar study utilised coding rubrics on YouTube video comments [37]. Similarly, another related research used a quantitative method to understand self-harm content on Flickr.

Secondly, this research extracted data from YouTube and Twitter social media platforms. The data retrieved from these platforms is extensive, unstructured, and has different properties. In the case of YouTube, videos about self-harm were extracted, including their comments. Similarly, this study retrieved tweets concerning self-harm from Twitter. Although using purely quantitative or qualitative research on these data could be achievable, both techniques are practical to address the research questions. Hence, the reason for the use of the mixed method in this study.

3.3 Research design

The previous chapter demonstrates that several existing studies sourced and analysed data from various social media platforms. However, this research was initiated by searching relevant online databases (see Table 2.2) to identify related studies. The idea was to review and understand the literature to identify critical areas of improvement and form research questions. As shown in Figure 3.1, the research gaps identified from the existing studies are labelled *P1* to *P4* to represent the research problems approach in this study. Accordingly, this study used different techniques to address the research questions.

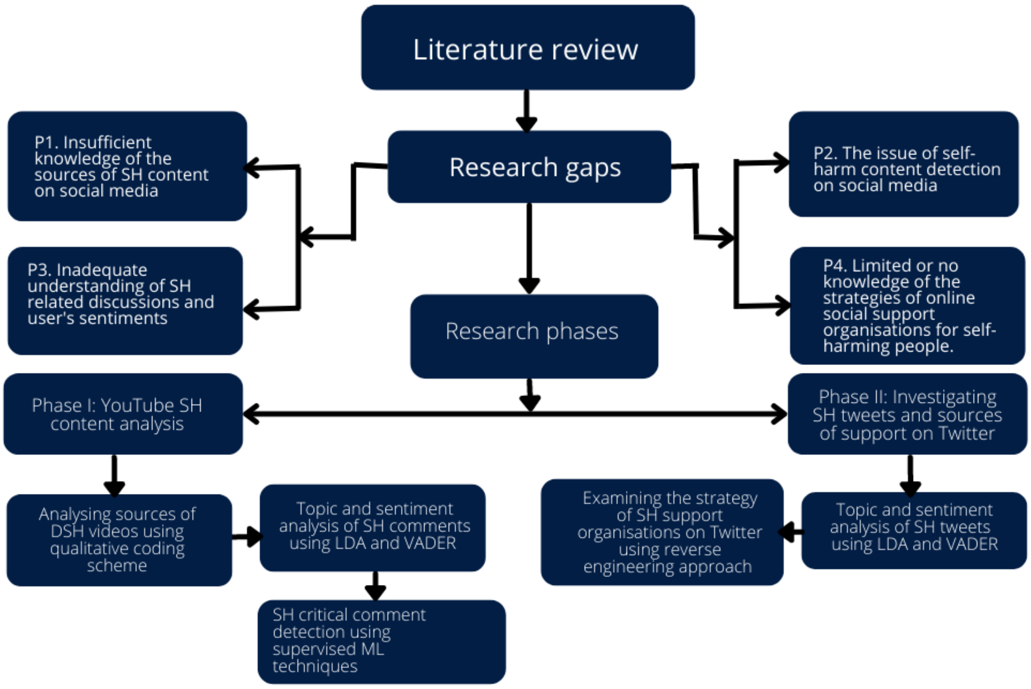

Literature review
Research gaps
P1. Insufficient knowledge of the sources of SH content on social media
P3. Inadequate understanding of SH related discussions and user's sentiments
P2. The issue of self-harm content detection on social media
P4. Limited or no knowledge of the strategies of online social support organisations for self-harming people.
Research phases
Phase I: YouTube SH content analysis
Phase II: Investigating SH tweets and sources of support on Twitter
Analysing sources of DSH videos using qualitative coding scheme
Topic and sentiment analysis of SH comments using LDA and VADER
SH critical comment detection using supervised ML techniques
Examining the strategy of SH support organisations on Twitter using reverse engineering approach
Topic and sentiment analysis of SH tweets using LDA and VADER

Figure 3.1: Research methods

Regarding the first research question, *Who uploads videos on YouTube concerning self-harm, and how are the videos rated?* Our study developed a coding scheme based on the nature of the data and qualitatively analysed the various sources of the videos.

Also, our study quantitatively analysed how viewers rated the videos. Similarly, our study approaches the second research question; *(What are the discussions surrounding YouTube videos presenting information about self-harm?)* using the Latent Dirichlet Allocation (LDA) topic modelling algorithm explained in Section 3.7.2. This algorithm is an unsupervised technique that allows the automatic detection of *"topics"* that appear in a document's collections [126], hence the purpose of its use in this research.

Moreover, this study uses the Valence Aware Sentiment Reasoner (VADER - discussed in Section 3.7.3) to address the third research question; *What sentiments do users express concerning videos about self-harm on YouTube?* due to its effectiveness in sentiment analysis of social media text [127]. Furthermore, this study uses the supervised ML techniques described in Section 3.5.1 to answer the fourth research question; *How can we detect critical comments from DSH videos on YouTube?*

Again, using the LDA topic modelling technique, this research study addressed the fifth research question; *What are people discussing regarding intentional self-harm on Twitter?* Meanwhile, this study uses a Reverse Engineering (RE) technique to answer the sixth research question - *What strategies support organisations operating on Twitter to facilitate and encourage self-harm recovery?* Similar to question two, our research analysed the seventh research question *(How do followers of the examined organisations opinionate on the information shared by the support accounts over time?)* using VADER.

Therefore, the design of this study is based on (1) the nature of the background issues and research questions, (2) the study's objectives and (3) the social media platforms under investigation. Hence, the doctoral research was designed and performed in two phases. The first phase study focused on YouTube to address research questions one to four, while the second phase focused on Twitter to answer research questions five to seven.

However, both platforms can not be used to answer all the study's research questions. This is because both platforms have different functionalities. Unlike Twitter, YouTube provides a search system for users to retrieve videos of various categories using query terms or phrases. The platform allows people to upload videos to their online channels. Videos posted by YouTube users could have comments of lengthy characters to allow viewers to share their views or though regarding the video. The findings from the YouTube study informed the second phase study. Some sources of videos about self-harm on YouTube (charity organisations) encourage users to follow them on Twitter. Thus, exploring the strategy of those organisations on Twitter in reaching out to self-harming users is crucial.

Additionally, people can post tweets (text, images or short videos) on Twitter. Unlike YouTube, the platform does not support long text tweets due to character limitations - not more than 280 characters. Hence, it is essential to analyse tweets and understand users' opinions. Furthermore, the content ratings of both platforms are different. YouTube provides anonymous ratings (*likes* and *dislikes*) for video content, while Twitter provide only a tweet *liking* feature. Consequently, in addition to the evidence confirming these platforms as ideal online spaces for self-harmers, the differences between both platforms suggest the need to study them differently.

3.4 Data collections and analysis

3.4.1 Application programming interface (API)

Application programming interface, commonly known as the API, can be defined as the instructions that facilitate access to web-based tools and software applications [128]. Software organisations may create their APIs and share them (securely) with private and public organisations, including individuals, to allow them to develop applications through their services. In other words, there is no need for developers to understand the implementation of an API; they utilise the interface to interact with the services. The use of API proliferated over the last decade, to the point that most popular social media platforms could not be a valuable data source for researchers without the APIs [128]. In the same vein, YouTube and Twitter have different API services that allow developers and organisations to access their services securely.

3.4.2 YouTube data API:

Google offers a plethora of APIs, and YouTube data API is one of them. Each of the APIs has a specific application in different areas.[1] This simplifies the process of developing web and mobile applications. YouTube is among the leading video-sharing social media platform on the internet.[2] While collecting data can be highly beneficial, it is essential to identify the most prevalent video channels and track their popularity, rates of likes and dislikes, and the number of views for each video.

This implies that YouTube enables developers to obtain rich information like video statistics and data about the video channels or uploaders. For the purpose of this dissertation, a Google developer API was created, and access keys to search for videos discussing self-harm on YouTube were obtained. As discussed in the YouTube data collection section 3.4.2, this study used five search terms and retrieved videos concerning deliberate self-harm.

Phase I: YouTube data extraction and analysis

Sourcing data from YouTube: YouTube is one of the social media platforms that facilitate the formation of an online community discussing self-harm. While the popular video search system provides quick access to videos related to intentional self-injury, there is a growing concern about the critical impact of self-harm content on social media. YouTube offer various interactive features that promote interactions (via comments) among viewers and video uploaders. However, this phase outlined the *step-by-step* procedure for retrieving the research data from the YouTube platform. In other words, the flowchart in Figure 3.2 illustrates the procedure used in collecting the data from YouTube.

Many researchers from different fields sourced data from YouTube and examined different issues. For instance, some researchers examined young individuals' practice of creating content on YouTube using content analysis [129]. Previous research concerning self-injurious behaviour and YouTube used keywords such as *"self-injury"* and *"self-*

[1] https://developers.google.com/youtube/v3/getting-started

[2] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

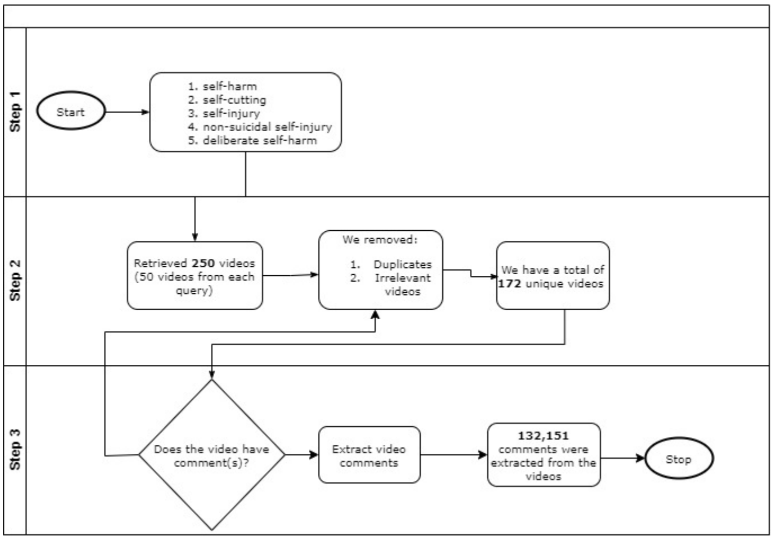

Step 1
Step 2
Step 3
Start
1. self-harm
2. self-cutting
3. self-injury
4. non-suicidal self-injury
5. deliberate self-harm
Retrieved **250** videos (50 videos from each query)
We removed:
1. Duplicates
2. Irrelevant videos
We have a total of **172** unique videos
Does the video have comment(s)?
Extract video comments
132,151 comments were extracted from the videos
Stop

Figure 3.2: YouTube data extraction flowchart

harm" in collecting data from the platform [37, 49]. This study extracted a set of videos about deliberate self-harm on YouTube. In other words, our research used five query terms *(1) self-harm, (2) self-injury, (3) cutting, (4) deliberate self-harm,* and *(5) non-suicidal self-injury* and retrieved videos discussing self-harm on YouTube.The purpose of using many search terms was to retrieve relevant content for the analysis. Meanwhile, the API developer account used for this study is limited to only fifty videos per search term. Therefore, a python script containing the API credentials was created and used to extract two hundred fifty videos *(n=250)* using the above key phrases.

This small number was extracted due to YouTube's API, which allows only fifty videos per query.[3] Meanwhile, duplicate and irrelevant videos were removed, such as those demonstrating haircut styles. Thus, resulting in 176 unique videos. Therefore, our goal was to examine the responses of those videos and answer research questions mentioned in Section 1.3 of Chapter 1. Unlike tweets on Twitter, video comments on YouTube have no character length. This allows users to express their opinion and share their experiences sufficiently.

YouTube data analysis: this section explained the step-by-step procedure followed in analysing the data. The idea was to understand the data and utilise the essential

[3] https://developers.google.com/youtube/v3/docs/search/list

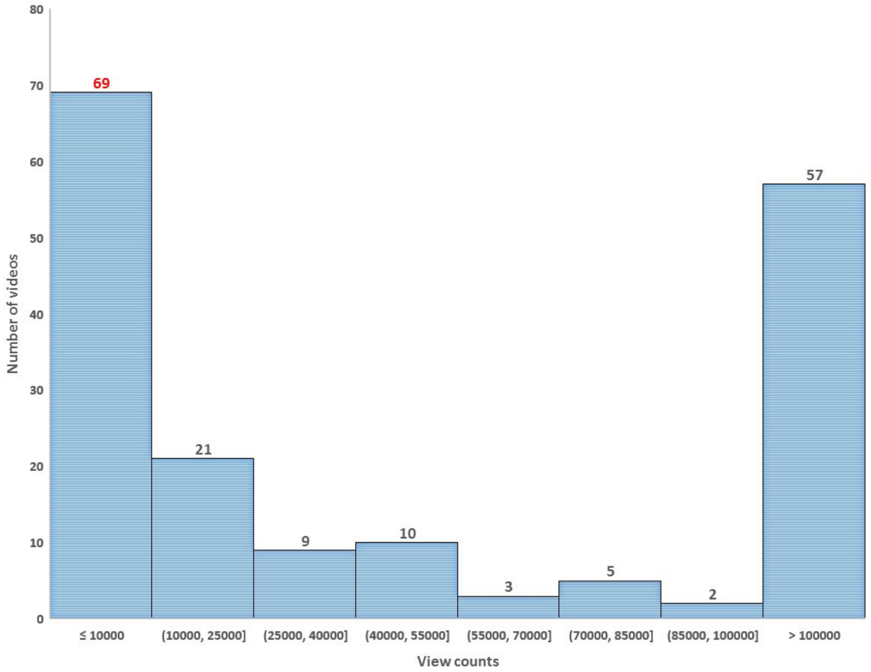


Figure 3.3: Distribution of video views - Most of the videos were viewed at most ten thousand times, and a considerable fraction received at least a hundred thousand views.

information to answer the research questions. Thus, the analysis takes into account YouTube's interactive features. The videos retrieved from YouTube were uploaded between 2010 to 2020 and received more than 18,807,869 views, see Figure 3.3. While the channels that uploaded the videos have 19,200,703 subscribers, the videos received 328,239 likes and 12,629 dislikes. These numbers indicated that many people had seen videos concerning self-harm on YouTube and participated in self-harm discussions through comments. Each video featured a presenter or a group of presenters. Since the literature shows a gender imbalance among people who are self-harming [15, 63], this dissertation analysed the data to understand the gender proportion of the presenters. Similar to the literature findings, the preliminary analysis shows that 66.67% of the video presenter were female as opposed to males with only 24.56%. Hence, males need to increase participation in reaching out to the audiences about self-harm on YouTube.

However, YouTube has a feature that allows users to assign a *category* to a video while uploading. YouTubers can choose a category that describes the type of video content. The category feature facilitates the video search and result ranking process.

For example, a video tutorial about learning python programming can be assigned in the education category. Our analysis examined the different categories assigned to the examined videos. Consequently, this dissertation found the most popular categories of assigned to the analysed videos by the uploaders. The goal was to bring these attributes to the attention of computer and information science researchers, particularly information retrieval system experts, to expand the ranking of relevant, helpful videos to the intended audiences to promote self-harm awareness and prevention. This analysis discovered the popular channel categories uploading videos about intentional self-harm. Among the top categories are people and blogs, education, and music.

Furthermore, YouTube incorporates features that allow viewers to rate video content. The *likes* and *dislikes* functionalities of the YouTube social media provide essential information to understand viewers' ratings. These features are crucial in improving the search and result ranking process. Accordingly, analysing the extracted data from YouTube required essential stages to fully understand and identify which aspects of the data would aid in answering the research questions and achieving aims and objectives. YouTube comments are written texts of audiences that viewed video content. The properties of our data include the commentators' names, the date and time the comment was posted, likes and dislikes associated with comments, and the comment's reply. Due to the terms and conditions of YouTube API, the commentators' names were removed from the analysis. Additionally, comment replies were removed to avoid deviating from the videos' main discussion and focused on other data properties.

3.4.3 Twitter API:

Twitter provides a number of API services to developers (third-party) to build applications that incorporate social network services such as retrieving tweets[4]. Because Twitter offers different API services, it is essential to determine which API to use when acquiring Twitter data. Most research investigations on Twitter depend primarily on using the Streaming API or the REST API [130].

Unlike the REST API, the streaming API is commonly used in quantitative studies

[4] https://developer.twitter.com/en/docs/twitter-api

using Twitter data. [130] Also known as a push-based service, the Streaming API streams data in real-time. To remain online and connected to the server, researchers may have to design and develop their applications [130]. The Streaming API is available in three different bandwidths: *(1) 'Spritzer', (2) 'Garden-hose',* and *(3) 'Firehose',* each of which can transmit one percent, ten percent or even a hundred percent of tweets over a specified period [130].

Even though the first bandwidth (Spritzer) is available for research purposes, other Twitter users may be granted permission to use it, subject to terms and conditions. The second bandwidth is often distributed to users who demonstrated a compelling need for enhanced access and to companies and licensed re-sellers. Like the search feature on Twitter web mobile applications, the streaming API aims to return relevant search results rather than providing everything from the Twitter social network, omitting some posts and user accounts from the results (Twitter Developers Blog, 2014). That led to early studies that examined Twitter being done through the REST API [130].

Meanwhile, the REST API differs from the Streaming API regarding the techniques available to users. The former is a pull-based interface, which means that a researcher or developer can build a programme that can effectively communicate (request and response) with the Twitter server. Accordingly, this study uses the REST API as it allows a researcher or developer (as a primary source) to obtain the data directly by using defined query terms or phrases without necessarily depending on the secondary source or purchasing the data [53]. This API allows access to previous data - returns tweets posted between six to nine days in the past. Moreover, the API provides user data and its social structure.

Although REST API use in searching tweets could save costs, it has a significant drawback of accessing limited historical data [53]. Despite its disadvantages, it is a practical technique for acquiring Twitter data. It allows researchers to obtain relevant data based on defined search terms to address their study's research questions. Hence, the purpose of its use in this study. Because only 100 tweets could be returned in a single request, and only a maximum of 180 requests could be performed in 15 minutes, it is unclear how Twitter returns previous data (tweets) posted six or nine days in the past. Consequently, this study iteratively passes the lowest tweet ID of the retrieved tweets to the maximum ID parameter in the subsequent call to avoid getting the same data results for any request.

Phase II: Twitter data extraction and analysis

Sourcing data from Twitter: Twitter has several advantages in terms of data analytics. Firstly, the short character length of tweets (280)[5] results in a highly homogeneous corpus. Secondly, access to large data is permitted due to the millions of tweets posted every day. Thirdly, the tweets are openly accessible for researchers or developers and can be extracted through API. However, as learned from [53], acquiring data from Twitter could be done in one of the three procedures mentioned below.

- Using the Twitter API
- Partnership with people who already have the data, or
- Purchasing the data.

Twitter offers various APIs (including premium and non-premium) for researchers and business organisations.[6] While these APIs allow ease of data extraction from Twitter, the size or volume of data to be extracted depends on the type of API used by the developer account. Therefore, this study created a Twitter developer account and used API (free version) to retrieve tweets posted using *self-injurious* hashtags.

This research study considered collecting Twitter data through the use of the API. Thus, this research chooses the first strategy due to its practicality in effectively extracting relevant data to answer the research questions. One of the procedure's benefits includes defining search phrases rather than depending on data from third-party sources [53]. The flowchart in Figure 3.4 shows the data collection process from Twitter. While using keywords or hashtags to extract data from Twitter, performing a quick search using the tags is always a good practice to check if it returns the expected or relevant data.

[5] https://developer.twitter.com/en/docs/counting-characters

[6] https://developer.twitter.com/en

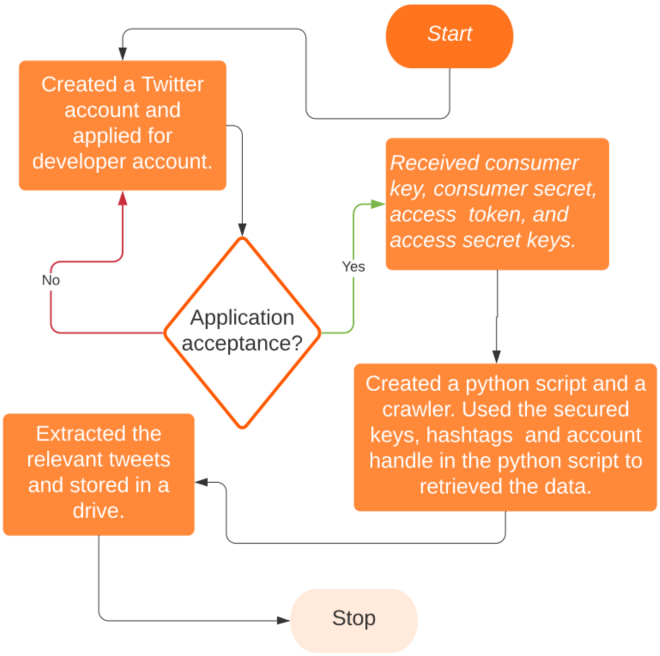

Start
Created a Twitter account and applied for developer account.
Application acceptance?
No
Yes
Received consumer key, consumer secret, access token, and access secret keys.
Created a python script and a crawler. Used the secured keys, hashtags and account handle in the python script to retrieved the data.
Extracted the relevant tweets and stored in a drive.
Stop

Figure 3.4: Tweets extraction flowchart

Therefore, before creating a custom crawler to retrieve the tweets, this study used the Twitter Archiving Google Sheet TAGS,[7] which is a cloud-based system that searches for tweets using a defined query or hashtag. The TAGS system was proposed by Martin Hawksey, an educational technologist [23]. Many researchers have widely used the tool to search and extract tweets using keywords or hashtags. The tool has been used [131] to understand the spread of misinformation on Twitter. Additionally, other researchers used TAGS to collect Twitter tweets and analysed the platform's influence on learning [132].

The idea of using TAGS was to have a reliable understanding of the appropriate hashtags to retrieve tweets concerning intentional self-harm. After one month (from 02/02/2020 to 07/03/2020) of data collection using the TAGS system and three popular self-injurious hashtags - *#selfharm*, *#selfinjury* and *#selfcutting*, a preliminary analysis of the data was performed. It was found that the *#selfcutting* hashtag is not returning tweets related to self-harm. Instead, the tag often returned tweets about cutting hairstyles. Therefore, the hashtag (including the data collected by tag) was

[7] https://tags.hawksey.info/

Table 3.1: List of Support Handles. Source: [5]

Organization	Twitter Handle	Followers count
Mind[a]	@MindCharity	461,665
Selfharm UK	@selfharmUK	4,409
Shout UK	@GiveUsAShout	9,476
Self injury Support	@sisupportorguk	3,216
LifeSIGNS	@LifeSIGNS	3,412
Mental Health Notes	@depressionnote	192,644
The WISH Centre	@TheWISHCentre	3,946
Samaritans[a]	@samaritans	132,939
Stop Self Harm	@stopselfharm	21,290
YoungMinds[a]	@YoungMindsUK	164,966

[a] Organizations listed by the NHS.

removed in preparing the data for analysis. Similarly, the *#selfinjury* hashtag was not considered in the second round of data collection as the tag tends to retrieve tweets related to accidental injuries rather than intentional self-injury.

Therefore, in the second round of data collection, this study extracted tweets concerning self-harm using only the *#selflharm* hashtag. As discussed in Section 3.4.3, one of the shortcomings of searching and retrieving tweets using a *keyword* or *hashtag* is the difficulty obtaining historical data - usually returns data of not more than nine days in the past [53]. Moreover, due to the nature of our research questions, this study retrieved data from Twitter in two cases. In the first case, tweets were crawled using the (*#selflharm*) hashtag. Meanwhile, to answer research questions explained in Section 1.3.2, this study extracted data directly from Twitter handles (see Table 3.1) of organisations supporting self-harming people as recommended by the United Kingdom National Health Service (NHS).[8]

Twitter data analysis in the previous section, this doctoral study reported the procedure for collecting data from YouTube social media and how the examined data were analysed. This section described how the Twitter data analysis was performed to gain more understanding of the data and answer the research questions effectively.

[8] `https://www.nhs.uk/conditions/self-harm/`

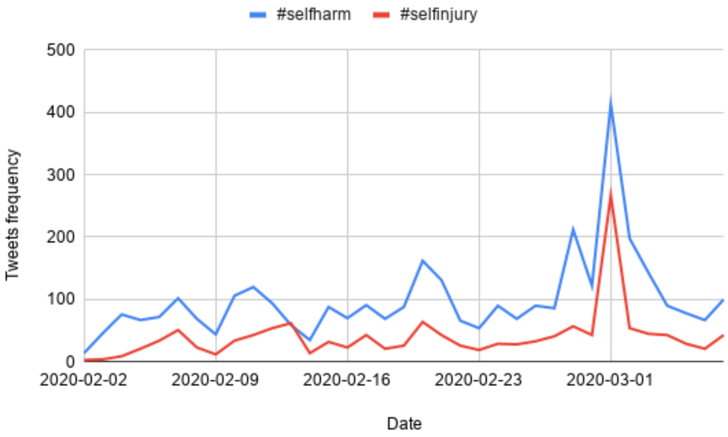


Figure 3.5: Frequency of tweets containing self-harm and self-cutting hashtags, source: [4]

An existing study highlights that research investigations performed on Twitter could be (1) Content analysis, (2) usage of social media, and (3) Future use of social media for research [133]. This dissertation focused on analysing self-injurious content (tweets containing self-harm) on Twitter.

Therefore, our analysis began with investigating the frequency of tweets generated by *#selfharm* and *#selfinjury* hashtags. The goal was to understand the popular hashtag used by online members in tweeting about self-harm on Twitter. It was observed that the number of tweets containing *#selfharm* was twofold those posted using the *#selfinjury* as shown in Figure 3.5. The line graph shows that many tweets containing both hashtags were posted on 01/03/2020. This date is globally recognised as a self-injury awareness day. Despite social distancing and lockdown caused by the coronavirus in 2020, the spike from the figure shows that various users participate in the online campaign to raise awareness about self-harm on Twitter.

By using the TAGS system, this research study retrieved tweets containing *#selfinjury* and *#selfharm* hashtags. After removing duplicates and irrelevant tweets, this study analysed 4,875 unique tweets from 8,523 posted by 3,421 tweeters. From the examined tweets, 1,394 were posted using the *#selfinjury* hashtag, and 3,481 tweets contained the *#selfharm* hashtag. Because this study is based on the English language, our analysis uses tweets found in English (up to 94.5%) and ignores the tweets written

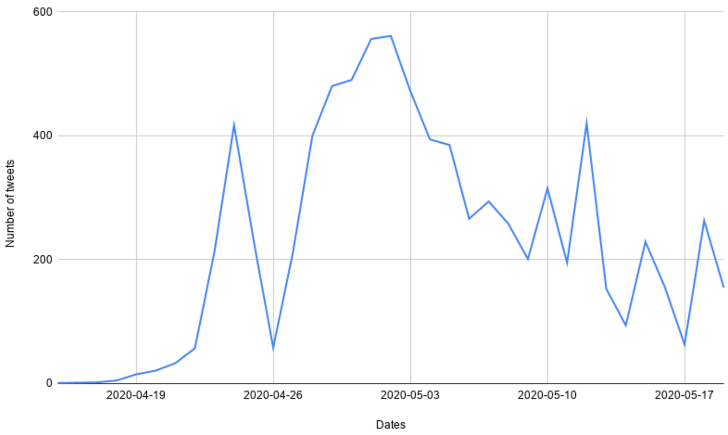


Figure 3.6: Frequency of tweets containing self-harm hashtag.

in other languages that are less than 5% of the data.

Consequently, our analysis indicated that the *#selfharm* hashtag is the popular tag used by people to post information related to intentional self-harm on Twitter. Therefore, to understand self-harm-related discussions on Twitter, this research specifically used the *#selfharm* hashtag in the second round of data collection. The idea was to obtain relevant data to address the research questions in the phase two study. Therefore, a custom Twitter data extraction crawler was developed. Hence, the API - including the access tokens, were used to extract a collection of 8,063 unique tweets. Different users between 15/04/2020 and 19/05/2020 posted these tweets, including 7,914 replies. Again, to ensure proper representation, this study considered tweets written in English, representing 98% of the entire set of tweets. As shown in Figure 3.6, the line graph shows fewer rate deviations from the analysed tweets. It is observed that over the examined period, there is a prevalence of tweets concerning self-harm on Twitter.

However, our study analysed the set of tweets extracted from the support handles listed in Table 3.1. Moreover, the table shows the names of the support organisations, the Twitter account handle, and the number of followers (at the time of data collection). Additionally, the data extracted from the support organisations were from two different periods. In the first period, tweets were extracted between 2019-08-22 to 2019-11-19. This period was before the first index of the coronavirus (COVID-19) in the United

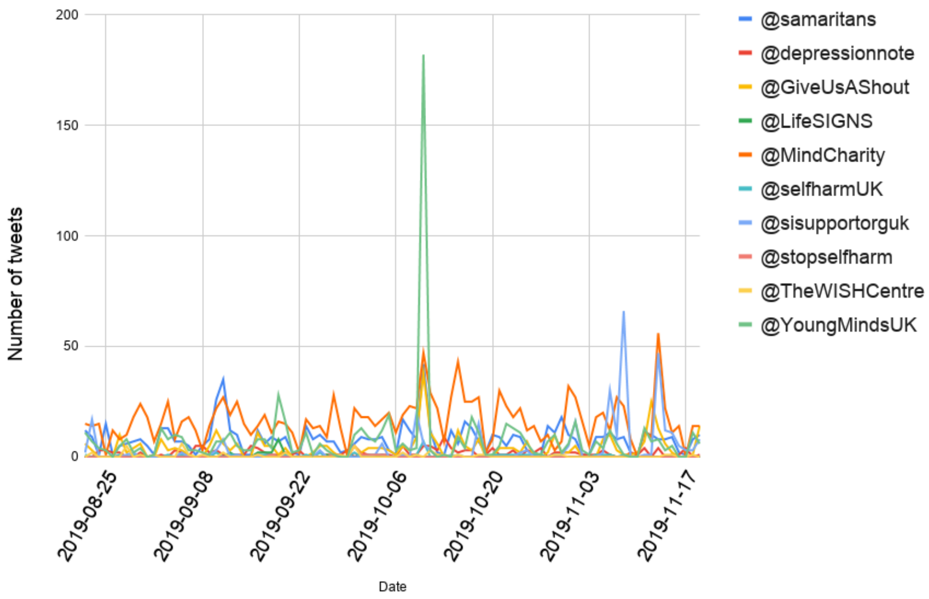


Figure 3.7: Tweeting behaviour of the support handles before COVID-19. [5]

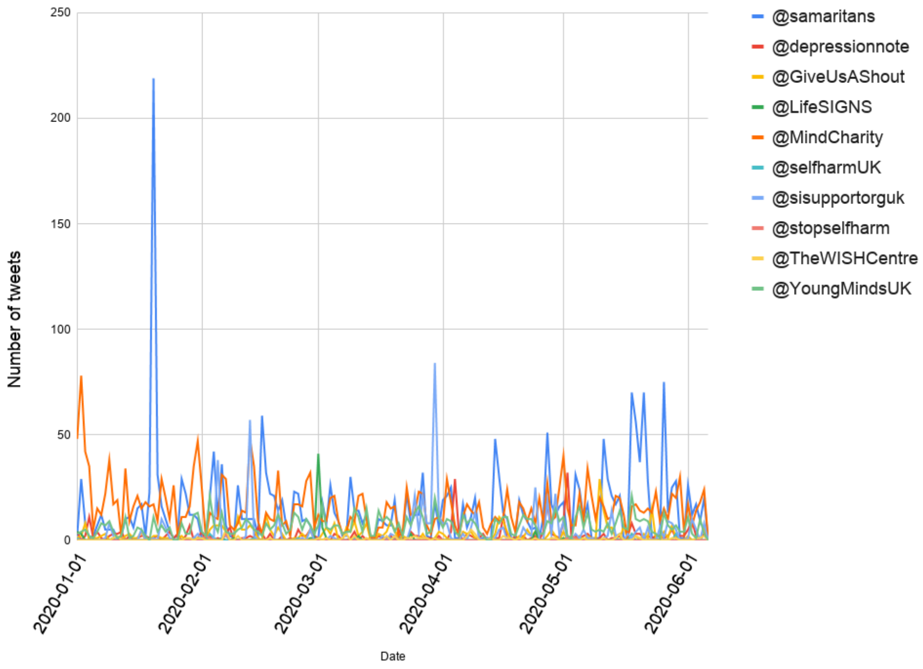


Figure 3.8: Tweeting behaviour of the support handles during COVID-19. Source: [5]

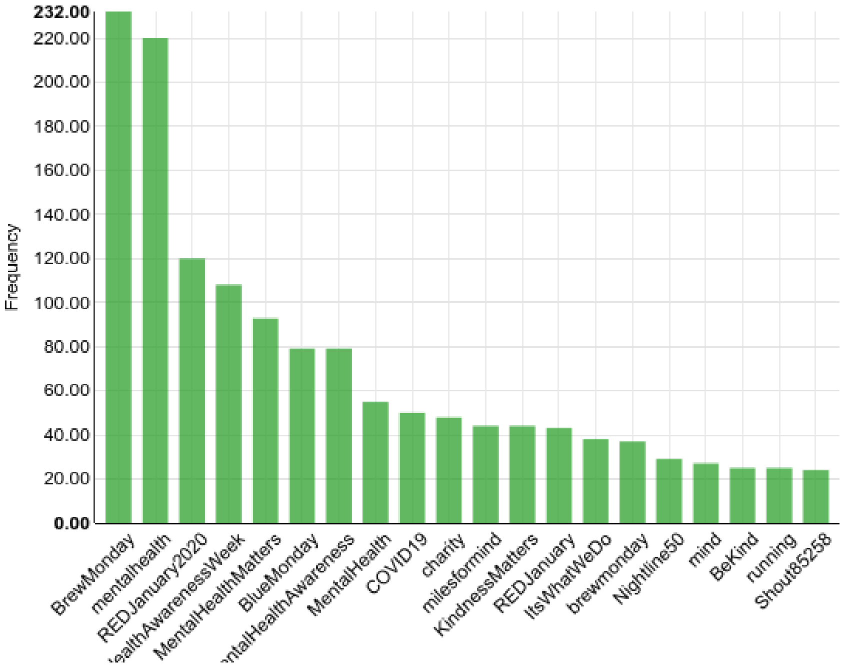


Figure 3.9: Top 20 hashtags from the support handles tweets. Source: [5]

Kingdom (UK). Meanwhile, the second period of the data collection was between 2020-01-01 to 2020-06-06. This period was during the COVID-19 as the first index of the virus was recorded in the UK on 23-01-20 [134].

Similarly, this study examined the tweeting behaviour of the support handles from the two periods. The data were analysed in two separate periods to investigate the support handles' behaviour (tweeting activity) before and during the global coronavirus (COVID-19) pandemic. As shown in Figure 3.7 and 3.8, all the examined Twitter handle posted tweets between the stipulated data collection period. The spike in the former figure came from the **YoungMinds** organisation on the 10th of October 2019. This date is recognised as the mental health awareness day globally.

Furthermore, in the latter figure, a higher rate of tweets from the **Samaritans** organisation was observed on the 20th of January 2020. This date was considered the most depressing day (blue Monday) of 2020. Consequently, our analysis shows that these support handles are socially engaging with many users from around the world on Twitter to raise mental health issues campaigns.

Moreover, our analysis discovered that in August and November 2019, the most socially active support organisations were *@MindCharity*, *@Samaritans*, and *@YoungMindsUK*. These handles were ranked in order of tweeting activities over the period. In early 2020, there was a change between the handles in terms of their tweeting behaviour. The *@Samaritans* organisation became more active, followed by the *@MindCharity* and *@YoungMinds* accounts. While Twitter allows users to post tweets that are not more than 280 characters, people tend to include hashtags in their tweets to facilitate information search or sharing among online users. Therefore, our analysis explored the top 20 hashtags from the set of tweets extracted from support organisations' accounts.

The top twenty hashtags used by the support organisations in posting self-harm and mental health-related tweets are depicted in Figure 3.9. The *#selfharm* and *#selfinjury* hashtags do not appear in the first twenty hashtags. Meanwhile, the *#BrewMonday* and *#mentalhealth* hashtags were the most frequent hashtags found from the organisation's tweets. While the hashtag *#RedJanuary* appeared in most of the handles' tweets, it was evident that these organisations remained socially active in increasing support and awareness concerning mental health issues. In addition, because our data collection period was during the global pandemic (**COVID-19**), the examined shared information related to the virus using the *#COVID-19* hashtag.

In contrast to the tweets from the support handles, this analysis examined the tweets posted using the *#selfharm* hashtag to determine tweeters' most frequently used hashtags while posting information about self-harm. As shown in Figure 3.10, our investigation discovered that users often use hashtags associated with anxiety, autism, and depression which are all associated with intentional self-harm. Additionally, tweeters often use hashtags like *#addiction* and the *#suicide* hashtag in the examined tweets. Even though self-harming people have no intentions of killing themselves, evidence confirmed that strong addiction to self-harm is associated with self-injury [55]. Therefore, our analysis indicates that users and support organisations share common hashtags to communicate about well-being and mental health issues associated with self-harm.

On the other hand, while manually examining the set of hashtags, our analysis found various unique hashtags online users use to post information about self-harm.

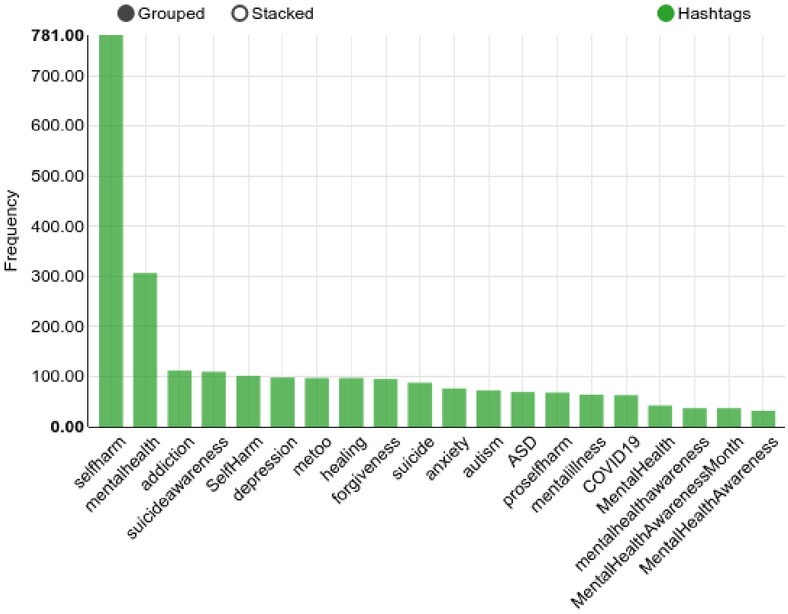


Figure 3.10: Top 20 hashtags from the user's tweets, source: [5]

The #shtwt hashtag is a clear example of how users use utilise ambiguous hashtags to avoid being notified or recognised by social media teams or people reporting online users sharing harmful content. However, it was observed that a new hashtag (*#shtwt)* that represents self-harm tweets was predominantly used by online members to communicate about self-harm. Consequently, this doctoral study crawled 12,102 tweets using the *#shtwt* posted between 18-02-2021 and 21-06-2021. Therefore, the idea was to investigate what people posted and discussed on Twitter using *#shtwt* and *#selfharm.*

This study retrieved top hashtags in the *#shtwt* dataset as depicted in Figure 3.11. Except for the *#shtwt* hashtag, it was observed that most users frequently used both *#shtwt* and *#edtwt* (eating disorder tweet) while posting information related to self-harm and eating disorders. Thus, the findings are similar to a recent study demonstrating the link between the duo on social media [39].

3.4.4 Data cleaning - Text processing

Cleaning data is one of the fundamental steps of data analysis. This research focused mainly on the textual data (tweets and video comments) relating to deliberate self-

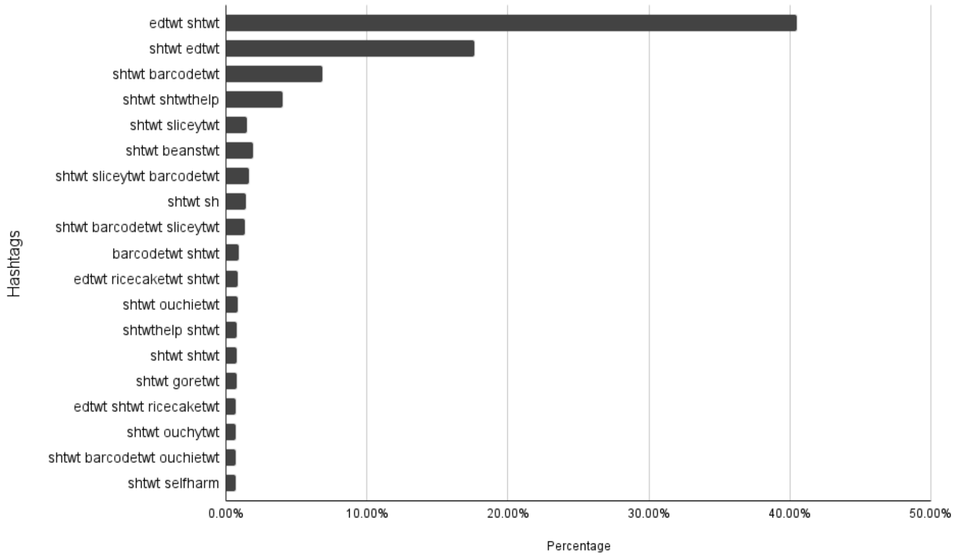


Figure 3.11: Top hashtags from #shtwt tweets

harm. Accordingly, this text collection requires proper cleaning for topic analysis, sentiment analysis, and machine learning tasks. However, before analysing the data, it is crucial to use the basic text processing techniques to clean it and prepare it for analysis. Thus, using the below traditional text cleaning and processing method is essential.

- Conversion to lower case
- Tokenization
- Stopwords removal
- Removing noise such as punctuations, tags and irrelevant character.
- Stemming
- Lemmatization

3.4.5 Ethical issue

This study considered the three critical components of a framework proposed by [135], which is an ethical guideline for research involving social media data.

Legalities

Before data collection, this research accepted the terms and conditions of using YouTube and Twitter. Although these conditions could be changed over time, they often included information about how one's data will be used by third parties, including researchers. Therefore, this poses an ethical issue regarding how to use and protect the privacy of users' data so that it does not reveals individuals' personal information or violate people's privacy. The Department of Computer and Information Sciences ethical committee was contacted regarding this issue. The ethical unit advised that submitting an ethical application is not required for this study as the data is publically available and accessible online on the platforms.

Privacy and risk

The use of social media data creates some concerns about whether the user-generated data is public or private. However, it is vital to remember that most social media users accept the terms and conditions of the platform they use. Anyone on YouTube can access videos considered in this study and no permission regarding who can comment on the videos. Similarly, the data extracted using self-injurious hashtags and support handles are publicly available on Twitter.

However, self-harm is a sensitive topic, and children could be involved in its discussion on social media. Thus, putting them and anyone who participated in the discussion at risk. Consequently, in the analysis of this study, the user's account handle and any identifiable information were concealed to protect users from risk.

Reuse and publication

This research takes into consideration the reuse of the data as some parts of the data will be used to report findings. The legalities, privacies and risks were considered in reporting the results of this research, including the research publications. Some parts of the data that present users' identifiable information were modified to ensure anonymity and protect users' privacy. Moreover, any sensitive information that could violate the users' privacy will be removed before sharing with a third party - future researchers. Also, interested researchers will be advised to provide evidence of accepting the conditions of using the social media platform before obtaining the data for their research.

3.5 Machine learning approach

Machine learning has advanced considerably over the last two decades and has evolved into designing powerful applications concerning computer vision, image recognition, and natural language processing [114]. Machine learning has significantly reshaped research in computer science and other disciplines, as the technique has changed how researchers analyse data (especially a vast amount of data) in unexpected ways. As a subarea of computer science and Artificial Intelligence (AI), Machine Learning (ML) uses algorithms and data to mirror how humans learn while continuously improving its accuracy [114]. Moreover, ML is a significant part of data science that is gaining popularity and overgrowing. While training algorithms to create classifications or prediction models utilising statistical techniques, the approach facilitates data mining initiatives to discover crucial insights [114]. Furthermore, there are various techniques of machine learning that can be used to build a classifier or prediction model, as briefly described in the next section.

3.5.1 Machine learning techniques

From the examined studies discussed in Chapter 2, about 10% used machine learning techniques in their studies. Even though this percentage is not sufficient to generalise the effectiveness of this method, evidence from computing fields confirmed its applicability to social media research [53]. Supervised and unsupervised machine learning techniques are state-of-the-art methods for analysing social media data. Similar studies demonstrated the effectiveness of these approaches on Flickr citewang2017understanding, and Instagram [24]. Consequently, the ML learning analysis performed in this study uses the below-supervised ML algorithms.

Supervised learning

Supervised learning uses labelled datasets to train models capable of effectively classifying data (as the input data is provided into the model) and predicting results [114]. The model's weights change until it is fitted appropriately, a process known as *cross-validation.* The cross-validation method assesses the ML model by training many models on a subset of the input data. The method evaluates the model's accuracy using data different from the data used in training [136]. While the technique indicates the model's ability to analyse unseen data, its goal is to detect overfitting. Accordingly, this study used the *cross validation* technique to validate the effectiveness of the proposed ML model.

Meanwhile, supervised learning guides models to create the desired output while using a training set. This training dataset includes both true and false results, enabling the model to develop and improve over time. In dealing with data mining issues, supervised learning could be categorised into two classes of problems: classification and regression [114]. Classification is a method that uses an algorithm to classify test data appropriately. It recognises different entities within the dataset and attempts to provide inferences on how they should be labelled or expressed [114]. The most popular classification techniques are random forest (RF), k-nearest neighbour (KNN), and support vector machines (SVM).

On the other hand, regression is a statistical technique that is used to determine the correlation between dependent and independent variables [137]. This method is frequently used to make predictions, and the common algorithms include logistic regression, polynomial regression, and linear regression [137], Even though both algorithms are supervised learning techniques used for prediction with labelled datasets, the distinction between them is how they are applied in ML tasks. Classification algorithms predict discrete values, such as male or female, while regression algorithms are used to predict continuous values, such as price or salary [137].

Classifying self-harm comments is a complex activity. It requires proficient effort from mental health and social computing researchers. Meanwhile, supervised learning methods employ a variety of classification algorithms and allow for labelling the data (comments) into the desired categories. Hence, the purpose of its use in this research. As discussed in the subsections below, some of these algorithms were used in classifying comments concerning videos presenting information about self-harm on YouTube. The idea was to compare the performance accuracy among the classifiers and build a model with good performance accuracy.

K nearest neighbour (KNN): the KNN is a non-parametric approach for classifying data points based on their distance, closeness, and relationship with other relevant data [137]. The KNN is based on the assumption that similar data points can be located nearby. Thus, it determines the distance separating data points, typically using Euclidean distance, while allocating a class based on the frequently occurring class [137].

Additionally, the KNN is applied in the study as the algorithm is easy to use and consumes a short computation time. This and many other advantages make it a preferred technique for data scientists. However, as the test dataset increases significantly, the processing time also increases. Therefore, it has become a less attractive supervised learning technique, although it is commonly used in the recommendation and image recognition systems [138, 139].

Random forest (RF): the RF is another supervised learning technique that is commonly applied for classification problems. The term "forest" refers to a group of unrelated decision trees combined to lower variation and produce more precise data predictions [137]. Moreover, this study applied RF due to its adaptability. The algorithm can be applied in regression and classification problems, and the relative priority it assigns to the input features is easily visible [137]. Random forest is also promising because it frequently uses the default *hyperparameters* results in accurate prediction [137]. Even though there are many *hyperparameters*, understanding them is often self-explanatory.

On the other hand, the primary disadvantage of random forest is that many trees might render the method inefficient for real-time prediction [137]. Generally, RF algorithms are quick to train though they are incredibly slow to generate predictions after training [137].

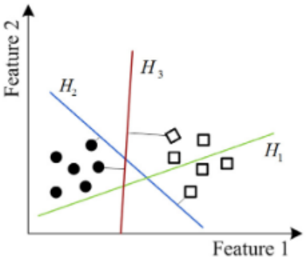


Figure 3.12: SVM Hyperplane, source: [6]

Support vector machines (SVM): SVM is one of the popularly recognised supervised learning techniques introduced by Vladimir Vapnik [115]. It is used for both the classification and regression of data. Intuitively, it is frequently used to solve classification problems by generating a hyperplane in which the distance separating two sets of data points is most significant [6, 115]. This hyperplane is the decision boundary that separates the classes of data points on either side of the plane. As shown in Figure 3.12, hyperplane *H1* does not distinguish the classes. Although there is a narrow margin, *H2* divides the two categories slightly, and *H3* classifies the two groups with a significant margin.

While *n* refers to the number of features, the goal of the SVM is to identify a hyperplane (using *n-dimensional* space) that categorises the data points accurately. In a collection of training sets labelled as two classes, a trained binary SVM classifier could be built to classify a new dataset into one of the two categories. Within the two classes, one hyperplane outlines the most significant margin. The distance separating the selected hyperplane and the closest data point of the individual class is enlarged [6]. Usually, categorising a dataset has been a nonlinear task.

Therefore, the linear SVM, also known as SVMLinear, could be expanded to a nonlinear version that executes nonlinear classification effectively using the kernel method. The dataset's initial instances are outlined in a dimensional area utilising the kernel function. The SVM determines the ideal hyperplane to categorise the data in the dimensional space effectively [6]. Although this study explained using SVM in a binary classification task, the supervised learning method can be applied to a multi-class problem. The analysis considered the issue of detecting critical self-harm responses on

YouTube as a multi-class problem. The objective was to compare the performance accuracy of the classifiers in binary and multi-class scenarios. The former group consists mainly of red and green comments, while the latter have amber, crisis, green, and red.

Unsupervised learning

The unsupervised machine learning method examines and groups unlabeled datasets using machine learning algorithms such as the K-means algorithm. Without requiring human annotation or labelling, these algorithms uncover hidden information or pattern from unlabelled datasets. One of the strengths of the unsupervised learning approach is the ability to identify similar and dissimilar patterns from a given dataset. Thus, making it a suitable approach for analysing and exploring data. The unsupervised learning techniques are used in performing (1) association rules (2) clustering, and (3) dimensionality reduction.

However, this research study performed topic analysis using the Latent Dirichlet Allocation (LDA - explained in Section 3.7.2) algorithm, an unsupervised learning technique for investigating topics from an extensive collection of documents [126]. The data examined in this thesis were textual comments from deliberate self-harm videos on YouTube and tweets containing self-injurious hashtags. Thus, using the LDA technique, this study explored topics users discuss concerning self-harm on YouTube and Twitter. The analysis and findings are discussed in Chapters 4 and 5.

3.6 Reverse engineering approach

Analysing a current system to gain design information is called Reverse Engineering (RE) [140]. That is to say, the task of recovering the design of a functional software system without accessing the code is called RE. The goal is to enhance re-documentation and product (software) services. Traditionally, the reverse engineering technique is often used in software engineering. A recent study has used the RE technique to understand the behaviour of Twitter bots. The researchers proposed a tool (Botscope)[9] and used

[9] https://github.com/bellobichi2/botscope

it to understand the behaviour of bots accounts on Twitter during the Brexit period in the UK [141]. The approach used by the authors yields promising results and can be applied to other related research - hence the purpose of its use in this research. Hence, the purpose of adopting the approach in examining the strategy of support accounts supporting self-harming individuals on Twitter. The objective is to analyse actions such as tweets and replies posted by the support handles and understand their campaign strategies.

3.7 Research tools

Due to the nature of our research questions and the data used, tools such as python's NLTK, LDA, and VADER were used. Existing studies used these tools to uncover hidden themes and sentiments from social media data, especially in computer and information sciences [142, 143]. These tools effectively work on python by installing and importing all the necessary and relevant libraries.

3.7.1 Natural Language Toolkit (NLTK)

The natural language toolkit, popularly known as the NLTK, is the cutting-edge platform for developing Python programs that interact with data representing natural human language. The platform offers user-friendly interfaces to more than fifty corpora and lexical resources. [10] These include WordNet [144] and a suite of libraries for performing basic text processing. Natural Language Processing (NLP) is increasing in popularity as its techniques and theories are applied in various disciplines. In the field of computer science, the use of NLP on human language data is gaining rapid attention. Social media content contained human language data such as comments and tweets. This data could reveal various information about the users, such as their feelings and mental well-being.

The techniques of NLP could infer individuals' mental states from their tweets on Twitter, comments on YouTube videos, and other textual posts on social media

[10] https://www.nltk.org/

platforms. The inferences could be utilised to provide online support and connect individuals struggling with self-harm to helpful information and support. People's writings on online digital spaces such as social media could be automatically processed. While analysing human writings, NLP methods infer what individuals speak and their opinion. The inferences obtained can be utilised to trigger messages or other activities. Although the approach is applicable to various fields, the technique is widely used in marketing. Business organisations analyse emails or Google query terms and social media content to generate adverts and other types of 'interventions' to understand or influence our preference for purchasing items online.

However, the NLP techniques are not limited to advertising and marketing purposes. The procedure could be used to understand social media users' mental states, such as predicting depressed users on social media [38] and detecting suicide posts [145]. Therefore, given the above brief discussion on python's NLTK as supported by relevant studies, this PhD study used the tool to analyse the tweets containing self-harm hashtags and responses from videos discussing self-harm on YouTube.

3.7.2 Latent Dirichlet Allocation (LDA)

The latent Dirichlet allocation (LDA) was proposed by [126], and the authors described the technique as a general probabilistic topic model for analysing text corpora. The LDA topic modelling methods are considered unsupervised ML techniques that employ probabilistic predictions in order to condense enormous amounts of text data into manageable topics [38]. The LDA approach is considered a hierarchical Bayesian model consisting of three levels. Each item or document in a collection is represented as a finite combination of topics. Therefore, every topic is modelled as an infinite combination of topic probabilities obtained from the hidden set of topics. The probabilities associated with each topic give a precise representation of a document. That is to say, terms with a high probability of emerging in concurrence with another word are assorted to produce a latent topic that qualitatively describes a content region in the text collection. These techniques continued to be used in a variety of contexts, including identifying common themes in product reviews [146] mapping themes throughout bodies of scientific literature [147], and identifying topics from social network's data [148]. As a result, the LDA probabilistic approaches are well suited for topic modelling analysis.

Another study used the LDA model and examined topics related to a mental health issue (dementia) on Twitter [149]. The investigators analysed Twitter discussions concerning dementia. However, the LDA's use in this dissertation is strongly inspired by the literature. Many studies highlighted the significance and efficacy of applying LDA to discover hidden topics from a large corpus [142,143,147]. This PhD study applied the LDA algorithm and identified the underlying themes related to self-injury discussions on YouTube and Twitter social media platforms.

Similarly, a research investigation extensively examined the effectiveness of utilising users' activity on social media to predict the degree of depression [150]. The purpose of their study was to determine the use of several features retrieved from Twitter users (such as topics from tweets and history of activities) to recognise depression. Their analysis examined the techniques for estimating depression severity that relies on objective data, such as the activity log data of social media users. The authors proposed a technique to detect a sign of depression among users through their written essays [150]. Using the LDA model [126], the researchers generate topics from the essays. By applying machine learning techniques and the detected topics as features, the researchers measured depression severity.

However, this demonstrated the significance of their research and the feasibility of using texts produced by an individual to evaluate the severity of depression in that individual. Therefore, as seen from the literature, LDA is a powerful probabilistic topic model that can be used to identify topics from extensive text collections. Intuitively, the use of this current state-of-the-art technique in this dissertation is strongly supported by the existing research.

3.7.3 Valence Aware Sentiment Reasoner (VADER)

This doctoral study used the valence aware sentiment reasoner VADER [127], one of the cutting-edge rule-based models for sentiment analysis of social media text. This section briefly described VADER, a validated gold-standard collection of lexical features, including related emotional intensity weights, developed using qualitative and quantitative methods [127]. The authors [127] combined these lexical characteristics by examining five generic principles that include grammatical and syntactical patterns for expressing and highlighting sentiment intensity. While their findings are positive, they demonstrated that VADER is an exceptional sentiment analysis tool that outperformed (and in many cases) eleven other well-recognised sentiment analysis techniques. In contrast to advanced machine learning algorithms, VADER is easy to implement and has notable benefits (reduces time and saves costs) that support its use in our analysis.

Firstly, it is both fast and computationally efficient without endangering accuracy. A text corpus that uses a short time to analyse using VADER may need several hours to analyse in an experiment involving supervised learning models like SVM because of training datasets. Secondly, the dictionary and rules employed by VADER are available, rather than being concealed in a black box accessible only to machines [127]. Consequently, the researchers examined and improved VADER. While presenting both vocabulary and the model, the rule-based approach makes the operations and performance of the sentiment analysis of the engine easier to understand and interpret for humans [127]. In line with this, the researchers believed that investigators from different fields working with the LIWC system could consider using VADER as the tool has been proved to be effective in detecting sentiments expressed via texts. Thirdly, VADER was self-contained and domain agnostic in adopting a human-verified sentiment lexicon and generic grammar as well as syntax rules. The rule-based model does not need an extensive training set and works well across domains. However, the authors demonstrated a straightforward, interpretable, and computationally cost-effective sentiment analysis technique. The approach is capable of producing effective results that defeat human raters.

As a rule-based sentiment analysis technique, VADER has been used in different settings and proved to be an effective tool for understanding sentiments expressed in a social media text. For example, a recent study applied VADER to predict the result of elections from Twitter data [151], and detect deviations in cryptocurrency through analysing user's comments [152]. While there are tools for analysing emotions and linguistic or psychological cues such as the linguistic enquiry word count LIWC [11], VADER has been used by mental health researchers in sentiment analysis concerning the affective micropatterns in tweets posted on the Twitter social network [153]. Nevertheless, the effectiveness of VADER, its simplicity and ease of use, and several strengths mentioned above make it suitable for this dissertation.

3.8 Summary

This chapter explained the research procedure adopted in this dissertation. The chapter discussed the step-by-step research procedure used in the dissertation towards achieving the study's aims and objectives. As seen in the literature review chapter, several research studies used various research methods in this field. While the dominant approach found from the existing studies is based on qualitative descriptive design, this doctoral research adopted a mixed-method approach in order to effectively answer the research questions. Existing evidence demonstrated the effectiveness of this method and its applicability in the context of social media and mental health research. Moreover, while the chapter provides an overview of the research design (based on phases), it also describes the current state-of-the-art tools applied in the present study. The next chapter describes the findings of the phase I study conducted on the YouTube social media platform.

[11] https://liwc.wpengine.com/

Chapter 4

Phase I: Investigating Self-injurious Content on YouTube

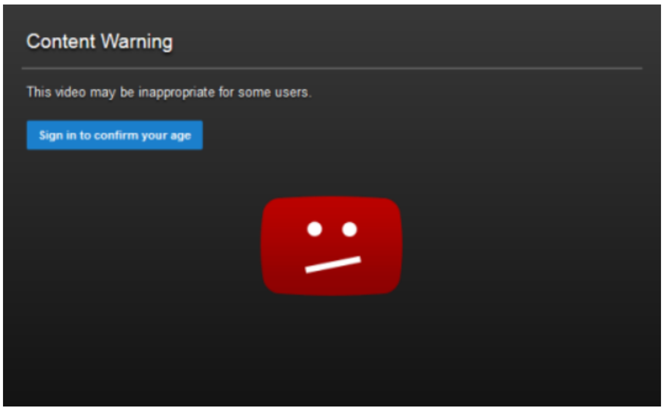
4.1 Introduction

In this dissertation, we review relevant studies and explained the research approach in the previous chapters. While highlighting the background of the study and the research questions addressed, the present chapter discusses the research findings of the phase one study conducted on YouTube social media.

4.2 Results

4.2.1 Sources of videos discussing self-harm on YouTube

As discussed in the previous Chapter - Section 3.4.2, the phase one study used YouTube's API to retrieve 250 videos about self-harm on YouTube using five different query terms. The entire responses (comments) from the unique videos were extracted for analysis. Section 3.4.2 of Chapter 3 provides a detailed explanation of the data collection procedure. However, a coding scheme (see Figure A.1 of the Appendix) was developed to answer the first research question. The idea of using the *codebook* was to guide the



Content Warning

This video may be inappropriate for some users.

Figure 4.1: YouTube's age restriction message

classification process.

Accordingly, the data extracted from YouTube provides valuable information to address our research questions. The data contained information about the video, such as the view counts, number of comments, and counts of likes and dislikes for each video. While the data included the title for each video, it lacked information (such as the channel's description) regarding the video channel or the YouTuber that uploaded the video. Therefore, this study manually searched YouTube social media and examined the sources of the videos. The name and description of each video channel were examined. Video sources were classified using the criteria described in Table A.1. The classification aimed to identify the distinct group of people (users) involved in sharing videos about deliberate self-harm on YouTube.

Similarly, our analysis examined video ratings from two perspectives. Firstly, viewers' ratings were analysed using the likes and dislikes associated with the videos. The idea was to understand how the audiences favoured or unfavoured the examined videos. Secondly, YouTube's content ratings were assessed for each video. The below list shows the critical elements of YouTube's ratings.[1] [2] The platform prohibits posting videos displaying self-harm behaviours. Therefore, it is essential to assess the nature of the video content as they were retrieved using queries containing self-injurious phrases. The objective was to explore the accessibility of those videos to children or under eighteen

[1] https://support.google.com/youtube/answer/146399?hl=en-GB#zippy=

[2] https://support.google.com/youtube/answer/4601348?hl=en

audiences and whether they contained violent graphic content.

- Strong language
- Nudity
- Sexual content
- Violence
- Drug use
- Flashing light

YouTube content moderators use the above content categories to assess video content. Figure 4.1 depicts a warning message on YouTube to indicate that video content could be upsetting or inappropriate to the audiences. Therefore, users are required to confirm their age before viewing the video. The criteria used by YouTube to ascertain a user's age could not be hundred percent reliable as people could register with false information. Meanwhile, users confirmed providing accurate information by agreeing to YouTube's terms and conditions. Hence, the study analyses the accessibility of the videos to underage audiences.

However, YouTube is a popular social media site that facilitates searching, sharing and viewing videos online. Regarding videos related to intentional self-harm on YouTube, there is a concern about sharing or uploading those videos as the content may encourage or discourage the behaviour [37]. To understand the sources of DSH videos on YouTube, our study analysed the video uploaders into any of the groups presented in Table A.1. Our analysis discovered the various groups of YouTubers sharing videos about self-harm. As seen from Figure 4.2, the vast majority of people (up to 56%) sharing videos about self-harm on YouTube are non-professionals. Most videos were uploaded by non-expert channels belonging to individuals (laypersons) who share information about self-harm.

Video channels representing organisations presenting self-harm-related information to support viewers struggling with self-harm occupied up to 12%. Meanwhile, channels

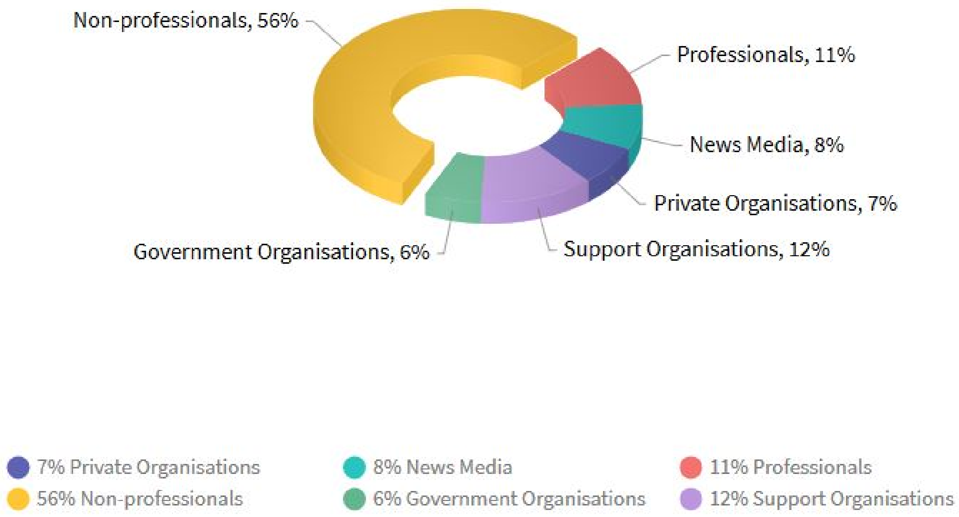


Figure 4.2: Sources of videos concerning self-harm on YouTube

managed by the group of medical experts and academic professionals represent only 11%. While private organisation channels account for up to 7%, only 6% of video channels represent government organisations. Accordingly, around 6% of the examined videos were from news media channels. Meanwhile, YouTube users can use different options to interact with the video-sharing site. While people can subscribe to video channels, they can rate videos by liking or disliking the content.

Similarly, people can socially interact with the video uploader or members of the online space through commenting. As illustrated in Figure 4.3, this study explored viewers' interactions from the identified group of channels. In other words, the view counts, likes, dislikes, and video categories from each group of uploaders were examined. From Figure 4.3, it is evident that each group of uploaders have a high number of subscribers. Majority of the viewers favoured (liked) videos uploaded by the *non-professionals*. Furthermore, the analysed videos posted by different channel groups, except those uploaded by government organisations, received significant responses. Compared to the other categories, the professional group received more dislikes than comments. Thus, indicating that viewers have unfavourably rated videos from this group.

Similarly, our investigation examined the various video content categories assigned by the group of video uploaders. The matrix table in Figure 4.4 depicts the rate of

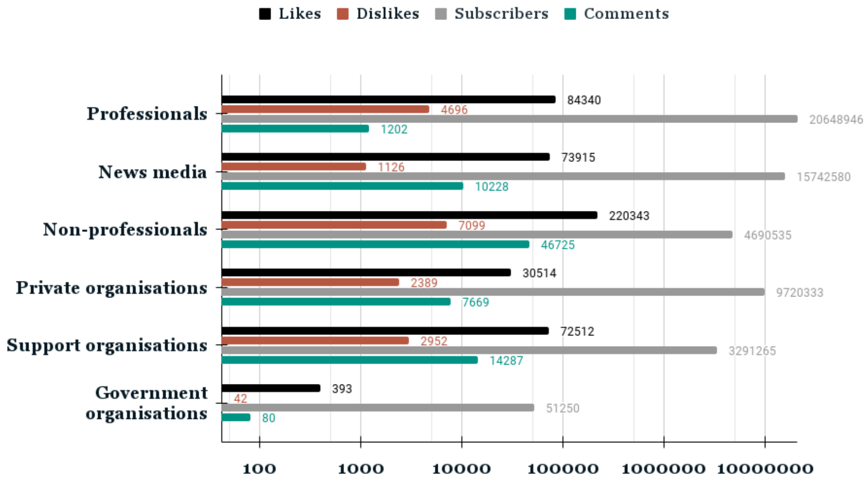
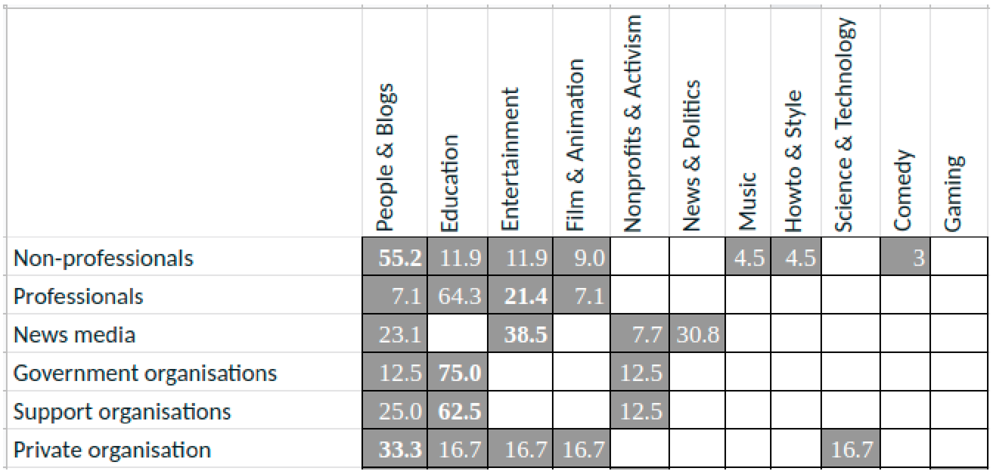


Figure 4.3: Frequency of interactions from different video sources



	People & Blogs	Education	Entertainment	Film & Animation	Nonprofits & Activism	News & Politics	Music	Howto & Style	Science & Technology	Comedy	Gaming
Non-professionals	**55.2**	11.9	11.9	9.0			4.5	4.5		3	
Professionals	7.1	64.3	**21.4**	7.1							
News media	23.1		**38.5**		7.7	30.8					
Government organisations	12.5	**75.0**			12.5						
Support organisations	25.0	**62.5**			12.5						
Private organisation	**33.3**	16.7	16.7	16.7					16.7		

Figure 4.4: Percentage video content category from different sources

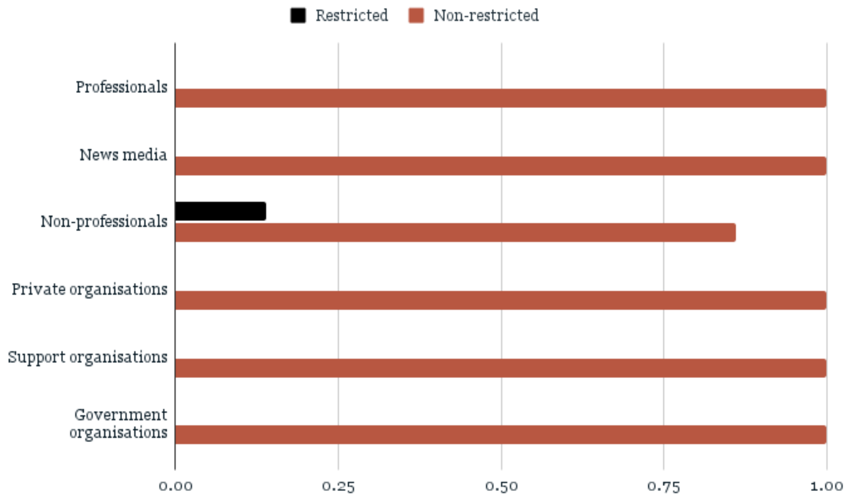


Figure 4.5: Accessibility of videos from different sources (in percentage)

content categories assigned by different uploaders. Numerous channels included at least one video in the Persons and Blogs category. Moreover, around 55.2% of the videos from *non-professional* sources were assigned to the People and Blogs category. This indicates that among the individual YouTubers worldwide, some who engaged in intentional self-harm make videos on YouTube to disclose their stories and increase awareness about the behaviour.

On the other hand, videos assigned to the Education category were found in all the sources except the *news media* group. Also, up to 75% of videos uploaded by *government organisations* were assigned the Education category. This percentage decreased insignificantly to 64.3% and 62.5% for *professionals* and *support organisations* channels.

Additionally, no video is assigned in the Entertainment category among the videos uploaded by the *government* and *support organisations*. Up to 30.8% of the *news and media* channels videos were found in the News and Politics category. This percentage slightly increased to 38.5% in the category of Entertainment. Consequently, these findings demonstrate that these groups of YouTubers socially engage with wider audiences on YouTube to enlighten them about self-harm and its related consequences.

Videos accessibility to underage viewers: Google regulates content to ensure safety and maintain standards. As one of their services, YouTube restricts access to specific videos to viewers under eighteen years. If a video contains violent or self-injurious content that could be triggering, YouTube signals, it as inappropriate and displays a warning message to the viewers, as shown in Figure 4.1. Therefore, this study explored the videos uploaded by different channels to identify contents that violate YouTube standards. In other words, our analysis discovered the rate of videos that contained graphic or violent content from the identified sources. The horizontal bar graph in Figure 4.5 shows the percentage of restricted and non-restricted videos from different sources.

YouTube social media restricted only 14% of those videos uploaded by non-professionals. This implies that those videos contained violent content or any attributes listed by YouTube.[3] [4] Therefore, those videos are not permitted to be viewed by children or people under eighteen years due to their graphic content. Additionally, self-harming viewers may find the videos inappropriate or triggering. Meanwhile, all the remaining uploaders shared videos that could be accessible by anyone, including children and individuals below eighteen years.

4.2.2 Self-harm discussions on YouTube

As a video-sharing site, YouTube allows viewers interactions through commenting on a video. People from various places around the world view self-harming videos on YouTube and discuss issues related to self-harm through comments. Hence, this doctoral study considered the comments extracted from the examined videos to address the second research question. This study employed the current *state-of-the-art* probabilistic topic model (LDA model) [126]. The goal was to investigate the video comments and understand the essential themes viewers discussed concerning intentional self-harm.

Even though YouTube video comments could have one or more replies, our experiment focused only on the video responses. Instead of investigating replies associated with the comments that could change the conversation or deviate from the topics of the

[3] https://support.google.com/youtube/answer/146399?hl=en-GB#zippy=

[4] https://support.google.com/youtube/answer/4601348?hl=en

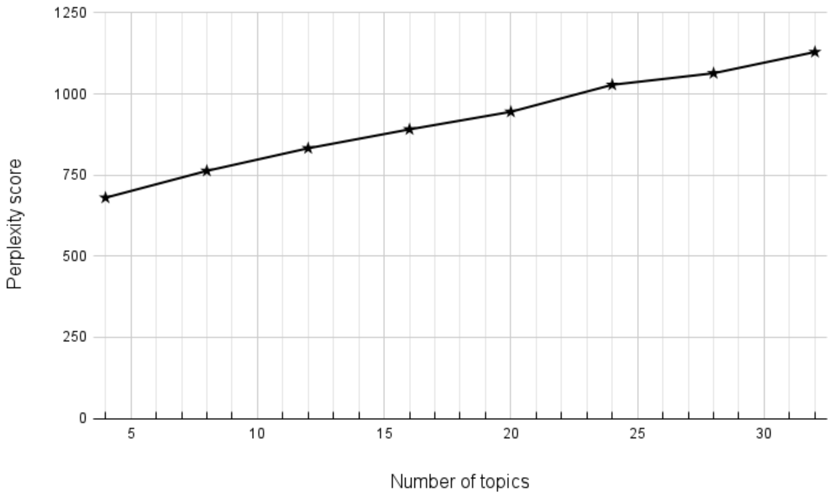


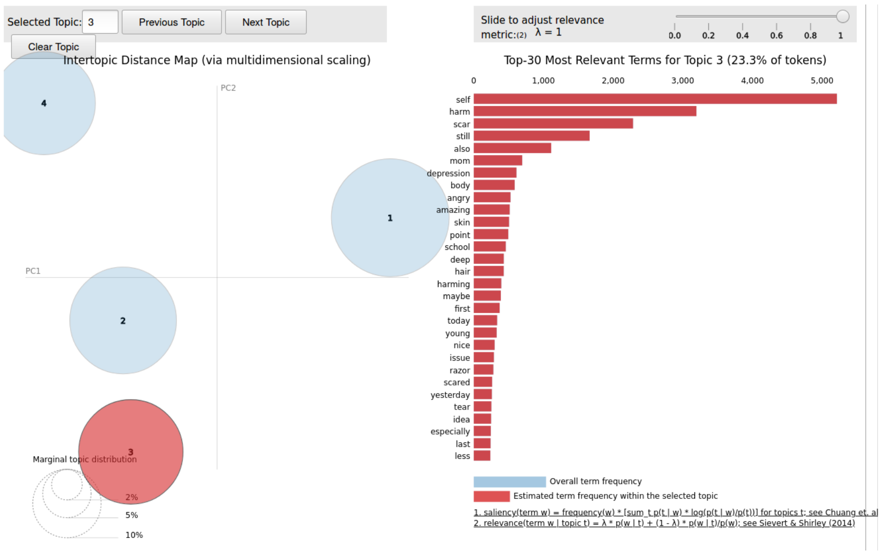
Figure 4.6: Choosing model's number of topics

discussions, our experiment focused only on the comments to understand the central discussions surrounding the set of videos. YouTube video comments are textual data and therefore require basic data cleaning. Because this is an NLP task, our analysis cleaned and processed the entire set of comments using the procedure explained in Section 3.4.4 of Chapter 3.

$$Perplexity(D_{est}) = exp\left\{ - \frac{\sum_{d=1}^{M} logP(w_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{4.1}$$

However, utilising the LDA technique to analyse extensive document collections requires choosing the appropriate number of topics [126]. Even though it is challenging to choose the number that best describes the corpus, our study used the *perplexity* metric presented in equation 4.1 and computed the performance of each model using a different number of topics. It is explicit from this equation that the absolute number of documents is expressed by M, and the probability of a word W appearing in a document is represented by P. In addition, the term N refers to the number of words contained within a particular document.

The idea of the perplexity metric is that the lower the perplexity score, the better the


Selected Topic: 3

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4.7: Visualising topic model using LDAVis [7]

model's performance [126]. However, this metric is essential to measure the appropriate number of topics in a large corpus. Consequently, our experiment grid searches for the best model using 4, 8, 16, 20, 24, 28, and 32 as the number of topics. While completing different rounds of experiments, the model with four topics achieved a good perplexity score, as presented in Figure 4.6. This study examined 27,520 video responses discussing self-harm on YouTube.

Accordingly, this study examined the comments corpus using the model and uncovered essential themes from the video responses. When analysing topics, it is crucial to consider their keywords and relevance (weights) [126]. Therefore, examining the frequency of specific terms in the video comments is essential. As seen in Figure 4.7, words that appeared in different topics, and those with a frequency more prominent than their weight, were observed as less significant. This study weighed the first ten words and interpreted the topics discussed by the commentators in Table 4.1. The first column of the table presents the topics that described the complete set of analysed comments. In this experiment, the video commentators were grouped based on similar topics. The idea was to identify the dominant topic discussed by the viewers.

Additionally, our experiment verified and visualised the results using LDAvis, a newly developed approach for understanding and visualising topic models [7]. LDAvis

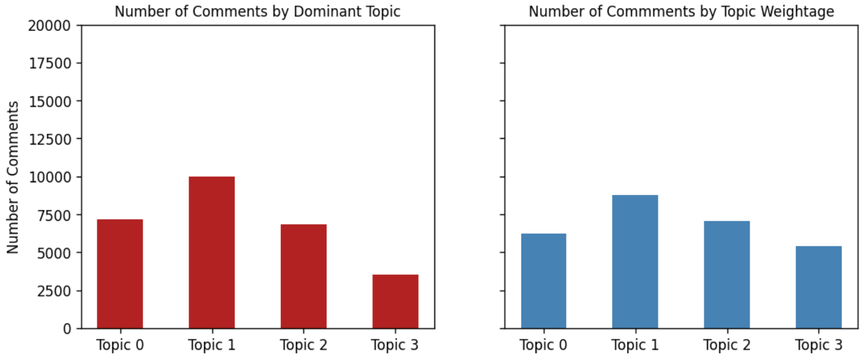


Figure 4.8: Dominant topics

is an online interactive system that enables more precise word inspections and topic connections. The LDAvis facilitates exploring the differences and relationships between topics interactively. From the left-hand side of Figure 4.7, the main topics of the model are shown in circles. The area surrounding each circled topic is proportionate to the length of the topics. The figure illustrates the distances between topics plotted in a two-dimensional space labelled using Principal Component Analysis (PCA). Thus, the LDA technique uses PCA to reduce N-dimensional vectors to two dimensions (x,y), representing PC1 and PC2 by default.

Meanwhile, on the right-hand side of the figure, the horizontal bar graph displayed the estimated word frequencies, including their weights in each topic and the complete corpus. When a topic is selected, its corresponding relevant terms are shown in a red colour horizontal bar graph. For example, when topic three was chosen, the figure shows the most relevant terms ranked according to their probabilities. Consequently, this process consists of (1) visualisation of the topics (using LDAVis), (2) obtaining the top most relevant terms for each topic and (3) labelling topics according to commentators with similar topics. Accordingly, the topic labels are as follows;

- Clean commentators
- At-risk audiences
- Self-harming users
- Appreciative commentators

Table 4.1: Topic interpretations and examples of comments

Topic	Unigrams	Example comments
Topic 0: Clean commentators	cut, people, know, stop, help, friend, month, want, year, clean	1. *"I started self-harm when I was 11 and I'm 12 now I've been doing it for 7 months and I'm 7 weeks clean right now"* 2. *"I'm clean 2 months and decided to stop hiding my scars, to stop hiding myself. I get a lot of comments daily and I'm already feeling ashamed of what I did."*
Topic 1: At-risk audiences	help, make, feel, thank, need, hard, care, pain, kill	1. *"I want to cut myself I use to be like this I wanna feel pain on my wrist I don't care - IDC JUST TEACH ME"* 2. *"I've been self-harming since I was 11 I'm almost 13. My cuts are deep and I need help. I tried killing myself two days ago by overdosing."*
Topic 2: Self-harming users	self, harm, scar, body, depression, skin, deep, razor, hair, scared	1. *"I self-harm and show my scars to prove the pain I have went through"* 2. *"I elbow my thighs until they bruise and the same with my arms and I wear long-sleeved sweaters and gym pants yes it hurts to walk but I've been doing it for 2 years ago."*

Topic 3: Appreciative commentators	get, really, video, love, people, understand, thing, could, life, go	1. *"Subbed. Because this video is really good. And I understand completely."* 2. *"I really appreciate that you made this video I am always so worried if I can show my scars or not but this helped a lot so thanks."*

While LDA allocates various themes to each document (comments), only a single topic could take up a significant portion of the document. Our investigation counted the total number of topics and identified the dominant discussed topic by the commentators.

As depicted in Figure 4.8, the theme representing comments indicating users could be at risk of self-harm and seeking help (32%) was the most widely discussed. Although topics representing self-harm experiences have been discussed at nearly the same rate (28%) as the at-risk topic, only a small proportion (nearly 18%) of comments appreciated the video content and thanked the uploader. Additionally, this percentage slightly increased to 22% for discussion regarding the self-harm recovery process. Thus, the finding adds to the existing knowledge about the nature of self-harm discussions on social media.

Most of the analysed comments to the examined DSH videos on YouTube consist of persons relating their NSSI experiences. A notable percentage of users who post online regarding self-harm communicate their self-harm behaviour. According to research in this field, people may post about their self-harm online to gain validation and acceptance [154] and to receive peer support from other self-harming individuals. The nature of the discussions surrounding the videos, such as the appreciative responses, demonstrates that commentators may have found relevant information described in the video.

Accordingly, when advising young self-harming individuals with hopeful and recovery-based messages, the information could be more helpful and influential if portrayed by their peers who have not only experienced but also recovered from self-harm. Existing evidence shows that the more the similarity between a message source (video presenter) and its audience (viewers), the more compelling the information [155]. As discussed in Section 4.2.1, this study discovered that most of the examined videos were from non-professional sources.

4.2.3 Comments sentiments analysis

It has been noted that apart from liking or disliking a video, it is difficult for viewers to react to videos using emojis on YouTube. The most straightforward way online members interact with the content uploader, presenter, or other audiences is through a comment. Unlike Twitter, with limited character length, YouTube allows viewers to express their opinion or react to a video using long character length. However, this study approached the third research question using the VADER rule-based sentiment analysis model. As a rule-based model, VADER classified the sentiment of a given text into negative, neutral or positive sentiments. The model uses specific rules to combine the effect of individual sub-text upon the perceived degree of the sentiment of a text at the sentence level [127]. These heuristics rules include:

- Punctuation
- Capitalisation
- Degree of modifiers
- Changes in polarity due to conjunctions, and
- Detecting negation polarity.

However, to calculate the sentiment score, VADER uses a compound score. This score is measured by calculating and adding the valence scores of all words in the lexicon. The computed scores are adjusted to ensure that they are within rules and normalise the scores between -1 and 1 [127]. Normalisation is achieved by using the below equation:

$$x = \frac{x}{\sqrt{x^2 + \alpha}}$$

The value of the valence score is represented by d, and α is the normalisation constant in which the default sore is 15 [127].

Since its introduction in 2014, VADER has been extensively utilised in social media data analysis to understand users' sentiments surrounding politics [141] and mental health research [156]. However, validation of VADER demonstrated that the rule-based sentiment technique works well in social media data. It has been confirmed that the tool exceeds human raters in analysing tweet sentiments [127]. In line with this, this study used VADER to examine the sentiments of viewers (based on their discussions) who commented on videos about self-harm on YouTube.

Studies on YouTube social media showed strong participation of community feedback via video responses [30, 157]. Investigating video comments could be an attractive data source for gaining implicit information about video commentators and their opinions. Sentiment analysis of video comments presenting self-harm information on YouTube can reveal viewers' interest and the role of the online community members in promoting or fighting self-harm. In contrast to other social media networks, such as Twitter, YouTube does not restrict the length of video comments to short charters. This allows various individuals to debate and voice their thoughts on various issues surrounding self-harm via video comments. Thus, this research investigation used VADER and calculated the sentiment score for each video comment. Accordingly, our study analysed the rate of sentiments expressed in the topics obtained from the discussions surrounding videos about self-harm on YouTube. The diagram in Figure 4.9 depicts the percentage score of sentiments across each topic. While red represents negative sentiments, green and yellow indicate positive and neutral sentiments of the analysed comments.

Our analysis found conflicting viewpoints on various topics discussed by the video commentators. Many commentators expressed negative sentiments on the second and third topics. Numerous reasons could be the underpinning factors that result in high negative sentiments on both topics. Disclosing and sharing self-harm-related experi-

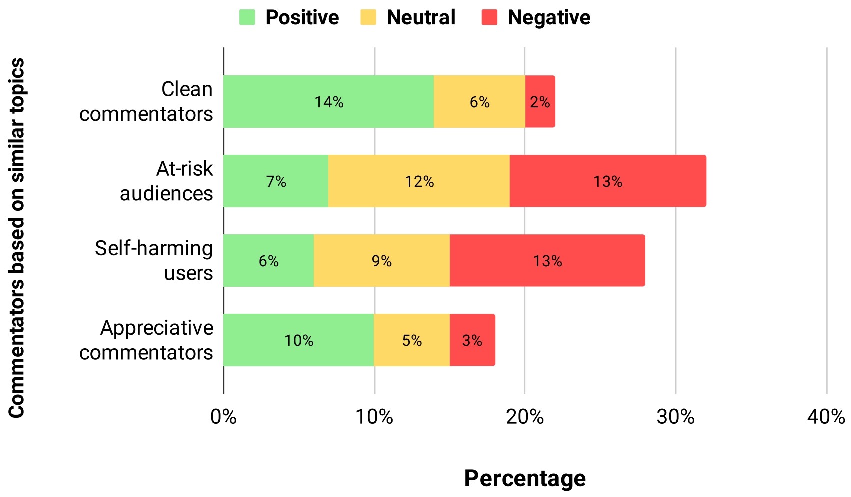


Figure 4.9: Topics' sentiment analysis

ences could be one of the contributing factors to high negative sentiments in the group topics. Even though some comments were neutral on topics two and three, only a low rate of the video responses represents positive sentiments.

Meanwhile, this analysis found many optimistic (positive) sentiments on topics one and four. This demonstrates how commentators exchanged positive posts to facilitate peer support and help-seeking. While appreciating the video content, the commentators from these topics used a pleasant tone to communicate or provide feedback to the video uploaders.

4.2.4 Detecting self-harm comments

As seen from the literature, some research studies from computer science-related disciplines used ML methods to analyse self-harm content (images) on social media. Hence, this study applied state-of-the-art ML techniques to detect self-harm comments from videos about self-harm on YouTube. Our analysis considered Supervised Learning (SL) algorithms (explained in Section 3.5.1) to classify the video responses. In addition to sufficient evidence from the existing studies, this study considered the SL method because the approach allows human experts to label the dataset to train and build an effective classifier.

Therefore, to complete the task of labelling the video responses, our study utilised a modified version of the coding criteria (see Table C.1) used by previous studies [158,159]. Two postgraduate researchers working on social media and mental health issues (self-harm and dementia) labelled the comments into one of the colour-coded categories described in Table C.1. The annotators were from the computer and information sciences department at the University of Strathclyde. Also, they have sound experience in labelling and categorising social media data. However, the annotators attended two meetings and discussed any issues associated with the data labelling. After completing the task, this study achieved a Kappa score of .81, indicating a stable agreement between the annotators [160].

Similarly, upon completing the annotation task, this study cleaned the set of comments and prepared the data for analysis. However, this section reported the results of different classifiers in detecting critical video responses that require urgent action from YouTube moderators. While proposing a machine learning model, the aim was to determine the accuracy of different classifiers in classifying video comments. Thus, this analysis began with annotating and preparing the data for the experiment. After extracting relevant features, a model was built and trained based on different classifiers to detect critical video responses.

Data cleaning and preparations

As described earlier in Section 4.2.1 this study randomly chose 10,000 comments from videos concerning self-harm on YouTube. Therefore, instead of creating a manual list of self-harm phrases, a lexicon of words from the set of video responses was created. The Term Frequency/Inverse Document Frequency TFIDF technique was used, and words that often appeared in the crisis and red classes and less in the amber and green categories were found (see Table C.1). This procedure establishes vocabulary used to differentiate among classes. This study analysed the *n-grams* of varying lengths (from one to three) and ranked the top 200 words using the TFIDF technique. Our research analysis explored these phrases and excluded non-self-harm related and duplicate words.

Even though evidence confirmed that self-harming individuals have no intent to end their lives [3], our analysis produced a set of terms indicating critical video responses showing possible intent of self-harm and suicide. Meanwhile, to achieve effective comments classification, a codebook (see Table C.1) was formed alongside researchers working on self-harm studies. In our previous study [156], two researchers categorised a sample of 2,000 comments into one of the classes described in Table C.1. During the analysis, the comments were not distributed equally for each class. Thus, to improve our analysis, this study increased the sample size to 10,000 comments, nearly 50% of the entire comments extracted from the examined videos. The same coding scheme was used in categorising the additional 8,000 comments.

Consequently, 2,923 and 2,302 video responses were classified in the *green* and *amber* classes. Also, the *crisis* and *red* group received up to 1,309 and 913 comments. This produced a total of 7,447 comments. Therefore, the remaining 2,553 that failed to fit into our coding scheme were categorised as ambiguous and ignored. *"I think it is happening :O."* is an example of ambiguous comment. In order to have a balance of sample representations, this study considered the minimum number (900) from each colour-coded classes, and produced a total of 3,600 video comments from the entire classes.

Moreover, the group of comments were also regrouped to form only two categories; *critical* and *non-critical* comments. The former consist of the *crisis and red comments* while the latter includes comments from the *green and amber* groups. Thus, making a total of 1,800 comments for each category. The idea was to improve our previous study [156] and build a model that could effectively detect comments that require further action from the YouTube content moderators. Thus, the analysis split the dataset into three. That is to say, about 70% of the dataset was used for training and 30% for testing and validation. Although there are different validation methods, this study used the *k-fold* cross-validation technique.

This technique ensures that the model can be generalised and work with real-world data, hence the purpose of its use in our analysis. However, this study used 5-fold cross-validation. The performance accuracy for each classification method was computed. The model with a high score of precision and recall was reported. The results are

Table 4.2: Comments classes proportion of linguistic cues

Class	**Users**	**verb**	**adverb**	**adjectives**
Amber	1,954	1.75	**2.33**	1.64
Crisis	1,024	**1.92**	1.43	1.83
Green	2,638	2.03	2.91	**2.97**
Red	913	**1.91**	1.21	1.25

shown in Table 4.3 and 4.4 for the binary and multi-class classification tasks.

Feature extractions

As explained earlier, this study utilised a set of video comments discussing self-harm on YouTube. This section explained the features used to train and evaluate machine classifiers for identifying critical responses associated with self-harm communications on YouTube. However, our experiment used the below features.

Linguistics features: research evidence confirmed that individuals' writing styles are connected with their psychological state of mind [161]. Therefore, the linguistic style of a comment could be tied to the commentator's mental and psychological states. Therefore, using the LIWC software,[5] this study utilised linguistics features from the examined comments to understand the differences between the language styles of different commentators. Accordingly, the ratio of the comments' distributions of verbs, adverbs, and adjectives was computed, as shown in Table 4.2. Thus, each class's different number of commentators and linguistic features were found.

Similarly, there was a difference in language cues between the *crisis* and *red* classes, with the former having a higher ratio of verbs, followed by adverbs and adjectives. Furthermore, this proportion is smaller than those in the *amber* and *green* categories, corresponding to research on word usage in suicidal messages [162]. Moreover, this analysis processed the comment texts using the *Word2Vec* technique that encodes (using numbers) individual words in a high-dimensional vector space representation [163]. The process encodes words with similar meanings that can be used synonymously in a given context as near vectors.

[5] https://liwc.wpengine.com/

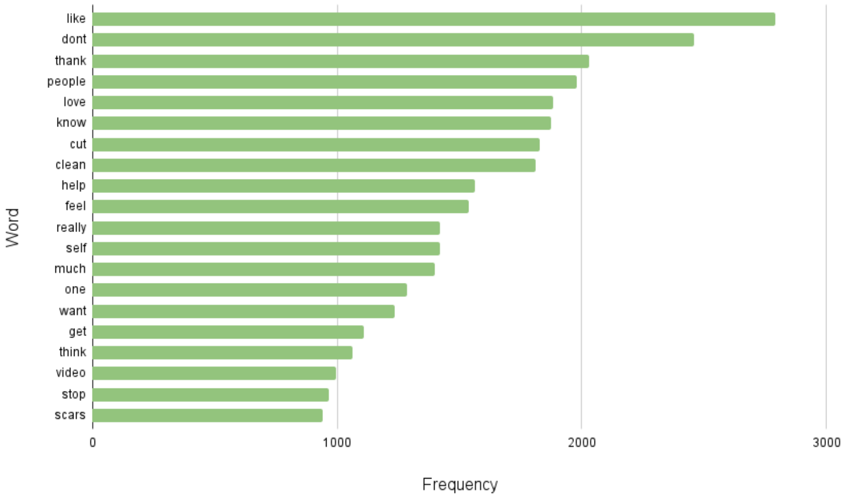


Figure 4.10: Top 20 terms in positive comments

Sentiment features: another essential feature that mirrors our identity is our vocabulary in textual writings [161]. Users' sentiments could be one of the unique ways self-harming individuals communicate their feelings and seek support online. While watching videos on YouTube, viewers can react to the videos by *liking* or *disliking* a video. To our knowledge, the platform did not have an emoji feature that users could use to express their emotions or opinion on a video. Therefore, the content of text comments on social media platforms like YouTube can describe how online audiences could accept or reject video responses.

This could be due to the terminology and language used in various video responses. For illustration, a ranked list (based on frequency) of terms considering positive and negative sentiment comments was formed. The idea was to understand the online audiences' influence on accepting or rejecting the examined comments. The sentiment feature affects the different categories of the video comments under investigation.

The Figures (4.10 and 4.11) illustrate the top twenty terms extracted from the positive and negative comments. The terms found in the positive comments with high-frequency show signs of commendation and recovery, such as *'thank', 'love'* and *'clean'*. Meanwhile, words such as *'cut'* and *'harm'* that appeared highly frequently in the negative comments indicate signs of DSH behaviours. In some negative comments,

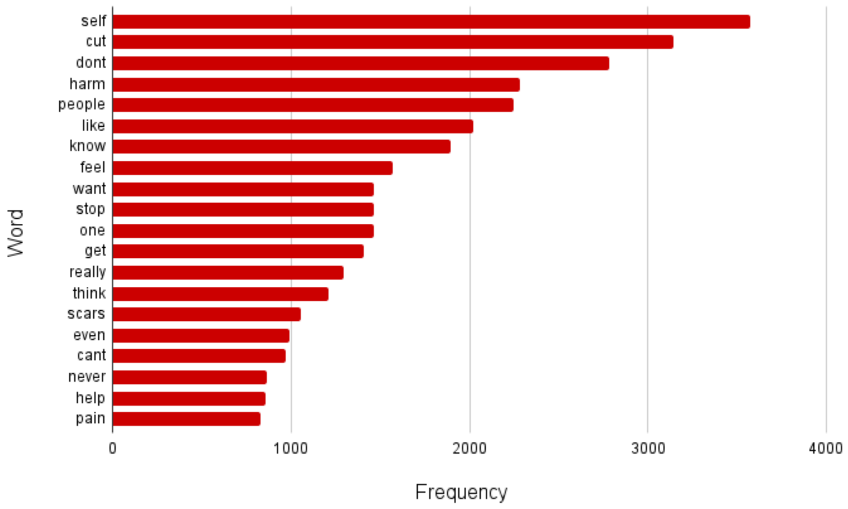


Figure 4.11: Top 20 terms in negative comments

words like *'help'* and *'pain'* could mean that the commentators voiced out to seek help from the online community. While both positive and negative responses could be acceptable or unacceptable to the online community, it is evident that negative responses may include words that could be triggering. However, this experiment found sentiment features crucial to training and building machine learning classifiers to detect critical comments requiring urgent attention.

Consequently, understanding commentators' sentiments through comments writings is an essential part of our experiment, as this feature can improve the classifiers' performance in detecting critical comments. Hence, this study utilised the VADER sentiment analysis model that surpassed several sentiment analysis techniques, including LIWC, SentiWordNet [127]. The graph in Figure 4.12 shows the rate of sentiments computed using VADER across the comments classes. Compared to positive sentiments, commentators in the amber, crisis and red categories expressed highly negative sentiments, especially the crisis group, which accounts for more than 60%. This result is identical to a prior study in which the sentiments of suicide notes were investigated [164].

Even though a high rate of positive sentiments (more than 50%) in the green comments category was found, the amber group also shows a notable amount of positive responses. The groups of comments exhibit different sentiment features essential in

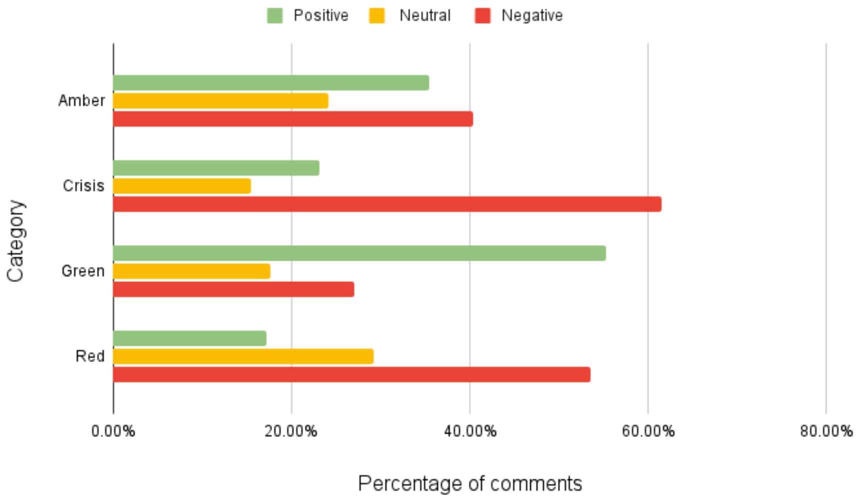


Figure 4.12: Comments sentiment analysis

training the classifiers. However, including the sentiment and linguistic features generated from our analysis, this study also used *state-of-the-art* approaches to extract classic text features. Therefore, each word was transformed and represented in a hundred-dimensional vector space using a pre-trained *word2vec*. The absolute representation of the feature is the summation of the vector of the entire representation of the words.

Critical comment detection

The preceding section discussed the sentiments of commentators responding to videos about self-harm on YouTube. While some of the examined comments revealed positive responses that could support self-harming users, the nature of the user's sentiments (high rates of negative responses) shows a possibility that some of the commentators could be at high risk of potential self-harm. Numerous video comments could cover critical responses requiring immediate attention due to the high number of user-generated comments on YouTube. Therefore, the results obtained from the sentiment analysis demonstrate that YouTube is a level playground for people battling with self-harm to voice out their feelings and seek support. With *state-of-the-art* machine learning techniques, it could be helpful to uncover and detect those hidden comments requiring urgent support or action from the YouTube content moderators.

While the comments under investigation are unstructured and complex to extract essential information, the examined comments contained information about the people that posted the comments, such as the commentator's identification (ID) number, the date and time the comment was posted, and the comment text. This study concentrated on the comment text, a subset of data that provides insight into commentators' views.

Experiment settings: this study analysed the set of comments from the *amber*, *green*, *crisis*, and *red* classes in two scenarios. The first scenario viewed the classification task as binary, while the second considered the issue a multi-class problem. In the first scenario, the *amber* and *green* classes were labelled as non-critical while the *crisis* and *red* groups represent the critical category. The psycholinguistic cues obtained from LIWC and sentimental features found in each category were utilised for training the model. Accordingly, these categories of video responses were represented in real numbers. For example, in a two-category attribute, supposing there is an n-category attribute; this implies that n numbers can be represented in a way that only one of those numbers or attributes in the category is represented as one and the remaining are zeros - (01, 10,). This study assumed that the labelled information of the video comments is C. Therefore, comments were represented as $C = \{c_1, c_2, c_3 ... c_n\}$ and n represent the actual number of comments.

Similarly, the set of comments was expressed in heterogeneous attributes conforming to m set of features. In the case of the examined video responses, the value of m is two, corresponding to psycho-linguistic and sentiment features. Therefore, the set features ware represented as $F = \{f_1, f_2, f_3, ..., f_m\}$ in which $fi \in \mathbb{R}^{ki}$ indicates the feature-length in the *ith* source and *ki* represent the size of the feature in *ki.* While Xi is a matrix representing the *ith* position, this analysis used $X = \{X_i \in \mathbb{R}^{n \times ki}\}_{i=1}^{m}$ to represent the set of data matrices. For individual comments, this experiment pulled a collection of features based on the sentiment and psycho-linguistic features to expand the standard features, as these features could not be attained by typical features but are capable of distinguishing between the green and red categories.

This experiment split the dataset into two. Seventy percent of the data was allocated to the training set, and thirty percent was used in testing and validation. The idea was to have a model that would effectively learn from the training set and make correct predictions without overfitting. Even though the process of labelling the video responses was tedious and time-consuming, it is believed that labelled data facilitate the model's learning and increase effective performance [165]. However, the prediction model was built using (1) the SVM with linear function, (2) KNN, and (3) RF-supervised learning techniques. Each data sample was defined as a vector of natural integers. This implies that any categorical attributes were converted to numeric data, and the results and performance evaluation of the classifiers are presented in Table 4.3 and 4.4.

Model's evaluation: in a supervised learning classification task, the essential metrics presented in equations 4.2, 4.3, 4.4 and 4.5 are used to evaluate a model's effectiveness and performance. The true positives *(tp)* occur when a prediction of a comment belongs to the correct class. In the case of true negatives *(tn)*, this happens when a comment that did not belong to the group was predicted, and indeed the comment is not a member of the class. Moreover, a false positive *(fp)* happens when a comment is predicted to be in a particular class, and the fact is that the comment could not be considered in the class. The false negative *(fn)* happens after predicting a comment as not a particular class member, while the comment is indeed an actual class member.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{4.2}$$

$$Precision = \frac{tp}{tp + fp} \tag{4.3}$$

$$Recall = \frac{tp}{tp + fn} \tag{4.4}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * tp}{2 * tp + fp + fn} \tag{4.5}$$

Table 4.3: Performance accuracy (in percentage) across the classifiers: first scenario

algorithm	**class**	**precision**	**recall**	**f1-score**
KNN	non-critical	0.68	0.67	0.67

	critical	0.73	0.61	0.72
SVMLinear	non-critical	**0.86**	0.81	0.81
	critical	**0.84**	0.53	0.79
RForest	non-critical	0.69	0.68	0.68
	critical	0.66	0.56	0.61

The experiment performed in this study was in binary and multi-class supervised learning scenarios. Table 4.3 illustrates the binary classification results, while Table 4.4 shows the findings from the four colour-coded classes. The performance of the three supervised learning algorithms used in the study is reported in both tables using the precision, recall and harmonic mean (f1) scores (in percentage). In this experiment, *false negative* and *false positive* costs are not equal because the objective is to classify comments that require critical attention (*true positive* comments) from the YouTube content moderators. In order to achieve this objective, this analysis focused on precision as the challenging task is identifying critical comments irrespective of the false negative values. However, from both scenarios, the performance of the Linear Support Vector Machine classifier (SVM-Linear) outweighs that of the remaining algorithms.

In the first scenario, while SVM-linear slightly outperformed KNN and RF, the later classifiers achieved similar performance in non-critical categories. That is to say, from Table 4.3, SVM-linear performed exceptionally well in classifying non-critical and critical categories with 86% and 84% precision compared to KNN and RF with similar performance accuracy. Furthermore, it was found that the KNN have better precision (73%) than RF in the critical class.

Table 4.4: Performance accuracy (in percentage) across the classifiers: second scenario

algorithm	**class**	**precision**	**recall**	**f1-score**
KNN	amber	0.58	0.49	0.53
	crisis	0.52	0.55	0.56
	green	0.65	0.58	0.61

	red	0.65	0.56	0.61
SVMLinear	amber	0.72	0.58	0.63
	crisis	0.74	0.63	0.68
	green	**0.77**	0.52	0.62
	red	**0.76**	0.69	0.72
RForest	amber	0.67	0.48	0.56
	crisis	0.65	0.52	0.58
	green	0.67	0.49	0.56
	red	0.64	0.57	0.50

Accordingly, in the second scenario, the SVM-linear demonstrates outstanding performance than both KNN and RF. Unlike the first scenario, RF performs slightly better than KNN. This algorithm achieved more than 70% accuracy across all classes. This analysis digs further into the core classes of interest - the crisis and red class (self-harm comments), the performance of all the classifiers changes as shown in Table 4.3 and 4.4. It was observed that there is a decrease in performance in both KNN and RF as opposed to SVM-linear. This could be due to difficulty classifying comments with self-harm (crisis) and suicide (red) features.

Moreover, the analysis contributes to the existing evidence concerning the nature of self-injurious content on social media. The sentiment and psycho-linguistic features extracted from the examined comments using VADER and LIWC were crucial in categorising critical video responses indicating self-harm signs. While using the same dataset in binary and multi-class settings, the sentiment and linguistic features used in the training of the models to detect self-injurious comments were practical in improving the performance. Although KNN and RF performed nearly at the same rate, this study observed that the former performed slightly better than the latter in the first scenario and vice versa in the second scenario. Hence, the results of the phase one study contribute extensively to the existing research [156] by improving the prediction performance. Thus, it is practical to understand that these algorithms, especially the

SVMLinear, demonstrate good performance on the analysed comments.

4.3 Summary

In this chapter, this doctoral study explained the research conducted on YouTube social media concerning self-harm video responses. The chapter discussed our findings of the sources of videos presenting self-harm information on YouTube. The study uncovered different topics viewers discussed through commenting on the examined videos. From the identified topics, the study analysed commentators' opinions and the accessibility of the analysed videos to young viewers - individuals under eighteen years. Similarly, this chapter described the performance of different classifiers that classified the video comments into colour-coded groups according to severity level.

Meanwhile, user-generated content on social media migrates from one platform to another. Self-injurious content such as short videos about self-harm on YouTube could be found on Twitter. Although support organisations like Samaritans provide support for self-injurers on YouTube by sharing videos, little is known about the sources of information providing support on social networks like Twitter. Also, users' attitudes and discussions surrounding self-harm are challenging to explore on a social network that allows a maximum of 280 limited character length. Consequently, the next chapter discussed the findings of the phase II study that examined a set of organisations supporting people with mental health issues (especially self-harm) on Twitter. to understand the impact of online social support for self-harming people.

Chapter 5

Phase II: Understanding self-harm discussions on Twitter and assessing the strategy of support handles

5.1 Introduction

This dissertation explained the phase one research study using data from YouTube social media in the previous chapter. The data examined from the previous study shows the connection between YouTube and Twitter social networks in disseminating information about self-harm. Various sources of videos discussing intentional self-injury can be found on Twitter. While self-injurious content could be shared between the two platforms, existing research shows the effect of Twitter, such as celebrity influence and the exchange of peer support.

Therefore, this chapter describes the findings of the second phase study performed on the Twitter social network. In particular, the study examined the strategy used by some organisations offering online social support through sharing information regarding self-harm and other mental health issues. The goal was to understand those organisations' impact in promoting self-harm awareness and prevention to inform best practices.

5.2 Results

The previous chapter discussed the findings of the phase one study. This chapter explained the results of the phase two study.

5.2.1 Self-harm discussions on Twitter

The current technological advancements and interactive social networking tools increase online social interactions among people from different backgrounds. Online social connections like Twitter allows people to post information (tweets) about their social and political interest or activities. Also, the mixture of the online networking platform facilitates sharing of health or mental health-related issues to increase awareness and improve positive well-being. However, existing studies demonstrate that Twitter is an online public space where self-harming people voice out their opinions.

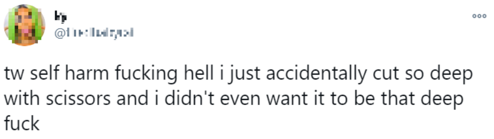
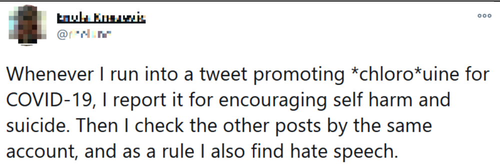
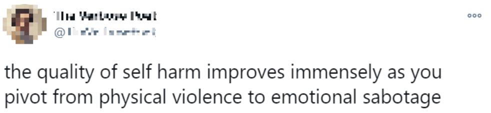
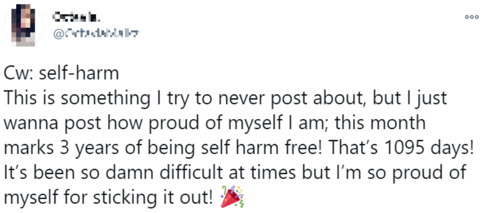
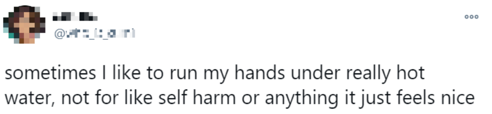
As explained in Section 3.4.3 of Chapter 3, this study extracted tweets using the *#selfharm* hashtag. The tweets were processed and prepared for analysis. Similar to the topic analysis performed in Section 4.2.2 of Chapter 4, cleaned tweets were considered and the LDA algorithm was used to uncover hidden themes discussing deliberate self-harm. Meanwhile, unlike YouTube video comments, tweets are short character lengths. Hence, it is essential to use trigrams (phrases that have only three words) to understand the context of the discussions and improve topic interpretability.

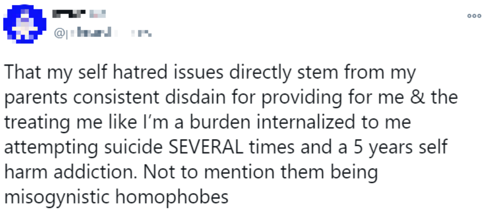
Consequently, this study found the different topics online users discussed regarding self-harm on Twitter. Even though different people joined the online discourse, this study grouped the discussants according to similar topics. The different groups of users who participated in the discussions are as follows;

- Inflicted
- Anti-self-harm
- Pro-self-harm

- Recovered
- At-risk
- Support seekers

Table 5.1: Some trigrams and corresponding themes from relevant tweets

Topic	**Trigrams**	**Example tweet**
inflicted	tw self harm, self harm scars, self harm suicide, self harm im, self inflicted injury, used self harm	 tw self harm fucking hell i just accidentally cut so deep with scissors and i didn't even want it to be that deep fuck
anti-self-harm	encouraging self harm, suicide self harm, people self harm, mentalhealth suicide awareness, suicide awareness selfharm, traumatic experience, selfharm addiction	 Whenever I run into a tweet promoting *chloro*uine for COVID-19, I report it for encouraging self harm and suicide. Then I check the other posts by the same account, and as a rule I also find hate speech.
pro-self-harm	people self harm	 the quality of self harm improves immensely as you pivot from physical violence to emotional sabotage
recovered	clean self harm, self harm free	 Cw: self-harm This is something I try to never post about, but I just wanna post how proud of myself I am; this month marks 3 years of being self harm free! That's 1095 days! It's been so damn difficult at times but I'm so proud of myself for sticking it out! 🎉
at-risk	like self harm	 sometimes I like to run my hands under really hot water, not for like self harm or anything it just feels nice

support seekers	selfharm addiction metoo, selfharm healing metoo, metoo forgiveness healing, autism selfharm asd	 [redacted] That my self hatred issues directly stem from my parents consistent disdain for providing for me & the treating me like I'm a burden internalized to me attempting suicide SEVERAL times and a 5 years self harm addiction. Not to mention them being misogynistic homophobes

For brevity, while comparing users' tweets, this study analysed the examined tweets and obtained the top *trigrams* to determine the meaning of the themes (see Table 5.1). The anti-self-harm group members fight against self-harm by reporting any Twitter account handle that encourages online members to self-harm or commits suicide. The anti-self harm tweeters actively engage in reaching out to self-harming users, advising them to stop self-harming, and educating them about the available helpful resources for support.

Furthermore, vulnerable tweeters seeking help voiced out concerns regarding developing an addiction to self-harm and recovery. Meanwhile, our analysis discovered fewer tweets from *pro-self-harm* and *at-risk* groups. Compared to other groups, the *pro-self-harm* and *at-risk* groups show significant differences as the duo exhibit a rare population that communicated their experiences on Twitter social networks. Moreover, some of the online members in the inflicted group frequently include trigger warnings **(tw)** in the tweet content while posting their self-harm experiences and the injuries *(scars)* that resulted from the self-harm behaviour. On the other hand, tweeters recuperating from self-harming behaviour communicated about becoming clean and free from intentional self-harm.

Accordingly, this study discovered that the *inflicted* category has the highest users, as illustrated in Figure 5.1. This category of tweeters accounted for up to 47% of the investigated users, showing that about half of the examined users mention their self-harm episodes and the resulting scars. *Anti-self-harm* users who discussed self-harm

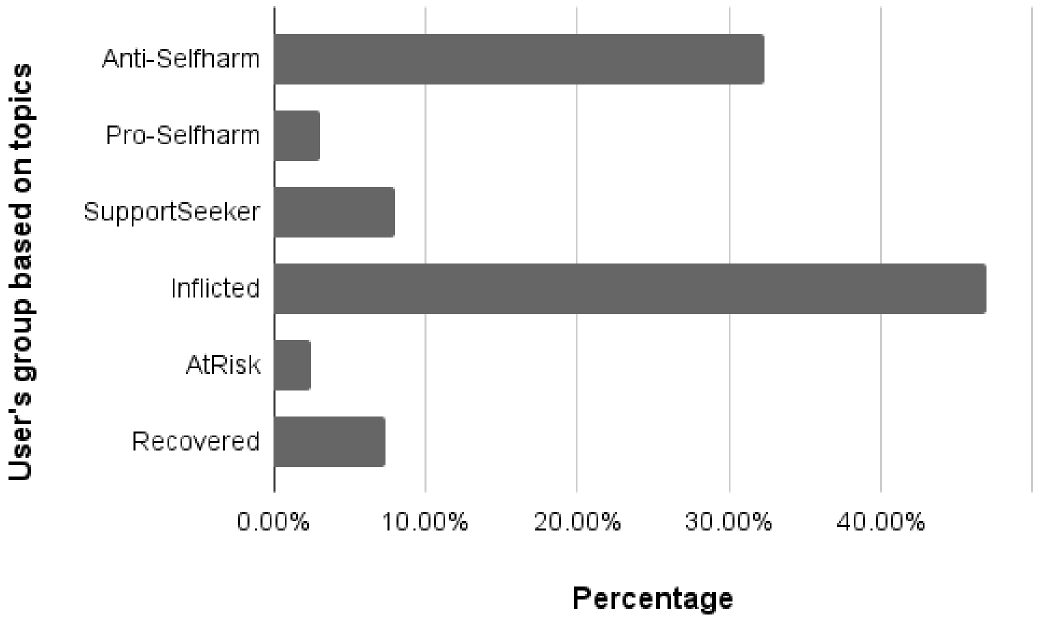


Figure 5.1: Categories of tweeters discussing deliberate self-harm using **#selfharm** hashtag

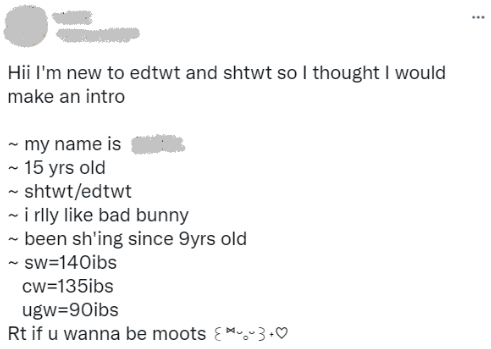
difficulties and consequences and strived for guidance and help from health experts obtained 32%. This rate decreased considerably to 3% for *pro-self-harm* and *at-risk users*, respectively. Consequently, this study found a concurrent rate distribution for individuals pursuing support and those who healed from self-harm.

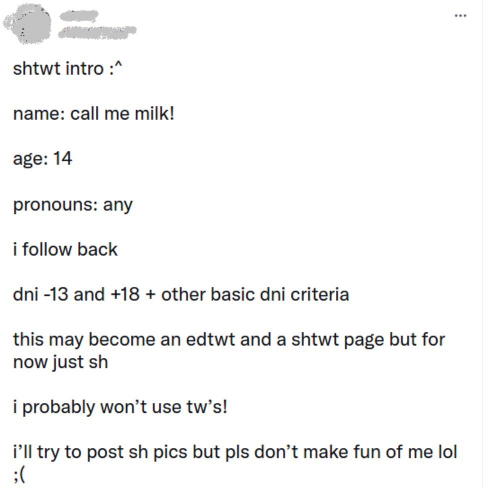
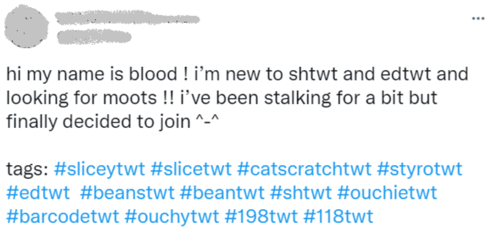
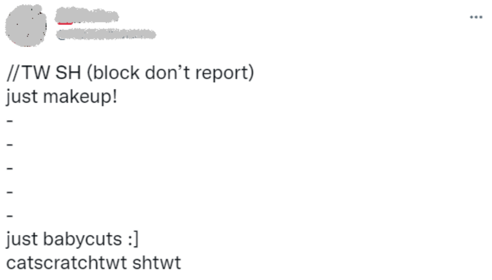
Furthermore, this study analysed the set of tweets retrieved using the *#shtwt* hashtag. Unlike the #selfharm hashtag, it was observed that users discussed different topics using the *#shtwt* hashtag. Thus, this analysis uncovered hidden themes discussed by users. Again, the experiment grouped online members according to similar themes, as shown in Table 5.2. As seen from the table, four major themes emerged as listed below.

- Self-harm and eating disorder
- Inflicted group
- Mutual followers
- Stop reporting

It was observed that some tweeters joined the discussions by disclosing information indicating that they newly started self-harming and also engaged with other mental health issues - eating disorders. Similar to a recent study [39], this shows that Twitter is an ideal platform for self-harm beginners, including individuals with early eating disorder habits. Moreover, similar topics (inflicted) were found in tweets posted using *#selfharm* and *#shtw* hashtags. The group of inflicted users revealed information concerning their self-injurious information. Although both users of *#selfharm* and *#shtw* hashtags tend to warn (using the trigger warning initials: tw) online members about the content they shared, it was observed that the inflicted users discussing self-harm using the *#shtwt* are more likely to incorporate images depicting self-harm and eating disorder in their discussions.

Table 5.2: Themes discussed using **#shtwt** hashtag

Topic	**Trigrams**	**Example tweet**
Self-harm and eating disorder	hi im new, im new shtwt, new edtwt shtwt, new just new, new shtwt edtwt hey im new, im new edtwt	 [redacted] Hii I'm new to edtwt and shtwt so I thought I would make an intro ~ my name is [redacted] ~ 15 yrs old ~ shtwt/edtwt ~ i rlly like bad bunny ~ been sh'ing since 9yrs old ~ sw=140ibs cw=135ibs ugw=90ibs Rt if u wanna be moots ɛ⋈˘◦˘ʒ⋆♡

Inflicted	tw self harm, ed sh twt, post sh pic	 [redacted] shtwt intro :^ name: call me milk! age: 14 pronouns: any i follow back dni -13 and +18 + other basic dni criteria this may become an edtwt and a shtwt page but for now just sh i probably won't use tw's! i'll try to post sh pics but pls don't make fun of me lol ;(
Mutual followers	im looking moots, rt like moots, like rt moots	 [redacted] hi my name is blood ! i'm new to shtwt and edtwt and looking for moots !! i've been stalking for a bit but finally decided to join ^-^ tags: #sliceytwt #slicetwt #catscratchtwt #styrotwt #edtwt #beanstwt #beantwt #shtwt #ouchietwt #barcodetwt #ouchytwt #198twt #118twt
Stop reporting	just new account, block dont report, old account got, old acc got, report just block	 [redacted] //TW SH (block don't report) just makeup! - - - - - just babycuts :] catscratchtwt shtwt

Accordingly, this study found a group of users tweeting about creating mutual social connections. While the literature shows that self-harming individuals participate in online-related activities to increase social connections and believe that they are not alone [28], this investigation found that some users sought to build online connections by following and retweeting self-harm content. Even though the literature emphasised that some online self-injurious content could be harmful and triggering [37], this analysis unveiled users posting harmful content to promote self-harm. Although they were being reported by online members fighting against self-harm, it was discovered that

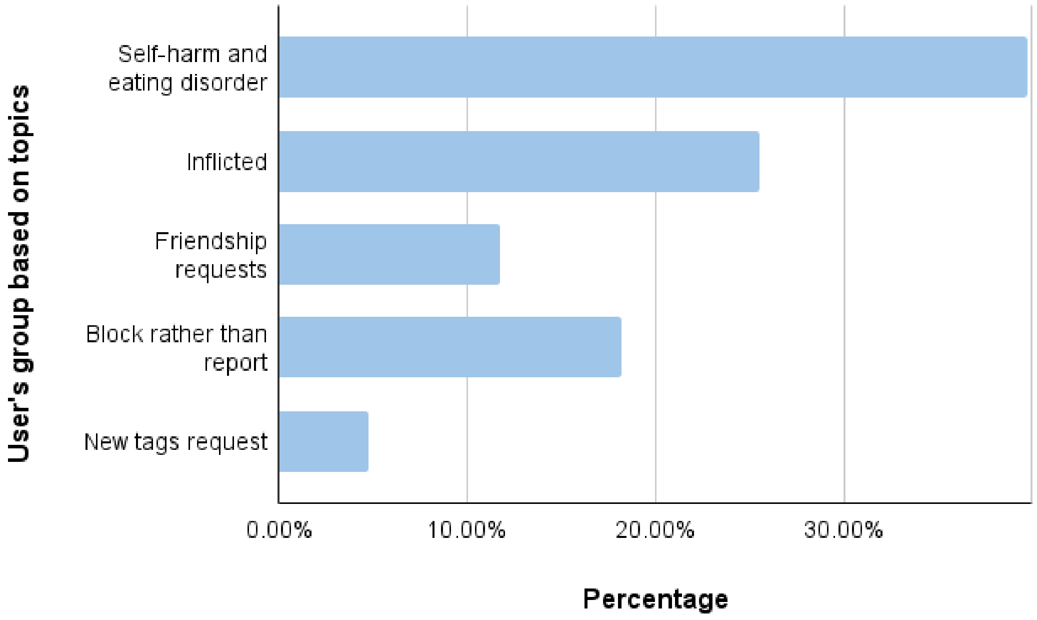


Figure 5.2: Categories of tweeters discussing deliberate self-harm using **#shtwt** hashtag

they engaged in the discussions pleading with the users to block them rather than reporting them directly to Twitter social networks. This has been found on several occasions.

Furthermore, this study analysed the discovered themes and found the most dominant topics discussed by the online members using the *#shtwt* hashtag. As seen in Figure 5.2, it was discovered that nearly 40% of the entire discussions were related to individuals that recently engaged in self-injurious behaviour and eating disorder habits. Like themes associated with the *#selfharm* hashtag, online users revealed their self-harming behaviours and experiences using the *#shstwt* hashtag. Meanwhile, there is a variation in the proportion of the inflicted theme between the two hashtags. Only 26% of the analysed topics from the *#shtwt* hashtag represented the inflicted tweeters.

On the other hand, themes relating to building social connections (friendship requests) account for only 12%, and nearly 5% of the discussed topics were concerned about seeking new self-injurious hashtags. Because Twitter's social network allows users to report any account sharing harmful content, some people tend to be reporting online accounts posting triggering content such as self-harming pictures. Consequently, this investigation found 18% of the discussed topics advising fellow users to block their online accounts instead of directly reporting them to the Twitter social network.

Thus, Twitter continuously monitors online users and suspends any account reported sharing dangerous self-harming pictures. Although Twitter limits tweets' length to only 280 characters, this study uncovered the hidden themes surrounding self-harm-related discussions on Twitter. Even though the tweet content is brief, it is evident that online users discussed crucial information, such as those promoting self-harm or fighting against self-harm and offering online social support.

5.2.2 Support handles strategies on Twitter

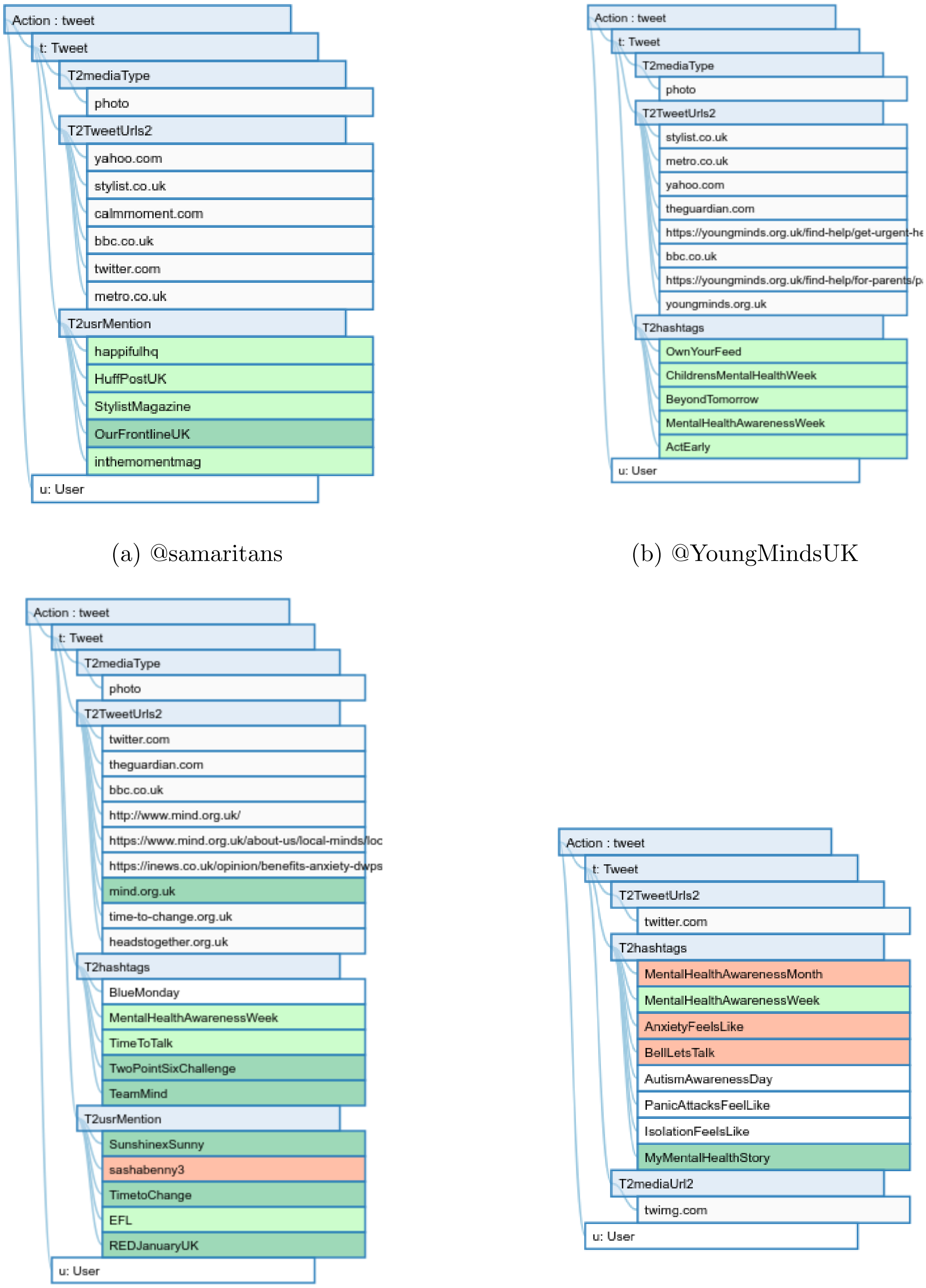
On Twitter, tweets posted by users have significantly impacted members of the online platform. Tweet's content contained essential information to understand the behaviour of the online account that posted the tweet. However, this study investigated the Twitter account handles of organisations supporting tweeters experiencing self-harm by analysing tweets crawled from those organisations. The investigation was conducted based on the tweets and replies posted by the organisations.

Tweeting behaviour of the support organisations

As discussed in Chapter 3, a novel reverse engineering approach proposed by [141] was used to understand the behaviour of the support organisations listed in Table 3.1. Thus, our analysis identified different strategies used by the support handles to disseminate information about self-harm and support vulnerable users. Although Figure 5.3 shows the underlying tweeting behaviour of the support accounts, the analysis focused on surface information, which is relevant to assessing the strategy of a Twitter account.

As seen from the figure, there are variations in the tweeting behaviour of *@YoungMindsUK*, *@Samaritans*, and *@MindCharity* support handles. The figures are colour coded. For example, green represents positive sentiments conveyed by the account handles, and red illustrates negative sentiments. Following this observation, a silent question could be raised as to why the *@depressionnote* account is more engaging than

Figure 5.3: Tweeting behaviour of support organisations.

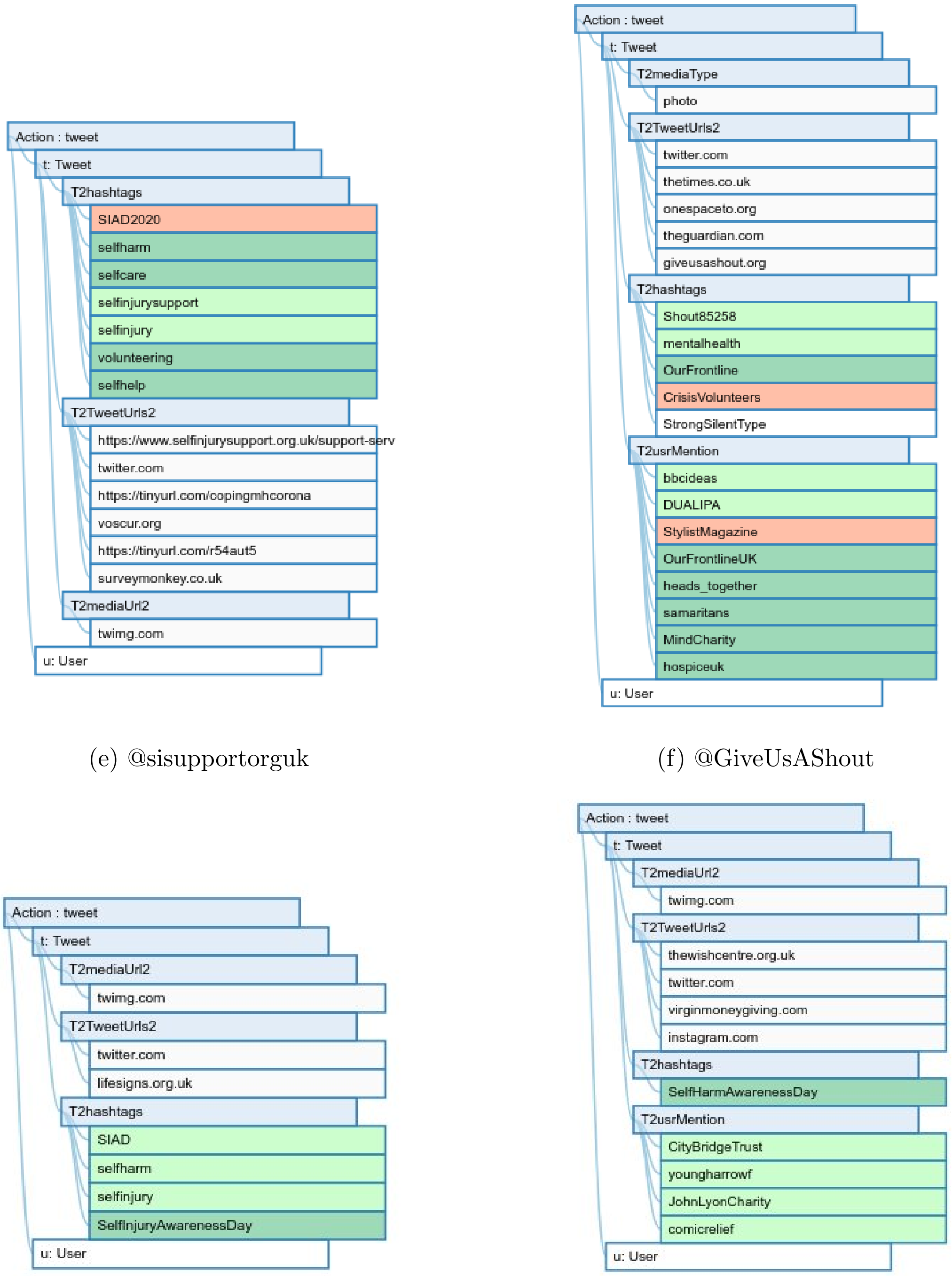


(a) @samaritans

(b) @YoungMindsUK

(c) @MindCharity

(d) @depressionnote

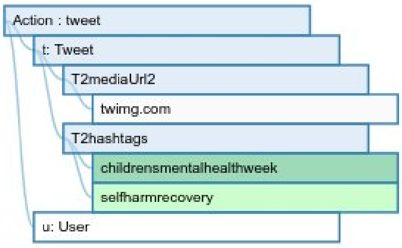


(e) @sisupportorguk

(f) @GiveUsAShout

(g) @LifeSIGNS

(h) @TheWISHCentre



(i) @selfharmUK

the other handles? Meanwhile, the *@depressionnote* utilises hashtags associated with anxiety. Similarly, the account's campaign strategy changes sentiments while using *#MentalHealthAwarenessWeek* and *#MentalHealthAwarenessMonth* hashtags.

In contrast to the *#MentalHealthAwarenessWeek* hashtag that elicits positive sentiments, the *#MentalHealthAwarenessMonth* produces negative emotions. The inclusion of hashtags associated with mental health attributes such as *#AnxietyFeelLike* and *#IsolationFeelLike* may contribute to social engagements between online members and the *@depressionnote* handle. Even though the result of this study contained the fundamentals of an account's tweeting pattern, our analysis focused on the outer information such as tweets and replies. The idea was to understand the strategies the examined handles are operating on Twitter effectively. Accordingly, as Figure 5.3 illustrates, the *@samaritans* organisation mostly tweets images and online links sourced from news organisations such as the BBC and Metro.co.uk and links online members to other helpful handles such as the *@happifulhq* and *@OurFrontlineUK* for further support. In contrast to the *@YoungMindsUK* and *@MindCharity* support handles, this study discovered that the @Samaritans account does not often use hashtags in its campaign strategies to increase social engagement with online users.

This study found the three most active organisations: (1) *@Samaritans*, (2) *@YoungMindsUK* and (3) *@MindCharity*. Even though the three active accounts consistently provide URLs, only the *@YoungMindsUK* and *@MindCharity* use the identical hashtag (*#MentalHealthAwarenessWeek*) to communicate information and raise awareness regarding mental health. Unlike the remaining organisations, *@MindCharity* incorporates the pattern of *@YoungMindsUK* and *@Samaritans*. Consequently, this could be the basis of the handle's growing popularity and followers engagement. *@GiveUsAShout's* strategies are unique among other support handles in that it links online members to *@samaritans* and *@MindCharity* support organisations. On the other hand, among the examined Twitter accounts, only the *@sisupportorguk* handle used the **#SIAD2020**, an abbreviated hashtag that represents self-injury injury awareness day which is 1st March every year[1], to communicate self-harm information to online users.

Like the *@sissupport* handle, the *@LifeSIGNS* account handle shares positive tweets containing self-harm and self-injury hashtags. Although the former communicated negatively with the **#SIAD2020** hashtag, the latter posted positive information using the **#SIAD** hashtag. Moreover, both *@LifeSIGNS* and the *@TheWISHCentre* handles exhibit similar self-harm campaign strategies. During self-injury awareness day, the *@LifeSIGNS* shared information about DSH using the *#selfinjuryAwarenesday* hashtag while the *@TheWISHCentre* account posted a campaign against self-harm the *#SelfHarmAwarenessDay*. Notably, both accounts expressed positive while raising awareness against intentional self-harm.

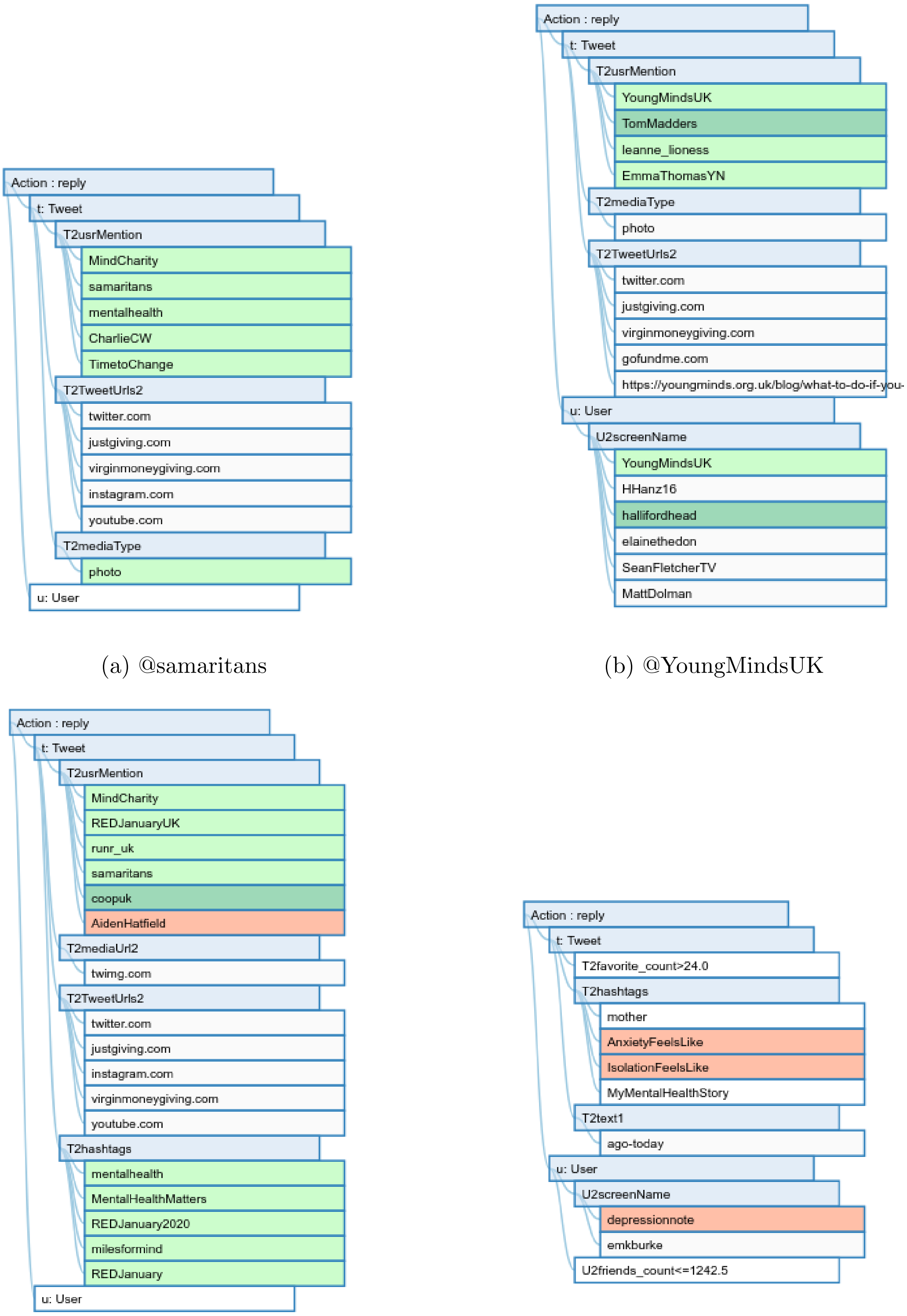
Consequently, out of the ten support handles examined in this study, only the *@stopselfaharm* account was found to be inactive; hence, its tweeting behaviour is not reported. However, one of the handles communication with the online community is responding to their tweet. This doctoral study analysed the replying pattern of the support handles on Twitter, as explained in the next section.

Support handles replying pattern

On Twitter, users can reply or respond to a tweet to participate in discussions. While support handles actively post tweets to increase mental health and self-harm awareness, followers of those organisations also post tweets or reply to support handles' tweets to join a campaign or discussion initiated by the organisation's account. Therefore, this study analysed the responses' behaviour patterns of the support accounts. The idea was to understand how the support organisations engaged (in terms of replies) with online members to provide support.

[1] http://www.lifesigns.org.uk/siad/

Figure 5.4: Support handles replying pattern

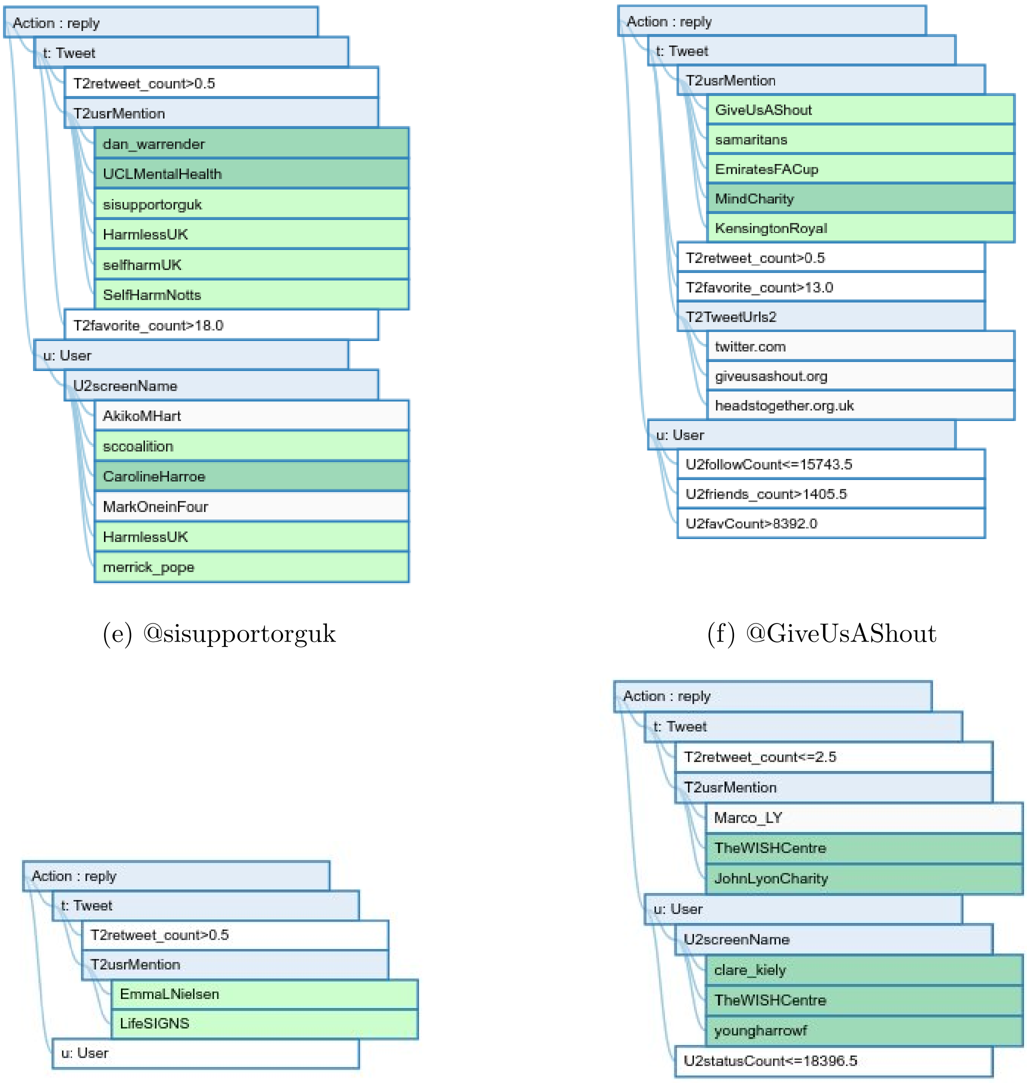


(a) @samaritans

(b) @YoungMindsUK

(c) @MindCharity

(d) @depressionnote



(e) @sisupportorguk

(f) @GiveUsAShout

(g) @LifeSIGNS

(h) @TheWISHCentre

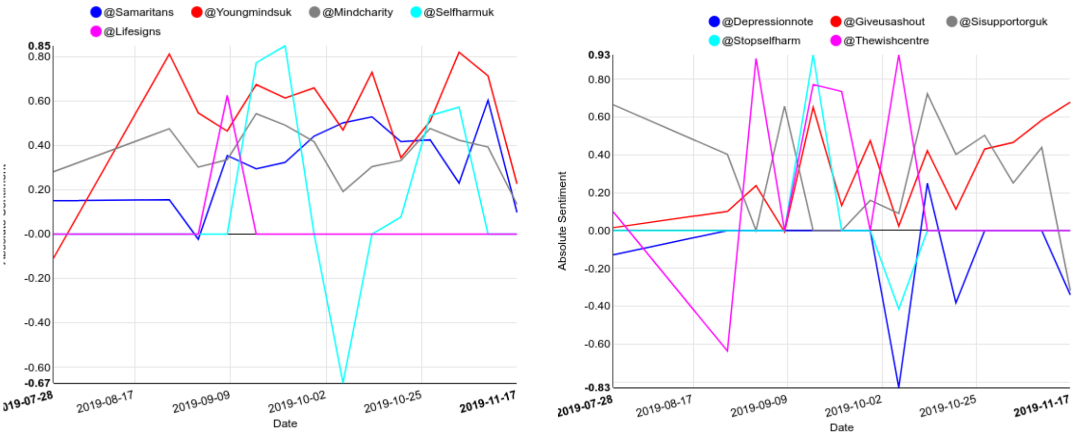
From our analysis, the *@samaritans* replied positively with tweets containing links to YouTube and Instagram social media platforms and mentioned other organisations like *@MindCharity*. Like the *@samaritans* handle, the *@YoungMindsUK* account often respond positively to tweets posted by online users—likewise, the *@MindCharity* and *@depressionnote* exhibit a similar pattern of responses associated with negative sentiments. While the former demonstrates responses to users' mentioned negative feelings, the latter replies to tweets expressing negative sentiments posted using the *#AnxietyfeelsLike* and *#IsolationFeelsLike* hashtags. Based on the analysed tweets, this could be attributed to the negative feelings associated with the COVID-19 self-isolation guidelines imposed by many nations to prevent the spread of the COVID-19 virus.

Moreover, similar to the *@GiveUsAShout*, the *@sisupportorguk* organisation demonstrate a unique replying strategy. The support organisation responded to discussions that mentioned *@HarmlessUk, @selfharmUK,* and *@SelfHarmNotts* support accounts combating self-harm. Furthermore, the *@GiveUsAShout* replied to tweets that involved the *@samaritans* and *@MindCharity*. Notably, this study observed that the responded information shared by most of the accounts showed positive sentiments. On the other hand, this study discovered that the replying patterns of the examined support accounts found @selfharmUK and @stopselfharm were significantly less active; therefore, no figure to illustrate the pattern they utilised in responding to tweets of the online users.

5.2.3 Self-harm support accounts and followers

Investigating how followers of the support accounts react to the information posted by the support handles on Twitter is paramount to understanding the influence of support organisations on vulnerable users. This doctoral study utilised the VADER rule-based sentiment analysis technique and examined the sentiments expressed by the users in communicating with the online community before and after the first index of the COVID-19 case in the UK. Thus, this study analysed the dynamic nature of followers' sentiments on information received from the support handles in two separate

Figure 5.5: Dynamic sentiments of followers towards the support accounts over time (before COVID-19)

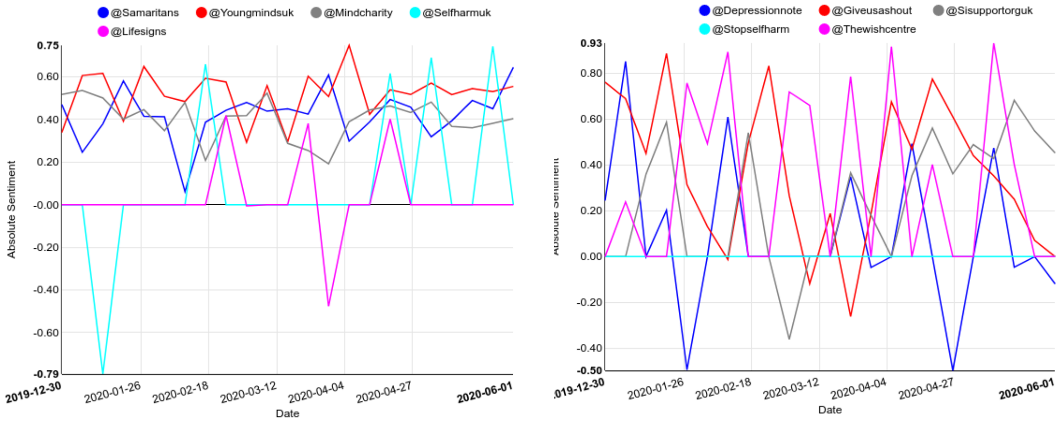


time periods - see Figure 5.5 and 5.6. Our experiment plotted the absolute sentiments of users in two graphs in order to improve readability. The figures displayed the sentiments (represented in the y-axis) of the online users over six months period. As seen in both figures, this study observed a few cases where followers expressed negative sentiments.

From the first graph on the left-hand side of Figure 5.6, it is observed that the *@samaritans, @LifeSIGNS* and *@MindCharity* received a high rate of positive responses from members of the online community. While the *@selfharmUK* and *YoungMinds* had instances of negative sentiments. The figure on the right-hand side shows several cases of negative sentiments, especially from the *@depressionotes* account handle. Even though *@TheWISHCentre* and *@stopselfharm* received negative views, both accounts, including *@GiveUsAShout* and *@sisupportorguk*, maintained positive sentiments within the period.

On the other hand, after the first case of COVID-19 in the UK, it was found that most support handles were socially active, reaching out to online members to lower anxiety and stress caused by the pandemic during the lockdown period. Looking at the first graph in Figure 5.6, most support accounts (except *@selfharmUk* and *@LifeSIGNS*) gained a significant amount of positive responses from the online users. Meanwhile, from the second line graph of the exact figure, it was observed that there were a few instances in which some accounts, especially the *@depressionotes*, received negative

Figure 5.6: Dynamic sentiments of followers towards the support accounts over time (during COVID-19)

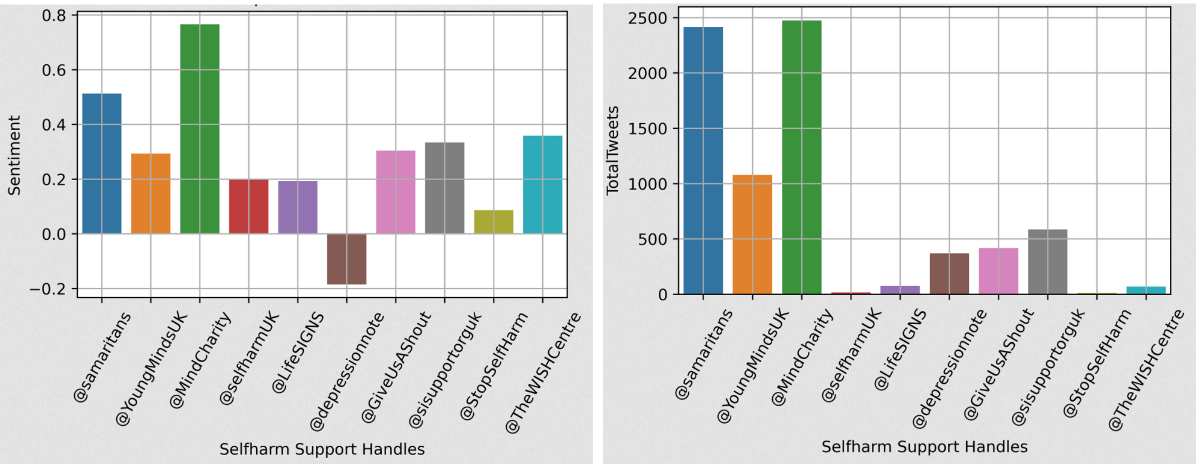


sentiments. As explained previously, this could be attributed to the account's tweeting behaviour involving anxiety-related hashtags. For example, *"Are you depressed?"* is a typical question heeded by helpful advice. However, users responded more positively to the remaining support handles.

Accordingly, this study examined the interactive features provided by the Twitter social network to promote social interactions among users. In other words, the interactive attributes between the support handles and users were examined. Figure 5.7 illustrates the online activity and sentiments expressed by the examined support handles. As seen in Figure 5.7b, this study found (1) *@samaritans* (2) *@YoungMindsUK* and (3) *@MindCharity* as the three most active handles information concerning deliberate self-harm. Even though the @*depressionnote, @GiveUsAShout* and *@sissupportorguk* were found to be functioning at a similar rate, it was observed that the *@StopSelfHarm. @SelfharmUk, @LifeSIGNs* and the *@TheWishCentre* were the less active handles on Twitter.

Moreover, this study uncovered the sentiments expressed by the support organisations in communicating or sharing self-harm related information with online members. In most cases, the fundamental role of the examined organisations in social networking sites, especially Twitter, is providing mental-health support and helpful information to self-harming individuals and users struggling with mental health issues. However,

Figure 5.7: Support handles engagement features

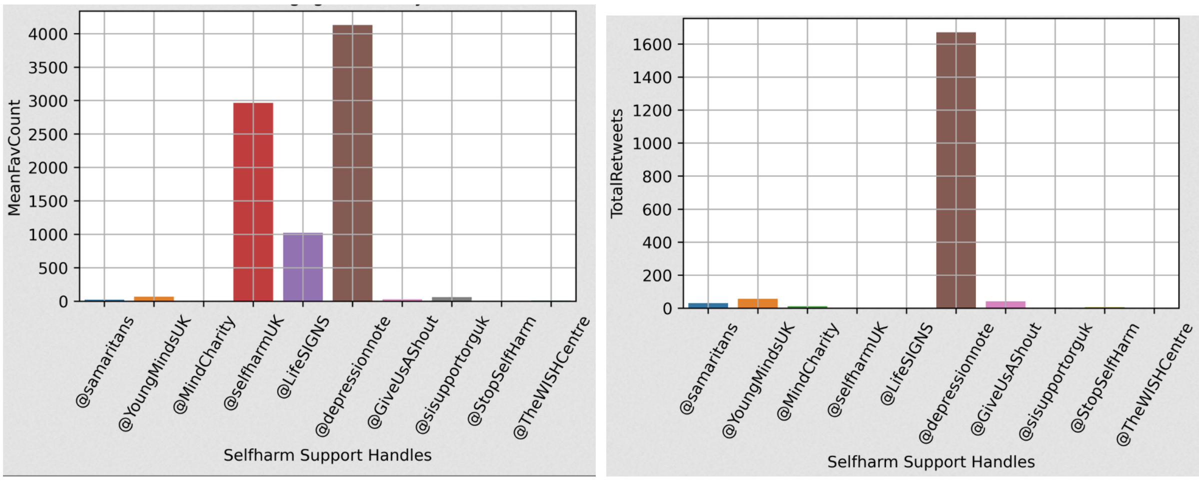


(a) Support accounts sentiments

(b) Online activity of the support handles

emotional support provided by the support handles has a significant impact on improving the mental well-being of online users seeking help. Looking at Figure 5.7a, except for the *@depressionnote* handle, most support accounts disseminated highly positive information, especially from the *@MindCharity* with online members.

Figure 5.8: Users engagement features



(a) User's likes

(b) User's retweets

Furthermore, how online users interacted with the support handle on Twitter (through user's *likes* and *retweets*) were investigated in this study. Even though it was discovered that the *@depressionnote* handle exhibits a negative tweeting pattern, it is surprising that the online members largely retweeted messages posted by the handle. In contrast to other support accounts, the *@depressionnote* handle received a significant number of retweets, as shown in Figure 5.8b.

Moreover, from Figure 5.8a, it was observed that the online members favourably rated (likes) information posted by the *@selfharmUK* and *@depressionnote* support organisations. Although numerous users liked tweets posted by the *@lifeSIGNS* account, this study discovered low *likes* ratings of the online users on information posted by the remaining handles. Consequently, this study demonstrates the impact of support organisations via Twitter social networks in combatting deliberate self-harm and improving positive well-being.

5.3 Summary

The previous chapter reported the findings of the phase one study of this doctoral thesis. This chapter discussed the findings of the phase two study conducted using the Twitter social network. The results of our research investigations show that Twitter is one of the platforms in which people participate in discussing intentional self-harm. While using real-world data from Twitter regarding self-harm, this dissertation uncovered the various themes associated with the discussions and identified and grouped people who participated. Similarly, the platform provides social support for self-harming individuals through support handles. The examined support accounts demonstrate positive patterns on Twitter to promote recovery and prevent self-harm.

This demonstrated the use of social media, especially Twitter, in fighting self-injurious behaviour and promoting positive well-being. Moreover, the chapter reported how support organisations interact with online members and vice versa to facilitate mental health support and increase self-harm awareness. However, the phase one and two studies of this research are associated with limitations. Therefore, the next chapter presents the research implications and limitations.

Chapter 6

Discussions

6.1 Introduction

The last two chapters reported the phase one and two research findings conducted on the YouTube and Twitter social networks. The present chapter starts with the problem statement addressed in the doctoral research. It then compares and contrasts the key findings of the two study phases. While the chapter explained the meaning and significance of the entire doctoral research findings, it also presents the research implications and limitations.

6.2 Background

The literature shows that our knowledge of digital content relating to self-harm is insufficient. It is crucial to analyse self-injurious content proliferating on social media spaces, particularly on video-sharing sites like YouTube and Twitter online networking space. Although the two platforms may have similar self-harm content, the sources of such content and people involved in sharing self-injurious resources online are not known [21, 40]. While self-harm posts on social media contain crucial information to understand users' attitudes or psychological states, it is vital to leverage the poster's sentiment and linguistic features to detect users needing critical support. Additionally, due to various forms of social support available on social networks for self-injurers [25, 33], understanding the reliability of sources of support and the pattern they used to supply information to online users is very challenging.

This research study tackles this challenge by analysing the strategy of organisations offering self-harm support on Twitter. However, in the study's discussion chapter, this thesis reported (in the next section) the key findings from the two platforms. The outcomes of this dissertation are essential in understanding the interplay between online social spaces and DSH. Additionally, the similarities and differences between the findings of these platforms will support our research community and stakeholders in developing policies and digital procedures and advising vulnerable online users on the sources of support available to prevent self-harm.

6.3 Comparison of the key findings

This research uses current *state-of-the-art* computing techniques to perform an in-depth analysis of self-harm related content posted on YouTube and Twitter. The objective was to (1) identify the key sources of the content, (2) understand the nature of the discussions surrounding the content, and (3) analyse the strategy of organisations supporting self-harming users on social media. One obvious connection between YouTube and Twitter is that both platforms support online social interactions worldwide.

While various online resources are available for self-harming people, such as special forums, the contemporary literature shows that researchers need to focus more on social media. Having discussed the findings of the YouTube and Twitter studies in Chapters 4 and 5, the following subsections will now discuss (in comparison) the critical findings of the phase one and two studies.

6.3.1 Sources of DSH videos on YouTube

In the case of YouTube, our investigation uncovered various groups of individuals participating in posting videos concerning self-harm. Different categories of sources involved in posting self-harm related videos on YouTube were found. In addition to professionals like medical doctors and psychiatrists, educational and medical institutions also participated in posting videos about self-harm on YouTube. Similarly, it was observed that the majority of the people that shared the examined videos were non-professionals, such as bloggers or social media influencers. Also, most of the analysed videos are accessible to children and individuals under eighteen years. However, examples of these sources can be found in Figure B.1 of the Appendix.

On the other hand, our investigations revealed that Twitter is a gateway to helpful information through key organisations fighting against non-suicidal self-injury that is widely accessed by self-harming people. Examples of those organisations are the Young Minds, Samaritans and Mind Charity. Moreover, unlike YouTube, it was observed that most of the support organisations are socially active on Twitter.

6.3.2 Common themes across both platforms

In phase one and two studies, themes surrounding self-harm discussions on YouTube and Twitter were examined. The idea was to understand what people communicate regarding non-suicidal self-injury on a larger scale. YouTube allows lengthy comments and responses to videos, while Twitter provides users with limited characters to create tweets. While both are textual content, users could detail their thoughts and views on YouTube instead of Twitter due to character length limitations. While analysing topics from both platforms, this study grouped users based on similar themes. However, this section discussed the common themes discovered from both platforms.

6.3.3 Recovered users

This research demonstrates that social media, particularly YouTube and Twitter networking platforms, are online spaces where people who recovered from self-harm communicate about their recovery process. The *clean commentators* and *recovered users* are examples of individuals that discussed their self-harm recovery. Although this study used an unequal dataset from the two platforms, it was discovered that most users commenting on YouTube videos reported that they recovered and cleaned from self-harm. Meanwhile, fewer tweeters communicate that they are free from self-harm. Even though this theme is typically common on both platforms, this study demonstrates that self-harm recovery-related information is more shared on YouTube than on Twitter.

Contrary to existing studies indicating that social media could normalise or reinforce self-harm [26, 76], this study shows that helpful information concerning self-harm recovery proliferated on YouTube and Twitter and could increase self-harm prevention on digital social networks. Consequently, these findings indicate that the success stories or information regarding self-harm recovery shared by the users could motivate other members of the online community that are self-harming users to seek help. On the other hand, some content of the examined platforms shows the presence of vulnerable users at-risk of self-harm.

6.3.4 At-risks users

In past studies, there are speculations that vulnerable people at risk of self-harm access social media. The investigation conducted by this study on YouTube and Twitter datasets revealed users prone to self-harm behaviour. Thus, the findings validate the claim that self-harm related communications on online social spaces could inform critical users at risk of potential injury. The topic analysis shows that more at-risk users on YouTube were twofold those on Twitter though the examined data is not equal in both cases. However, the findings of this doctoral study indicate that the discussions linking to at-risk individuals could be a unique opportunity for social media teams to connect them with appropriate support available online.

6.3.5 Self-harming users

The phase one and two analysis illustrates how self-harming users communicated about their behaviour and experiences. Similar to the existing research [31], this demonstrates that social media is an online platform through which self-injurers establish social connections and voice out their behaviour. In contrast to YouTube, self-harm related discussions on Twitter revealed a few themes promoting self-harm (pro-self-harm) in which online members encourage people to self-injure. Other topics found on Twitter that were not found in the YouTube study are anti-self-harm and support seekers. The former are those tweeting and fighting against self-harm behaviour, while the latter is a group of users engaged in help-seeking discussions.

The anti-self-harm group posts information that offers social support to users struggling with self-harm. Thus, this finding challenges the general public's perception, including the mass media, of the claim that self-harm content on social media is entirely harmful, as helpful content is overwhelmed on both platforms, especially Twitter.

6.3.6 Common sentiments

In this doctoral study, the opinions of people communicating deliberate self-harm on YouTube and Twitter were examined using the VADER sentiment analysis technique. Although both platforms have textual data suitable to analyse sentiment, this study acknowledged that the two platforms have different character lengths to support users interactions and communications. Intuitively, YouTube allows users to comment on a video with not more than 10,000 characters instead of Twitter, with only a 280 character limit. Therefore, this allows YouTube users to express their feelings in detail while tweeters only briefly communicate their thoughts using fewer characters. However, the sentiment analysis results from both platforms demonstrate a mixed-views from the online members discussing self-harm.

Positive opinions

This study performed sentiment analysis of the video responses (comments) and grouped commentators according to common themes in the phase one study. Most video commentators demonstrate positive attitudes towards online members, especially those (1) who appreciate the video content and thank the uploader and (2) those who were discussing their recovery process to support others. Similarly, the second phase study found a significant rate of positive opinions from Twitter users. Thus, the finding indicates the positive views of online members towards self-harming individuals. Self-harmers tend to suffer from a terrible state of mind resulting from anger; therefore, the observed positive attitudes toward self-harming individuals could help regulate emotions, increase support and potentially prevent self-harm.

Negative opinions

Although this study found a high rate of positive views from the online members across both examined platforms, it was observed that some users opinionated negatively. In the phase one study, at-risk audiences and self-harming groups of commentators shared many negative responses. This could be due to disclosing traumatic experiences and difficult feelings associated with their self-harming behaviour. Consequently, these findings demonstrate that social media is one of the critical places to look for self-harming individuals as the majority of them tend to isolate themselves and do not attend clinical support.

In other words, the literature shows that the problem of self-harm can be seen as a *'tip of an iceberg'* because only a few of the people who are self-harming are seeking medical support [14,55]. Thus, the findings of the sentiment analysis uncovered several users discussing self-harm related issues and experiences, especially on YouTube. Therefore, medical experts should consider using social media in mitigating self-harm by reaching out to wider audiences with helpful information.

6.3.7 Detecting critical responses

As discussed in chapter 3, this study retrieved a set of videos about intentional self-harm from YouTube. In order to create a model to detect critical responses, this doctoral study extracted comments from the retrieved videos. The video responses were cleaned and annotated using a colour-coded scheme representing the comment's criticality. However, the comments were analysed using state-of-the-art sentiment and linguistic analysis techniques (VADER and LIWC). While the machine learning model was created using different supervised learning algorithms, it achieved a significant performance accuracy, especially with linear support vector machines.

Although the model was built in two different scenarios; (1) binary class and (2) multi-class settings, it was found that the former achieved a reasonable rate of accuracy, and this is similar to previous research on the Flickr dataset [31]. Thus, the findings illustrate that the binary model automatically detects critical video responses requiring further moderation from the media team. Even though the multi-class model performance accuracy is not equal to the binary class model, the tradeoffs between the duo models require psychological and clinical investigations.

6.3.8 The strategy of self-harm support organisations on social media

In chapter 5, this doctoral thesis reports the findings of the phase two study that examined the strategy of support handles providing mental health support, especially for people struggling with self-harm. As a first analysis of its kind, this study examined the pattern of support handles supporting online individuals that are self-harming. In other words, this study investigated a set of organisations operating on Twitter that the NHS UK recommends to support and provide helpful information to self-harming people. Among these organisations, this study found the three most active support accounts; (1) *@YoungMindsUK,* (2) *@Samaritans* and (3) *@MindCharity* that are operating on Twitter. Although the remaining support handles were less active, the analysis conducted in this study discovered that most of the organisations shared self-harm-related information to increase awareness and support the online community, especially users who are self-harming.

Consequently, this finding demonstrates the positive use of social media as most of the analysed support handles engaged in fighting self-harm behaviour. The outcome of this study indicates that Twitter is being used as an online medium to facilitate digital campaigns against self-injurious behaviour. Additionally, the examined organisations exhibit different tweeting strategies while posting information about self-harm on Twitter. Although some of the organisations used links representing international news media sites and other social networking platforms like Instagram and YouTube, it was discovered that most of the support handles tend to share positive information with online users. Therefore, the perceived positive information could help regulate self-harming users' feelings and improve positive well-being. Despite the ongoing arguments or belief that social media serve as an online space that promotes deliberate self-harming behaviours that could lead to suicide, the study's findings demonstrate the positive use of online social platforms like Twitter to prevent self-harm and increase awareness.

6.3.9 Users reactions to support organisations

The literature emphasises that peer support in digital social spaces facilitates recovery [33]. Also, a survey study conducted on e-platform of self-harming people demonstrates that peer support plays a significant role in mitigating their social isolation [42]. Another analysis on Twitter informed us that self-injurers could obtain support from others [33]. However, Twitter provides different ways for users to interact with online members. Hence, this study investigated support organisations' strategies for fighting against DSH by raising awareness and offering online social support on Twitter. The goal was to understand the influence of those organisations on online users and inform best practices to support decision-making.

In addition to understanding the tweeting behaviour of the support organisations, the second phase study examined how the online users (followers of the support handles) engaged and perceived information with the support organisations. Although the concerned support handles posted an unequal number of tweets over time, it was observed that they remained socially active by interacting with online users on Twitter. Consequently, the analysis of this study found positive attitudes towards the information shared by the support handles, indicating that most of the examined users found the support organisations helpful in sharing positive information to support vulnerable users.

Accordingly, another essential attribute of the Twitter social network that denotes users' indirect feedback is the *'like'* interactive feature. Even though the analysis shows that the *@MindCharity, @Samaritans,* and *@YoungMindsUk* were the most active organisations on Twitter, it was discovered that the online users favourably rated most of the information tweeted by the *@SelfharmUk, @LifeSIGNS* and *@depressionnote* support accounts.

6.4 Implications

The literature shows the benefits and risks of sharing self-harm related content on social media. Among the most reported benefits are (1) relief from social isolation, (2) encouraging self-harm recovery, and (3) peer support [21, 33]. Interacting with online peers and disclosing their behaviour and experiences is one of the most noted advantages of online networking tools among self-harming people. Accordingly, this has also been confirmed by the analysis conducted in this research. However, several studies stressed the impact of social media platforms on self-harming people. As discussed in Chapter 2, the literature acknowledged the positive and negative effects of social media regarding self-harm. Some studies urged that experts from mental health areas and carers need to assess the online social behaviour of self-harming individuals [37, 111]. The research study conducted by [166] established functional criteria for assessing the nature and extent of online activities regarding self-harm behaviour, particularly in clinical contexts.

On the other hand, the quality of support available to social media users that are self-harming may directly impact their likelihood of seeking medical assistance. According to a previous research investigation, the YouTube and Twitter social networks can support self-harming individuals [33, 37]. From the examined dataset, YouTube users frequently watch and rate videos regarding self-harm information appreciatively. Even though the data does not contain the uploader's demographic information, the analysis shows that most of the video uploaders were females. This corresponds to previous investigations concerning self-injurious online related engagements [76, 105, 167] and conflicts with a study that found little or no difference in gender self-injurers [13].

However, consistent with cultivation theory [168], the study's findings indicate that frequently accessing self-harm videos on YouTube could fortify self-harm. Intuitively, the view counts and how the videos were viewed could be particularly relevant to this theory, implying that people who recognise and react to the information conveyed in the video may have received notable self-injurious reinforcement [48].

Moreover, like a contemporary analysis on Instagram [45], this study discovered that the analysed social platforms could help self-harm users by providing access to reliable information shared by professionals and support organisations. Consequently, the study's findings aligned with previous works concerning the nature of self-injurious content on social media [48], particularly on YouTube.

The outcome of the phase one study builds on the existing evidence of the views and opinions of people viewing deliberate self-harm videos on YouTube. Although the result confirmed previous research findings demonstrating the usefulness and dangerous nature of online posts related to self-injury [37], the analysis provides insights into user-generated content promoting self-harm recovery. Furthermore, the experiment conducted in the study contributes to the existing knowledge on using current *state-of-the-art* supervised machine learning techniques to detect critical self-injurious content.

In other words, some attributes of this doctoral study raise concerns about the currently accepted standard of moderating self-injurious content on social media. A closer examination of the self-injurer's social media posts revealed patterns that show early signs of self-harming behaviours. Using supervised ML methods and probabilistic topic modelling techniques results in identifying users who reported their mental health state and communicating about the nature of their self-harm behaviour to the platform's community. Hence, these techniques are essential in identifying vulnerable users at risk of potential harm.

Therefore, *what does this imply for social media and computing research that focuses on self-harm detection?* Thus, the mechanisms employed by our research community for detecting, predicting, and characterising self-harm related content are not very intelligent enough to automatically identify user accounts of people who could be at risk of self-harm or encouraging self-injurious behaviours that could lead to suicide. Although this is not to argue that the existing approaches are not practical and require further investigations, it implies that the examined platforms should consider a proactive approach in detecting vulnerable users and triggering content promoting self-harm—this can help reduce the negative impact of social media on self-harming individuals.

Meanwhile, the available informal sources of support on social networks could provide opportunities to reach out to self-harming individuals to reduce *stigma* and increase significant involvement with mental health professionals. Regarding the issue around young self-harming individuals (including children) who utilise online spaces to discuss and view self-injurious content is the possibility of a *'normalising effect'* [21]. Therefore, social media usage among those individuals may influence their views of intentional self-harm, resulting in the continued adoption of harmful coping practices and the possibility of initiating different acts of self-harming behaviours.

The second phase of this doctoral study analysed the Twitter accounts of some organisations supporting self-harming people. The sentiment analysis of the support handles and online users (followers) demonstrates that the Twitter online community recognised the support and awareness provided by the support organisations as positive. Additionally, this research study contributes to previous works on using social media sites to communicate and share self-harm-related content [21, 36]. The research findings are helpful to our research community and relevant stakeholders by increasing our knowledge and understanding of the possible effect of discussing self-harm on Twitter. While it is critical to have open and informed discussions regarding self-harm, its representation on social media, and its potential consequences, this study offered practical insights into the nature of self-harm-related discussions on Twitter. This could support medical practitioners in strategising for adequate care and support. For example, they may use the support handles to reach out to self-harming individuals on Twitter and offer sufficient support.

Similarly, from the perspective of engagement, the behaviour of the examined support accounts showed critical areas for development. Compared to less active accounts, the most active handles increase self-harm awareness and share positive or inspirational messages to facilitate support among vulnerable users. Meanwhile, the opinions or views shared by the followers show valuable feedback to the support organisations. For example, the organisation operating the *@depressionnote* account on Twitter needs to review its campaign pattern in order to increase online user engagement and positive message exchange among users.

Furthermore, this study could benefit the support organisations by refining their social movement strategy on Twitter about mental health problems, especially self-harm.

Through working with medical experts, the support handles can play a significant role by informing the general public about the dangers and risks of self-harm in a digital social manner. Similarly, the second phase study reaffirms the importance of clinicians collaborating with support organisations to fight against self-harm and other mental health crises.

6.5 Limitations

Numerous limitations of the study were acknowledged. In the phase one study, this research study retrieved a set of videos about self-harm using five different search phrases. Due to the type of YouTube API used, the study is limited to fifty videos per query term. This set of videos did not represent the entire collection of videos presenting self-harm information on YouTube. Therefore, the findings of the phase one study cannot be generalised to all videos discussing self-harm on YouTube. Similarly, the analysed data do not have demographic information about the video commentators.

Thus, it is difficult to confirm if the users who commented on the examined videos are young people, even though existing research found that young individuals are the primary users of social media [20]. Moreover, in the phase one study, comments were randomly picked and manually annotated to one of the classes described in Table C.1. Consequently, the study utilised a limited number of the entire comments and built predictive models to detect critical responses. Similarly, because the nature of self-harm content differed between videos and comments, it will be critical for researchers with a medical background to analyse the quality of information presented in both cases.

Another limitation of the phase one study is that not all viewers responded to the examined videos because YouTube allows users to watch a video without necessarily commenting on it. Therefore, the analysis conducted in this study did not represent the entire audience (users) that view and comment on the analysed videos. Furthermore, there are some limitations associated with the second phase study.

The second phase study examined the nature of the discussions around self-harm on Twitter. It was unclear whether the users who participated in the discussions (especially those at-risk of self-harm) were children or young people, as the literature shows that most of the people who are self-harming are young individuals. Although existing evidence confirmed that females are self-harming at a higher rate than males, the analysed Twitter dataset lacks sufficient information to understand the gender differences of the users tweeting about DSH on Twitter.

Another limitation of the second phase study is that the analysis focused on the tweet's textual content rather than images or video associated with the tweets. Although not every self-harming user posts information or communicates about his or her self-harming behaviour on Twitter, it is acknowledged that the analysed user accounts do not represent the entire self-harming tweeters discussing self-harm on Twitter. As a result of this limitation, the findings of the phase two study can not be generalised to all Twitter users who are self-harming.

Additionally, the second phase study investigated the tweeting behaviour of some organisations providing online social support to self-harming users. Even though similar organisations offering the same services are operating on Twitter, this study is limited to only ten support handles recommended by the NHS in the UK. Therefore, the strategies used by the examined support accounts cannot be applied or generalised to the entire set of organisations reaching out to vulnerable self-injurers on Twitter. However, both phases analysed textual data posted in English; therefore, the data did not represent the entire record of video responses or tweets concerning self-harm because other languages on both platforms were not assessed.

Consequently, the experiment conducted in this doctoral study did not analyse the voices of self-harming users who used non-English language to communicate with online members. Furthermore, while this study used the LDA technique, one of the current *state-of-the-art* probabilistic topic analysis approaches, it is acknowledged that the probabilistic topic approach is associated with some limitations, such as choosing the number of topics that describes the entire text corpus and uncorrelated themes [126, 169]. Hence, online members posting about self-harm on the examined platforms may disagree with some identified or interpreted topics.

As explained in Chapter 3, the data collection procedure used keywords and self-injurious hashtags. One of the disadvantages of this procedure is that it could retrieve irrelevant data. An example of this issue was found when users used *#self-cutting* to communicate hairstyle instead of self-injurious behaviour. Additionally, in the case of Twitter, it was discovered that online members may have specified the hashtag used to retrieve data without writing any statements about self-harm.

Consequently, this is one of the broader constraints of utilising social media data for research purposes. Data sources such as surveys or interviews allow researchers to direct participants to respond in unique and appropriate ways regarding the research study. Meanwhile, the benefit of using social media data is that it allows online users to express information related to their self-harming behaviour that they might not feel comfortable discussing in a research survey or interview.

Chapter 7

Conclusions

7.1 Introduction

This chapter concludes by providing a concise description of how the study's research questions were addressed.

7.2 Research questions

7.2.1 Who uploads videos on YouTube concerning self-harm, and how are the videos rated?

From the analysis conducted in the phase one study, it was found that different groups of users participated in sharing videos related to self-harm on YouTube. The study discovered that professional individuals such as medical experts and academics were involved in uploading and disseminating videos about intentional self-harm. While most of the video presenters were female, it was observed that most of the analysed videos were from a group of non-professional YouTubers. Other sources of the examined videos were government and non-government organisations, self-harm support organisations and news media agencies.

Furthermore, most of the assessed videos were accessible to general audiences, including children and young people. Consequently, the findings suggest the need for medical experts to increase participation on YouTube by creating and sharing video content to support self-harm recovery and prevention. Also, it is essential to continue assessing videos posted by non-professionals to ensure online safety.

7.2.2 What are the discussions surrounding YouTube videos presenting information about self-harm?

The viewer's responses were analysed, and essential themes associated with the comments were uncovered. The video commentators discussed topics related to (1) self-harm behaviour, (2) self-harm recovery, (3) at-risk audiences and (4) appreciative users. Online users responded to the videos with information concerning their self-harming behaviour and the object or tool they used to self-injure. Accordingly, stakeholders, psychiatrists and other medical professionals must collaboratively work to control the effective use of social media by self-harming individuals.

Moreover, despite disclosing self-injurious information, some commentators empowered online members with awareness and tips on becoming clean from self-harming. The proportion of video comments on each identified topic was examined, and it was discovered that most of the video responses were centred around self-harm recovery discussions. Even though some discussions were related to commentators at risk of intentional self-harm, it was found that viewers discussed appreciating the information presented in the videos.

7.2.3 What sentiments do users express concerning videos about self-harm on YouTube?

Online audiences viewing videos related to DSH on YouTube opinionated differently. The analysis conducted in this research revealed users' positive and negative sentiments who responded to videos concerning DSH on YouTube. Intuitively, the audiences' opinions were analysed based on the identified themes. While online members responded more negatively in topics related to at-risk and self-injurious behaviour, viewers reacted very positively in themes regarding self-harm recovery and appreciative comments.

Overall, the commentators' sentiments show a significant positive response rate. Meanwhile, high negative sentiments could increase the risk of self-harm and possibly suicide. Therefore, the sentiment analysis demonstrates the positive views of users, which could be attributed to peer support. Hence, social media moderators should continuously investigate users' sentiments to ensure online safety.

7.2.4 How can we detect critical comments from DSH videos on YouTube?

Video commentators expressed their feelings which are attributed to their psychological well-being. In order to detect critical comments indicating that the commentator could be at risk of self-harm, this study manually annotated a set of video responses into different categories (colour-coded) representing the severity level of the comment. The idea was to employ supervised learning methods and build a classifier capable of detecting critical video comments. However, the experiment performed in this study incorporated the sentiment and linguistic features trained and built models capable of detecting comment responses showing signs of intentional self-harm.

Although the analysis was performed using (1) binary and (2) multiclass classification scenarios, it was discovered that both scenarios are effective in detecting responses indicating at-risk commentators. While building the model with different supervised learning algorithms: K-NN, SVM Linear, and Random Forest, it was observed that the model built with SVM Linear achieved a high-performance accuracy in detecting at-risk responses.

Thus, indicating that the algorithm effectively outperforms the other algorithms on the examined datasets. However, as technology continues to evolve, it is vital for social media researchers, designers, psychologists, and medical experts to collaborate on developing new interventions to combat the proliferation of harmful content on social media.

7.2.5 What are people discussing regarding intentional self-harm on Twitter?

Unlike YouTube, self-harm-related discussions on Twitter are limited to only 280 characters due to Twitter's restrictions. Although the platform allows people to post tweets using emojis, images, links and videos, this study focused on the textual content of the tweet in order to understand what tweeters discussed regarding intentional self-harm. As seen in Chapter 3, this study crawled tweets using different self-injurious hashtags and examined the hidden topics discussed in the retrieved datasets.

However, this study utilised the LDA technique to address this question and discovered the themes surrounding self-injurious discussions on Twitter. While analysing the topics tweeters discussed, people who participated in the discussions were grouped according to similar themes listed below.

- Inflicted users
- Pro-self-harm
- At-risk users
- Support seekers
- Anti-self-harm
- Recovered users

Given the above-identified topics discussed by a different group of users, it was found that the *inflicted* and *anti-self-harm* tweeters are the majority of groups that dominate the entire discussion. Subsequently, the recovered, pro-self-harm and at-risk tweeters were found with fewer topics. Therefore, unlike YouTube, there is a need for medical experts, psychologists, support organisations and mental health professionals to increase using Twitter to provide support to vulnerable users at risk of potential harm.

7.2.6 What strategies support organisations are operating on Twitter to facilitate and encourage self-harm recovery?

This research examined the behaviour of Twitter accounts of organisations offering support for self-harming users. As discussed in Chapter 3, this research considered the list of organisations (recommended by the NHS) providing online social support to self-injurers, specifically on Twitter. The data retrieved from the support accounts was at the beginning of the global pandemic (COVID-19). Hence the reason for analysing the support handles before and during the first index of COVID-19 in the UK. Consequently, it was discovered that the *@MindCharity, @Samaritans* and *@YoungMindsUk* were the three active accounts that frequently interact with the online community on Twitter.

Most of the investigated support handles exhibit similar strategies in reaching out to the target individuals and the general public to support positive well-being. Although the *@depressionnote* account was more engaging than the other handles, it was observed that some support organisations used a unique pattern in sharing information related to self-harm. For example, the @Samaritans integrates web links and images and refers users to other online support sources.

Moreover, the strategy used by most of the examined accounts involved (1) using a positive attitude while communicating with online members, (2) tweeting with news media and social network URLs, (3) incorporating unique hashtags to increase awareness, especially during special dates recognised globally, to campaign against self-harm or mental health-related issues. However, the findings suggest the need for the examined organisations, especially those that are less active on Twitter, to re-strategise their online campaign pattern.

Although media links could provide access to useful information, it will be crucial for support organisations to consider incorporating web links (in their tweeting strategy) from professional bodies like the NHS. Access to links from medical institutions could provide helpful information and improve organisational's best practices.

7.2.7 How do followers of the examined organisations opinionate on the information shared by the support accounts over time?

The experiments performed in the second phase study considered analysing the interactions between the support accounts and online users (followers). The online activities of the support handle with users were assessed in which they showed an effective rate of positive interactions. Although most of the support accounts communicated positively, it was observed that the *@depressionnote* handle slightly exhibits negative opinions.

On the other hand, the user's sentiments regarding the information posted by the support accounts were analysed, particularly before and during the COVID-19 pandemic. In both cases, the sentiment expressed by the online users on the information posted by the support organisations indicated a significant level of satisfaction.

Consequently, the analysis demonstrates a significant degree of followers' satisfaction or positive views regarding the information they accessed from the support organisations on Twitter. Therefore, those organisations must continue providing helpful information to the online community to increase mental health awareness and the dangers of self-harming behaviours.

7.3 Contributions of the study

Consistent with existing research involving social media and mental issues, this doctoral study utilised raw data retrieved from the YouTube and Twitter social networks. The contributions of this research work are fourfold. Firstly, this doctoral study found the various sources of self-harm content on social media. These include a group of professional users from academic and medical fields and non-professional individuals promoting and against self-injurious behaviour. Secondly, this research provides valuable contributions by analysing self-harm related discussions and views of online users on popular social media platforms using current *state-of-the-art* probabilistic techniques. Thus, the research uncovered several themes associated with self-harm discussions from different users. Notably, this study found a hidden topic linked to a particular group of users that could be at risk of self-harm.

Thirdly, this study proposed an automatic technique to detect critical responses from users at risk of potential self-harm using machine learning techniques. Meanwhile, various sources of support for self-injurers operating on Twitter were examined. Most of the support handles exhibit positive strategies in increasing self-harm awareness and promoting self-harm prevention. It was observed that most of the examined support accounts demonstrate positive social engagements with online users. However, while social media could be dangerous for people battling self-harm, this study demonstrates that it could also be a valuable online space for self-harming users to obtain helpful information to facilitate support and recovery.

7.4 Future work

Despite the significant contributions of this study, researching self-harm content on social media is in an infant stage because of social networks' dynamic and technological variations. Therefore, there are several crucial areas of improvement in this area of research. In the first phase study, the focus was on the video responses rather than the actual content of the videos, which could directly impact the target audiences. While analysing responses to videos about DSH is one of the strategies for determining the impact of such videos, there could be other relevant procedures.

Thus, future researchers in this area need to examine how videos related to self-harm on YouTube are perceived by self-harming individuals. Accordingly, an effort is required to explore the viewing attitudes of online users because it is vital to explore whether the first, subsequent, or repeated viewing of such video content could directly influence the target audiences. Moreover, the nature of self-harm varied across individuals and seeking clinical support was not commonly encouraged.

Therefore, there is a need for clinical experts, psychiatrists, and psychologists to examine the quality of information posted on social media platforms, especially by those claiming to be medical professionals. While users discussed topics related to self-harm recovery, and most of the expressed sentiments were positive, social media and computing researchers need to study how helpful online information could contribute to minimising other risks and dangers of social media regarding DSH.

Meanwhile, the second phase study analysed self-harm content on Twitter. The examined support organisations demonstrate positive use of Twitter to support users suffering from self-harm and other mental health problems. Information science researchers may consider investigating other sources of information offering online social support to self-harming users. There could be various underlying factors underpinning self-harm-related information sharing on social media. Therefore, further effort is needed to explore the motives of posting self-harm information on social media.

Moreover, studying the general characteristics of self-injurious content spanning different social networking tools is crucial. This is because self-harm information found in text tweets could be different from images or videos presenting self-harm. For example, in the case of Twitter, computer science researchers should consider investigating the multimedia aspects of the tweets, including the web links, as they contain critical information. Additionally, as seen from the findings of the second phase study, online users tend to use different self-injurious hashtags to communicate about self-harm.

Therefore, our research community, including social media companies, must be aware of the new ambiguous hashtags about self-harm constantly evolving in social spaces. Hence, there is a need for computer science researchers to study the creation and use of ambiguous hashtags. The idea is to propose a technique to predict future ambiguous tags that online users could produce automatically. Twitter moderates, detect and remove content violating their platform's terms and conditions.

Although users can report online members sharing harmful content, social media companies should be aware of the different tactics used by the online community to violate the platform's terms of use and share harmful content. Therefore, studying how encoded self-injurious content proliferates from one social network to another is essential. Furthermore, while young people are incredibly active on social networks, little is known about the excessive use of social networks concerning self-harm and how such platforms affect online individuals' psychological well-being or mental health issues like self-harm. Hence, mental health and computer science researchers need to collaboratively research to investigate the relationship between self-harm and the excessive use of social media, particularly among self-harming individuals.

In conclusion, based on the results of this study, it is essential to draw the attention of social media companies to continue to monitor user-generated content. Doing this could improve online safety and minimise self-harm risk, as the behaviour is significantly associated with suicide. Even though the open nature of social media presents numerous issues in guaranteeing high-quality and trustworthy content, future research should consider investigating self-harm content on social media sites like **TikTok**, which young people currently dominate.

7.5 Thesis summary and reflections

This chapter discusses the study, the research implications and its limitations. While highlighting the contributions of the research study, it then concluded the doctoral thesis by summarising the research questions and several possible recommendations for future investigations. Intentional self-harm is a critical public health crisis. Platforms like YouTube and Twitter are online platforms where people discuss self-harm, report their episodes of self-harming and share their related views and thoughts about the behaviour. Thus, it was necessary to examine self-injurious content posted on those platforms to analyse the possible effects of such content and understand the influence of those platforms concerning self-harm behaviour.

Contrary to many related studies, this doctoral study was expected to uncover information disclosed on the examined online spaces that attributed to the effects of self-injurious content on online users and understand the impacts of social media platforms in preventing the proliferation of such content. Although people may have negative views or attitudes toward self-harming individuals, it was expected that most users from the examined platforms might demonstrate positive feelings toward users struggling with self-harm and offer informal support to facilitate recovery. Also, this research study sought to understand the positive impact of online support provided by organisations supporting self-harming people. While self-injurious information could be dangerous or helpful to online members [170], this study assumed that online users might perceive information from other users and support organisations as empowering and helpful.

Consequently, this research undertook an in-depth analysis of self-injurious content posted on YouTube and Twitter social networks in order to uncover important insight regarding the various sources of such content [170], the nature of critical content that needs urgent attention [156], hidden themes discussed by online members [4, 5, 170], views and opinions of the online community about self-harm, and the effects of online social information perceived through organisations fighting deliberate self-harm [5].

However, this study considered a mixed-method approach. The purpose of this approach is twofold. Firstly, the approach was informed by the existing studies on mental health and social computing. Secondly, the approach was adopted due to the nature of our research questions outlined and explained in Section 1.3. Similarly, different phases were considered while conducting this study to achieve aims and objectives. Although the analysed data was unstructured, this study found hidden information or themes associated with self-harm discussions [5, 170].

Similarly, campaign strategies used by support organisations to combat self-harm were found. Also, the views of the online community concerning DSH were discovered [5]. Finally, to a large extent, the findings of this study go beyond what was initially expected. While discovering these, the findings contradict other studies' views (including traditional media) that online social platforms could encourage or normalise self-harming behaviour [37]. Thus, the findings increase our understanding of the positive influences of using social media to fight self-injurious behaviour.

Appendix A

Coding Scheme for Classifying Video Sources

Coding is a frequently used procedure in qualitative research for categorising data into understandable themes, or subjects [171]. A *code* is a meaning obtained from texts that illustrate or represent the theme associated with the texts [171].

Table A.1: Video channels classification criteria

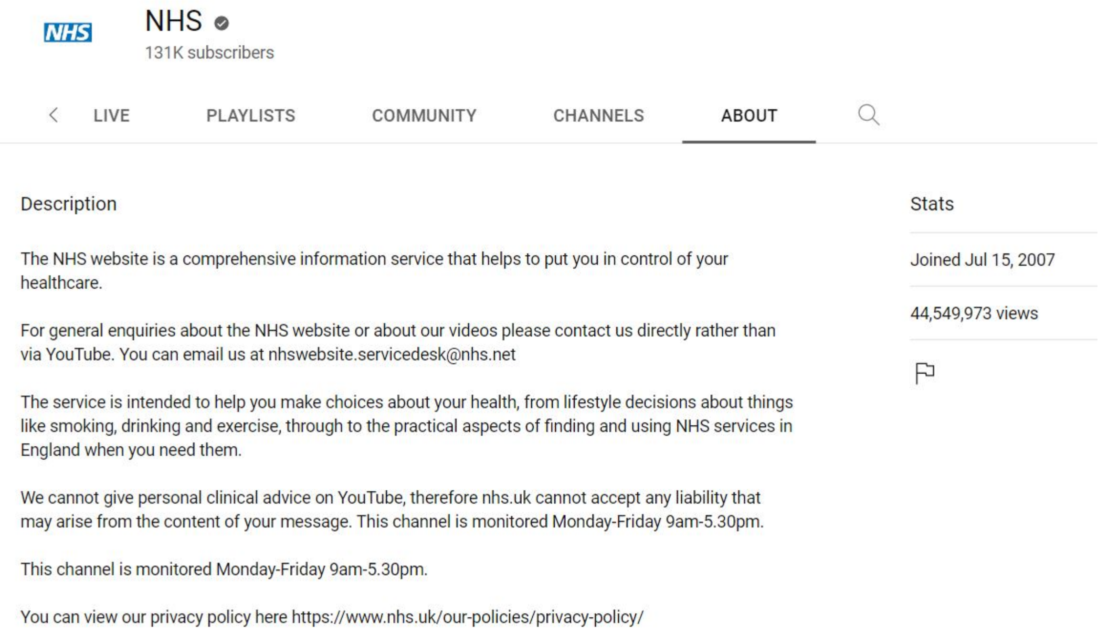
Category	**Code**
Professionals	**001**
Description	This group of channels represents experts that are creating and uploading videos concerning self-harm. As reported in their channel's description, these are a group of Professional YouTubers from academic or medical institutions. Examples could be psychologists, medical professors, psychiatrists and other related experts.
Non-professionals	**002**

Description	Video channels described the uploader and the video presenter as the same person with no medical or academic training to support self-harming people. An example of channels representing this group could be a blogger or social activist who raised awareness about self-harm.
News media	**003**
Description	This category represents a group of channels managed by traditional news media companies, including local and international news agencies that uploaded videos concerning intentional self-harm.
Government org.	**004**
Description	Our analysis considered channels representing government-sponsored organisations, such as educational and medical institutions, that posted self-harm videos on YouTube in this group.
Private org.	**005**
Description	This class is primarily for private companies that own video channels advertising and promoting mental health applications, mainly to self-harming people.
Support org.	**006**
Description	This group of channels belonged to organisations offering mental support on YouTube. Examples of those organisations are the Samaritans in the United Kingdom, who created self-harm videos to increase awareness, provide support, and encourage recovery.

Appendix B

Examples of video channels

Figure B.1



Description

The NHS website is a comprehensive information service that helps to put you in control of your healthcare.

For general enquiries about the NHS website or about our videos please contact us directly rather than via YouTube. You can email us at nhswebsite.servicedesk@nhs.net

The service is intended to help you make choices about your health, from lifestyle decisions about things like smoking, drinking and exercise, through to the practical aspects of finding and using NHS services in England when you need them.

We cannot give personal clinical advice on YouTube, therefore nhs.uk cannot accept any liability that may arise from the content of your message. This channel is monitored Monday-Friday 9am-5.30pm.

This channel is monitored Monday-Friday 9am-5.30pm.

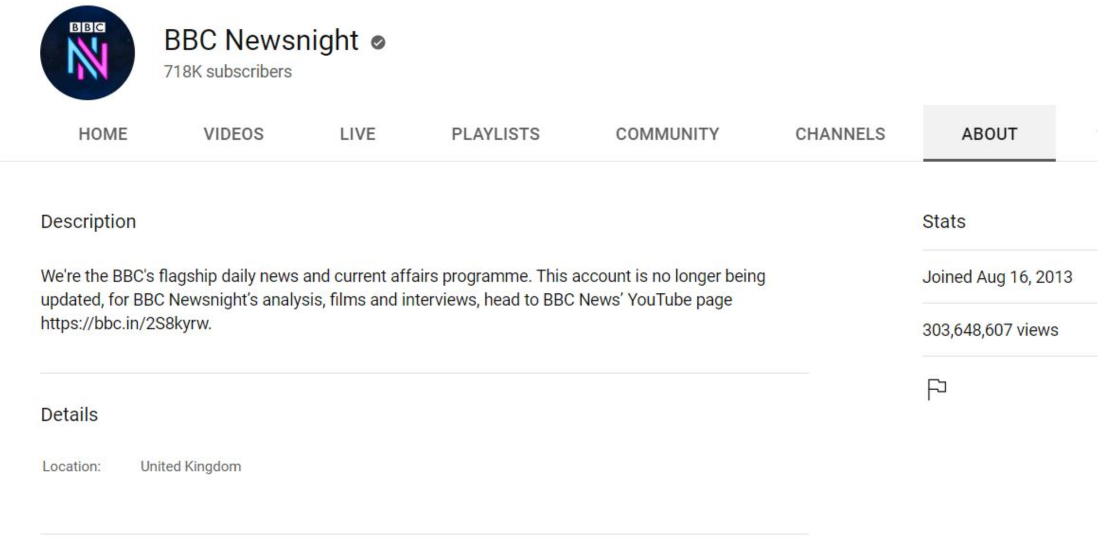
You can view our privacy policy here https://www.nhs.uk/our-policies/privacy-policy/

Stats

Joined Jul 15, 2007

44,549,973 views

(a) Government Organisation



BBC Newsnight

718K subscribers

Description

We're the BBC's flagship daily news and current affairs programme. This account is no longer being updated, for BBC Newsnight's analysis, films and interviews, head to BBC News' YouTube page https://bbc.in/2S8kyrw.

Details

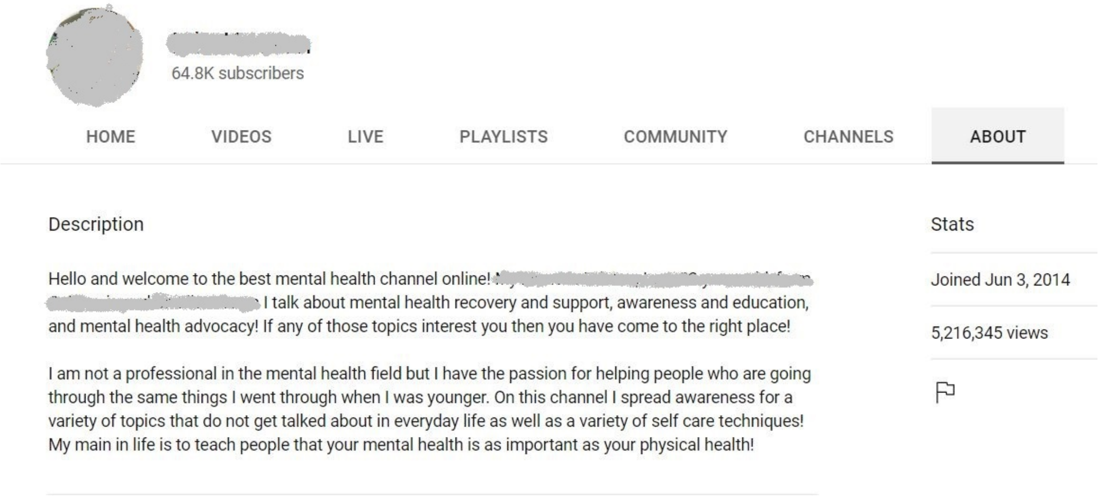
Location: United Kingdom

Stats

Joined Aug 16, 2013

303,648,607 views

(b) News Media



[redacted]

64.8K subscribers

Description

Hello and welcome to the best mental health channel online! [redacted] I talk about mental health recovery and support, awareness and education, and mental health advocacy! If any of those topics interest you then you have come to the right place!

I am not a professional in the mental health field but I have the passion for helping people who are going through the same things I went through when I was younger. On this channel I spread awareness for a variety of topics that do not get talked about in everyday life as well as a variety of self care techniques! My main in life is to teach people that your mental health is as important as your physical health!

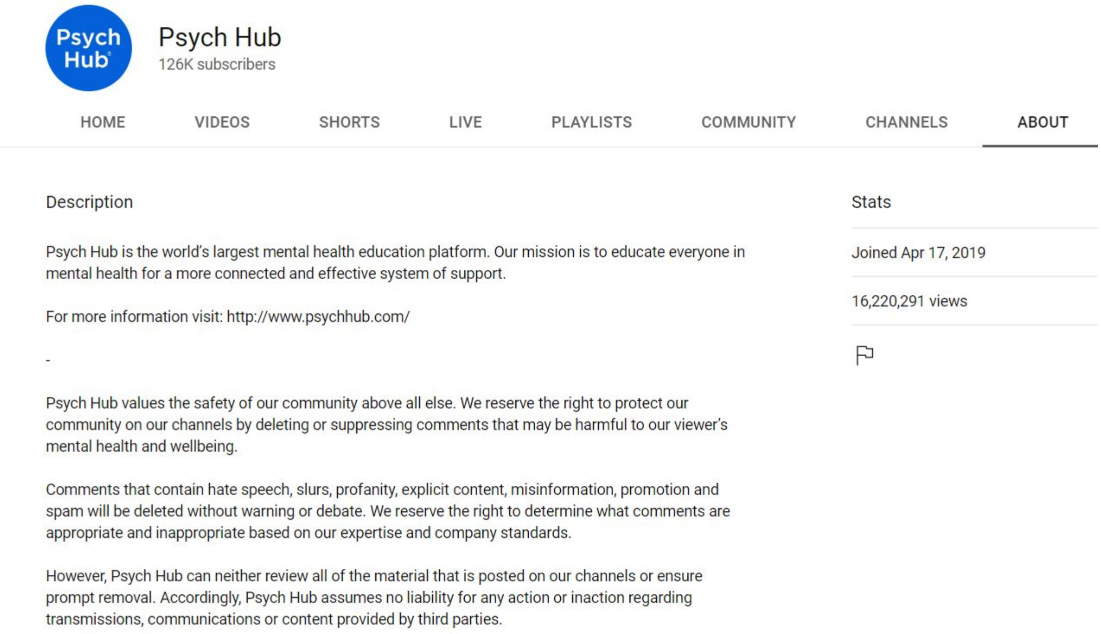
Stats

Joined Jun 3, 2014

5,216,345 views

(c) Non-professional

Appendix B. Examples of video channels



Psych Hub
126K subscribers

Description

Psych Hub is the world's largest mental health education platform. Our mission is to educate everyone in mental health for a more connected and effective system of support.

For more information visit: http://www.psychhub.com/

-

Psych Hub values the safety of our community above all else. We reserve the right to protect our community on our channels by deleting or suppressing comments that may be harmful to our viewer's mental health and wellbeing.

Comments that contain hate speech, slurs, profanity, explicit content, misinformation, promotion and spam will be deleted without warning or debate. We reserve the right to determine what comments are appropriate and inappropriate based on our expertise and company standards.

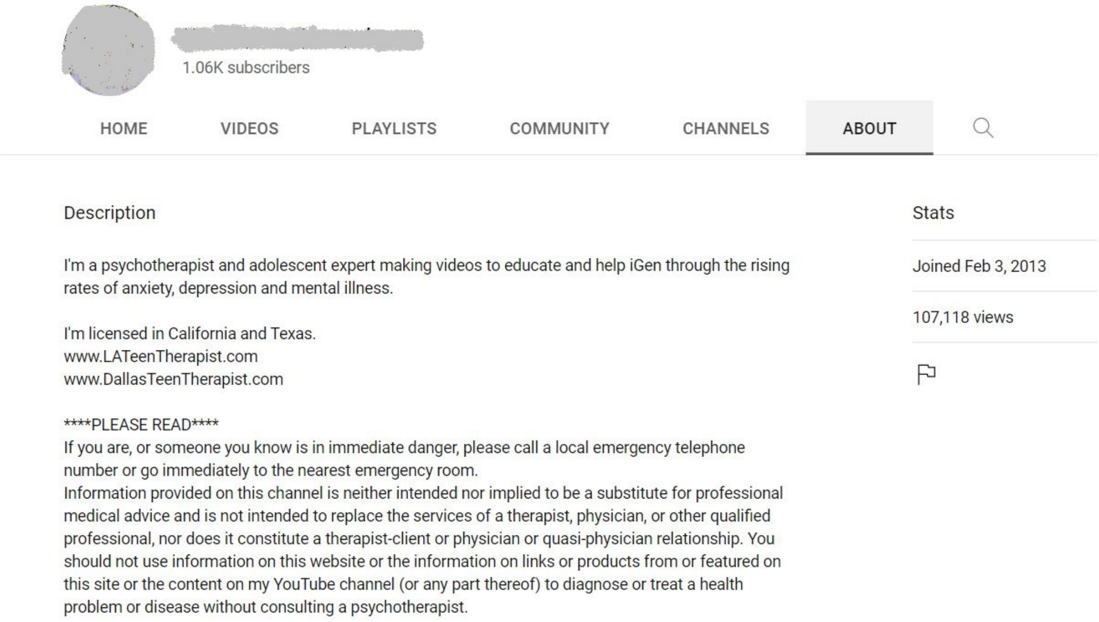
However, Psych Hub can neither review all of the material that is posted on our channels or ensure prompt removal. Accordingly, Psych Hub assumes no liability for any action or inaction regarding transmissions, communications or content provided by third parties.

Stats

Joined Apr 17, 2019

16,220,291 views

(d) Private Organisation



[redacted]
1.06K subscribers

Description

I'm a psychotherapist and adolescent expert making videos to educate and help iGen through the rising rates of anxiety, depression and mental illness.

I'm licensed in California and Texas.
www.LATeenTherapist.com
www.DallasTeenTherapist.com

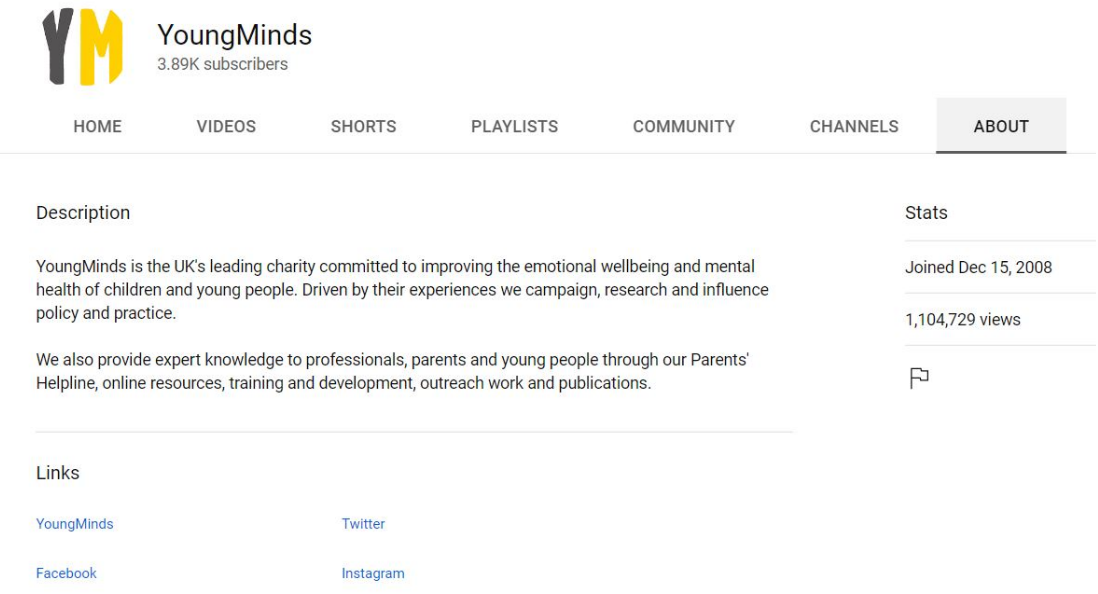
****PLEASE READ****
If you are, or someone you know is in immediate danger, please call a local emergency telephone number or go immediately to the nearest emergency room.
Information provided on this channel is neither intended nor implied to be a substitute for professional medical advice and is not intended to replace the services of a therapist, physician, or other qualified professional, nor does it constitute a therapist-client or physician or quasi-physician relationship. You should not use information on this website or the information on links or products from or featured on this site or the content on my YouTube channel (or any part thereof) to diagnose or treat a health problem or disease without consulting a psychotherapist.

Stats

Joined Feb 3, 2013

107,118 views

(e) Professional (expert)



Description

YoungMinds is the UK's leading charity committed to improving the emotional wellbeing and mental health of children and young people. Driven by their experiences we campaign, research and influence policy and practice.

We also provide expert knowledge to professionals, parents and young people through our Parents' Helpline, online resources, training and development, outreach work and publications.

Stats

Joined Dec 15, 2008

1,104,729 views

Links

YoungMinds

Twitter

Facebook

Instagram

(f) Support Organisation

Appendix C

Coding Scheme for Comments Annotation

Table C.1: Comments annotation guide

Class	Description	Example comment
Amber	These are video responses that do not require immediate action from YouTube moderators. An example of this comment may include self-disclosure or expression of complicated emotions associated with mental health issues like anxiety or depression.	*"I suffer with depression and I think of self-harm but I do not do it but hearing this made me smile with a real smile in a long time"*

Crisis	While some of the audiences may disclose when and how they self-injured, the **crisis** group of comments indicates that the viewer is self-harming and seeks support. Also, responses describing the video as a *triggering* content were considered in this group.	*”Well my parents do not care...they noticed but they just did nothing. My mom said once I'm gonna kill you if you do that and I continued and even when she noticed she did literally nothing. I should also say that I had two ways to self-harm cutting and stitching...yes I stitched my body with an actual needle and thread so it was pretty visible.”*
Green	This group of comments represent clean responses and requires no further action from the YouTube moderators. Examples of these comments include peer support, advice, and recovery-related responses.	*”I'm glad people like you do care about people like us. the video is right, we do need someone to listen. :) thank you. ”*
Red	The red class comments represent video responses showing suicidal thoughts among viewers and therefore need urgent attention from the YouTube moderators. While this category of comments shows the feeling of worthlessness by the commentator or reporting a clear suicide signal, they could also encourage suicide.	*”if cutters are constantly rolling in their own misery and contemplating on killing themselves, then do it already! nobody can stop you. It is your own life, there is nothing wrong with killing yourself. You will be doing the world a favour”*

Bibliography

[1] M. K. Nock, *Understanding nonsuicidal self-injury: Origins, assessment, and treatment.* American Psychological Association, 2009.

[2] M. Nock, "Why do people hurt themselves? new insights into the nature and functions of self-injury." *Current Directions in Psychological Science*, vol. 18, no. 2, pp. 78–83, 2009.

[3] K. Hawton, K. E. Saunders, and R. C. O'Connor, "Self-harm and suicide in adolescents," *The Lancet*, vol. 379, no. 9834, pp. 2373–2382, 2012.

[4] M. A. Alhassan and D. Pennington, "Investigating non-suicidal self-injury discussions on twitter," in *International Conference on Social Media and Data Mining-ICSMDM*, 2021.

[5] M. A. Alhassan, I. Inuwa-Dutse, B. S. Bello, and D. Pennington, "Self-harm: detection and support on twitter," *arXiv preprint arXiv:2104.00174*, 2021.

[6] Y. Lei, "Individual intelligent method-based fault diagnosis," in *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery.* Butterworth-Heinemann, 2017, pp. 67–174.

[7] C. Sievert and K. Shirley, "Ldavis: A method for visualizing and interpreting topics," in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.

[8] N. Madge, A. Hewitt, K. Hawton, E. J. d. Wilde, P. Corcoran, S. Fekete, K. v. Heeringen, D. D. Leo, and M. Ystgaard, "Deliberate self-harm within an interna-

tional community sample of young people: comparative findings from the child & adolescent self-harm in europe (case) study," *Journal of child Psychology and Psychiatry*, vol. 49, no. 6, pp. 667–677, 2008.

[9] M. M. Greaves and C. Dykeman, "A corpus linguistic analysis of public tumblr blog posts on non-suicidal self-injury," 2019.

[10] N. Madge, A. Hewitt, and K. Hawton, "Wilde ejd, corcoran p, fekete s, heeringen kv, leo dd, ystgaard m: deliberate self-harm within an international community sample of young people: comparative findings from the child & adolescent self-harm in europe (case) study," *J Child Psychol Psychiatry*, vol. 49, no. 6, pp. 667–77, 2008.

[11] N. C. C. for Mental Health (UK *et al.*, "Self-harm: the short-term physical and psychological management and secondary prevention of self-harm in primary and secondary care," 2004.

[12] K. Skegg, "Self-harm," *The Lancet*, vol. 366, no. 9495, pp. 1471–1483, 2005.

[13] S. P. Lewis and N. L. Heath, "Nonsuicidal self-injury," *CMAJ*, vol. 185, no. 6, pp. 505–505, 2013.

[14] M. K. Nock and A. R. Favazza, "Nonsuicidal self-injury: Definition and classification." 2009.

[15] R. Brunner, M. Kaess, P. Parzer, G. Fischer, V. Carli, C. W. Hoven, C. Wasserman, M. Sarchiapone, F. Resch, A. Apter *et al.*, "Life-time prevalence and psychosocial correlates of adolescent direct self-injurious behavior: A comparative study of findings in 11 european countries," *Journal of Child Psychology and Psychiatry*, vol. 55, no. 4, pp. 337–348, 2014.

[16] D. Ougrin, "Commentary: Self-harm in adolescents: the best predictor of death by suicide?–reflections on hawton et al.(2012)," *Journal of child psychology and psychiatry*, vol. 53, no. 12, pp. 1220–1221, 2012.

[17] K. Hawton, H. Bergen, N. Kapur, J. Cooper, S. Steeg, J. Ness, and K. Waters, "Repetition of self-harm and suicide following self-harm in children and adolescents: Findings from the multicentre study of self-harm in england," *Journal of child psychology and psychiatry*, vol. 53, no. 12, pp. 1212–1219, 2012.

[18] C. Haw, K. Hawton, K. Houston, and E. Townsend, "Correlates of relative lethality and suicidal intent among deliberate self-harm patients," *Suicide and Life-Threatening Behavior*, vol. 33, no. 4, pp. 353–364, 2003.

[19] G. C. Patton, C. Coffey, S. M. Sawyer, R. M. Viner, D. M. Haller, K. Bose, T. Vos, J. Ferguson, and C. D. Mathers, "Global patterns of mortality in young people: a systematic analysis of population health data," *The lancet*, vol. 374, no. 9693, pp. 881–892, 2009.

[20] M. Duggan, J. Brenner *et al.*, *The demographics of social media users, 2012*. Pew Research Center's Internet & American Life Project Washington, DC, 2013, vol. 14.

[21] M. P. Dyson, L. Hartling, J. Shulhan, A. Chisholm, A. Milne, P. Sundar, S. D. Scott, and A. S. Newton, "A systematic review of social media use to discuss and view deliberate self-harm acts," *PloS one*, vol. 11, no. 5, p. e0155813, 2016.

[22] C. Biernesser, C. J. Sewall, D. Brent, T. Bear, C. Mair, and J. Trauth, "Social media use and deliberate self-harm among youth: a systematized narrative review," *Children and youth services review*, vol. 116, p. 105054, 2020.

[23] P. A. Cavazos-Rehg, M. J. Krauss, S. J. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, R. Grucza, and L. J. Bierut, "An analysis of depression, self-harm, and suicidal ideation content on tumblrjacob2017influence," *Crisis*, 2016.

[24] S. Chancellor, Z. Lin, and M. De Choudhury, "" this post will just get taken down" characterizing removed pro-eating disorder social media content," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 1157–1162.

[25] A. Marchant, K. Hawton, A. Stewart, P. Montgomery, V. Singaravelu, K. Lloyd, N. Purdy, K. Daine, and A. John, "A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown," *PLoS one*, vol. 12, no. 8, p. e0181722, 2017.

[26] S. P. Lewis and Y. Seko, "A double-edged sword: A review of benefits and risks of online nonsuicidal self-injury activities," *Journal of clinical psychology*, vol. 72, no. 3, pp. 249–262, 2016.

[27] A. Topping, "Self-harm sites and cyberbullying: the threat to children from web's dark side," *The Guardian*, vol. 10, 2014.

[28] K. J. Mitchell and M. L. Ybarra, "Online behavior of youth who engage in self-harm provides clues for preventive intervention," *Preventive medicine*, vol. 45, no. 5, pp. 392–396, 2007.

[29] K. Daine, K. Hawton, V. Singaravelu, A. Stewart, S. Simkin, and P. Montgomery, "The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people," *PloS one*, vol. 8, no. 10, p. e77555, 2013.

[30] S. P. Lewis, Y. Seko, and P. Joshi, "The impact of youtube peer feedback on attitudes toward recovery from non-suicidal self-injury: An experimental pilot study," *Digital health*, vol. 4, p. 2055207618780499, 2018.

[31] Y. Wang, J. Tang, J. Li, B. Li, Y. Wan, C. Mellina, N. O'Hare, and Y. Chang, "Understanding and discovering deliberate self-harm content in social media," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 93–102.

[32] Y. Seko, "Picturesque wounds: A multimodal analysis of self-injury photographs on flickr," in *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, vol. 14, no. 2, 2013.

[33] C. Emma Hilton, "Unveiling self-harm behaviour: what can social media site twitter tell us about self-harm? a qualitative exploration," *Journal of clinical nursing*, vol. 26, no. 11-12, pp. 1690–1704, 2017.

[34] B. Mars, J. Heron, L. Biddle, J. L. Donovan, R. Holley, M. Piper, J. Potokar, C. Wyllie, and D. Gunnell, "Exposure to, and searching for, information about suicide and self-harm on the internet: Prevalence and predictors in a population based cohort of young adults," *Journal of affective disorders*, vol. 185, pp. 239–245, 2015.

[35] L. Michelmore and P. Hindley, "Help-seeking for suicidal thoughts and self-harm in young people: A systematic review," *Suicide and Life-Threatening Behavior*, vol. 42, no. 5, pp. 507–524, 2012.

[36] N. Shanahan, C. Brennan, and A. House, "Self-harm and social media: thematic analysis of images posted on three social media sites," *BMJ open*, vol. 9, no. 2, p. e027006, 2019.

[37] S. P. Lewis, N. L. Heath, M. J. Sornberger, and A. E. Arbuthnott, "Helpful or harmful? an examination of viewers' responses to nonsuicidal self-injury videos on youtube," *Journal of Adolescent Health*, vol. 51, no. 4, pp. 380–385, 2012.

[38] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[39] J. A. Pater, O. L. Haimson, N. Andalibi, and E. D. Mynatt, ""hunger hurts but starving works" characterizing the presentation of eating disorders online," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 1185–1200.

[40] E. M. Miguel, T. Chou, A. Golik, D. Cornacchio, A. L. Sanchez, M. DeSerisy, and J. S. Comer, "Examining the scope and patterns of deliberate self-injurious cutting content in popular social media," *Depression and anxiety*, vol. 34, no. 9, pp. 786–793, 2017.

[41] J. Whitlock, J. Muehlenkamp, J. Eckenrode, A. Purington, G. B. Abrams, P. Barreira, and V. Kress, “Nonsuicidal self-injury as a gateway to suicide in young adults,” *Journal of Adolescent Health*, vol. 52, no. 4, pp. 486–492, 2013.

[42] C. D. Murray and J. Fox, “Do internet self-harm discussion groups alleviate or exacerbate self-harming behaviour?” *Australian e-journal for the advancement of mental health*, vol. 5, no. 3, pp. 225–233, 2006.

[43] S. Jacoby and E. Ochs, “Co-construction: An introduction,” pp. 171–183, 1995.

[44] S. A. Sumner, S. Galik, J. Mathieu, M. Ward, T. Kiley, B. Bartholow, A. Dingwall, and P. Mork, “Temporal and geographic patterns of social media posts about an emerging suicide game,” *Journal of Adolescent Health*, vol. 65, no. 1, pp. 94–100, 2019.

[45] M. A. Moreno, A. Ton, E. Selkie, and Y. Evans, “Secret society 123: Understanding the language of self-harm on instagram,” *Journal of Adolescent Health*, vol. 58, no. 1, pp. 78–84, 2016.

[46] S. Sharkey, J. Smithson, E. Hewis, R. JOnES, T. EMMEnS, T. Ford, and C. OWEnS, “Supportive interchanges and face-work as’ protective talk’in an online self-harm support forum,” *Communication & medicine*, vol. 9, no. 1, p. 71, 2012.

[47] F.-Y. Tseng and H.-J. Yang, “Internet use and web communication networks, sources of social support, and forms of suicidal and nonsuicidal self-injury among adolescents: Different patterns between genders,” *Suicide and Life-Threatening Behavior*, vol. 45, no. 2, pp. 178–191, 2015.

[48] S. P. Lewis, N. L. Heath, J. M. St Denis, and R. Noble, “The scope of nonsuicidal self-injury on youtube,” *Pediatrics*, vol. 127, no. 3, pp. e552–e557, 2011.

[49] S. P. Lewis and A. K. Knoll, “Do it yourself: Examination of self-injury first aid tips on youtube,” *Cyberpsychology, Behavior, and Social Networking*, vol. 18, no. 5, pp. 301–304, 2015.

[50] A. Yates, A. Cohan, and N. Goharian, “Depression and self-harm risk assessment in online forums,” *arXiv preprint arXiv:1709.01848*, 2017.

[51] J. Adams, K. Rodham, and J. Gavin, “Investigating the “self” in deliberate self-harm,” *Qualitative health research*, vol. 15, no. 10, pp. 1293–1309, 2005.

[52] S. Rasmussen, K. Hawton, S. Philpott-Morgan, and R. C. O’Connor, “Why do adolescents self-harm?” *Crisis*, 2016.

[53] Z. C. Steinert-Threlkeld, *Twitter as data.* Cambridge University Press, 2018.

[54] S. Greenwood, A. Perrin, and M. Duggan, “Social media update 2016,” *Pew Research Center*, vol. 11, no. 2, pp. 1–18, 2016.

[55] M. K. Nock, “Self-injury,” *Annual review of clinical psychology*, vol. 6, pp. 339–363, 2010.

[56] S. A. S. Germain and J. M. Hooley, “Direct and indirect forms of non-suicidal self-injury: Evidence for a distinction,” *Psychiatry research*, vol. 197, no. 1-2, pp. 78–84, 2012.

[57] D. Owens, J. Horrocks, and A. House, “Fatal and non-fatal repetition of self-harm: systematic review,” *The British Journal of Psychiatry*, vol. 181, no. 3, pp. 193–199, 2002.

[58] S. A. S. Germain, *Expanding the Conceptualization of Self-Injurious Behavior: Are All Forms of Self-Injury Created Equal?* Harvard University, 2011.

[59] K. Hawton and A. James, “Suicide and deliberate self harm in young people,” *Bmj*, vol. 330, no. 7496, pp. 891–894, 2005.

[60] M. K. Nock and R. C. Kessler, “Prevalence of and risk factors for suicide attempts versus suicide gestures: analysis of the national comorbidity survey.” *Journal of abnormal psychology*, vol. 115, no. 3, p. 616, 2006.

[61] N. Kapur, J. Cooper, R. C. O'Connor, and K. Hawton, "Non-suicidal self-injury v. attempted suicide: new diagnosis or false dichotomy?" *The British Journal of Psychiatry*, vol. 202, no. 5, pp. 326–328, 2013.

[62] G. J. Lengel, B. A. Ammerman, and J. J. Washburn, "Clarifying the definition of nonsuicidal self-injury: Clinician and researcher perspectives." *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, vol. 43, no. 2, p. 119, 2022.

[63] M. K. Nock, E. B. Holmberg, V. I. Photos, and B. D. Michel, "Self-injurious thoughts and behaviors interview: development, reliability, and validity in an adolescent sample." 2007.

[64] C. M. Jacobson and M. Gould, "The epidemiology and phenomenology of non-suicidal self-injurious behavior among adolescents: A critical review of the literature," *Archives of Suicide Research*, vol. 11, no. 2, pp. 129–147, 2007.

[65] L. Doyle, M. P. Treacy, and A. Sheridan, "Self-harm in young people: Prevalence, associated factors, and help-seeking in school-going adolescents," *International journal of mental health nursing*, vol. 24, no. 6, pp. 485–494, 2015.

[66] S. Curtis, P. Thorn, A. McRoberts, S. Hetrick, S. Rice, and J. Robinson, "Caring for young people who self-harm: A review of perspectives from families and young people," *International journal of environmental research and public health*, vol. 15, no. 5, p. 950, 2018.

[67] E. D. Klonsky and A. Moyer, "Childhood sexual abuse and non-suicidal self-injury: meta-analysis," *The British Journal of Psychiatry*, vol. 192, no. 3, pp. 166–170, 2008.

[68] M. K. Nock, "Actions speak louder than words: An elaborated theoretical model of the social functions of self-injury and other harmful behaviors," *Applied and preventive psychology*, vol. 12, no. 4, pp. 159–168, 2008.

[69] G. Scoliers, G. Portzky, N. Madge, A. Hewitt, K. Hawton, E. J. De Wilde, M. Ystgaard, E. Arensman, D. De Leo, S. Fekete *et al.*, "Reasons for adolescent deliberate self-harm: a cry of pain and/or a cry for help?" *Social psychiatry and psychiatric epidemiology*, vol. 44, no. 8, pp. 601–607, 2009.

[70] J. Nesi, T. A. Burke, H. R. Lawrence, H. A. MacPherson, A. Spirito, and J. C. Wolff, "Online self-injury activities among psychiatrically hospitalized adolescents: prevalence, functions, and perceived consequences," *Research on child and adolescent psychopathology*, vol. 49, no. 4, pp. 519–531, 2021.

[71] S. Mythily, S. Qiu, M. Winslow *et al.*, "Prevalence and correlates of excessive internet use among youth in singapore," *Annals Academy of Medicine Singapore*, vol. 37, no. 1, p. 9, 2008.

[72] K. Y. McKenna, S. Sprecher, A. Wenzel, and J. Harvey, "Myspace or your place: Relationship initiation and development in the wired and wireless world," *Handbook of relationship initiation*, pp. 235–247, 2008.

[73] R. Kraut, M. Patterson, V. Lundmark, S. Kiesler, T. Mukophadhyay, and W. Scherlis, "Internet paradox: A social technology that reduces social involvement and psychological well-being?" *American psychologist*, vol. 53, no. 9, p. 1017, 1998.

[74] A. L. Giordano, L. A. Lundeen, K. L. Wester, J. Lee, S. Vickers, M. K. Schmit, and I. K. Kim, "Nonsuicidal self-injury on instagram: Examining hashtag trends," *International Journal for the Advancement of Counselling*, vol. 44, no. 1, pp. 1–16, 2022.

[75] K. Becker, M. Mayer, M. Nagenborg, M. El-Faddagh, and M. H. Schmidt, "Parasuicide online: Can suicide websites trigger suicidal behaviour in predisposed adolescents?" *Nordic journal of psychiatry*, vol. 58, no. 2, pp. 111–114, 2004.

[76] S. P. Lewis and T. G. Baker, "The possible risks of self-injury web sites: a content analysis," *Archives of suicide research*, vol. 15, no. 4, pp. 390–396, 2011.

[77] C. Moss, C. Wibberley, and G. Witham, “Assessing the impact of instagram use and deliberate self-harm in adolescents: A scoping review,” *International journal of mental health nursing*, 2022.

[78] J. Smithson, S. Sharkey, E. Hewis, R. Jones, T. Emmens, T. Ford, and C. Owens, “Problem presentation and responses on an online forum for young people who self-harm,” *Discourse studies*, vol. 13, no. 4, pp. 487–501, 2011.

[79] A. M. Kaplan and M. Haenlein, “Social media: back to the roots and back to the future,” *Journal of Systems and Information Technology*, 2012.

[80] A. KAPLAN, “Users of the world, unite! the challenges and opportunities of social media,” *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[81] D. R. Neal, *Social media for academics: a practical guide.* Elsevier, 2012.

[82] E. Dean, C. A. Hill, and J. Murphy, *Social Media, Sociality, and Survey Research.* Wiley, 2014.

[83] S. Rasmussen, K. Hawton *et al.*, “Adolescent self-harm: a school-based study in northern ireland,” *Journal of affective disorders*, vol. 159, pp. 46–52, 2014.

[84] E. S. Messina and Y. Iwasaki, “Internet use and self-injurious behaviors among adolescents and young adults: An interdisciplinary literature review and implications for health professionals,” *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 3, pp. 161–168, 2011.

[85] A. M. Memon, S. G. Sharma, S. S. Mohite, and S. Jain, “The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature,” *Indian journal of psychiatry*, vol. 60, no. 4, p. 384, 2018.

[86] I. Cataldo, B. Lepri, M. J. Y. Neoh, and G. Esposito, “Social media usage and development of psychiatric disorders in childhood and adolescence: a review,” *Frontiers in psychiatry*, vol. 11, p. 508595, 2021.

[87] R. C. Brown, T. Fischer, D. A. Goldwich, and P. L. Plener, "“i just finally wanted to belong somewhere”—qualitative analysis of experiences with posting pictures of self-injury on instagram," *Frontiers in psychiatry*, vol. 11, p. 274, 2020.

[88] O. Santesteban-Echarri, M. Álvarez-Jiménez, J. Gleeson, and S. M. Rice, "Social media interventions for adolescents and young people with depression and psychosis," in *Technology and adolescent mental health.* Springer, 2018, pp. 187–205.

[89] J. M. Duggan, N. L. Heath, S. P. Lewis, and A. L. Baxter, "An examination of the scope and nature of non-suicidal self-injury online activities: Implications for school mental health professionals," *School Mental Health*, vol. 4, no. 1, pp. 56–67, 2012.

[90] T. Vente, M. Daley, E. Killmeyer, L. K. Grubb *et al.*, "Association of social media use and high-risk behaviors in adolescents: cross-sectional study," *JMIR pediatrics and parenting*, vol. 3, no. 1, p. e18043, 2020.

[91] F. Arendt, S. Scherr, and D. Romer, "Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults," *New Media & Society*, vol. 21, no. 11-12, pp. 2422–2442, 2019.

[92] R. A. Record, K. Straub, and N. Stump, "# selfharm on# instagram: examining user awareness and use of instagram's self-harm reporting tool," *Health communication*, 2019.

[93] J. Nesi, S. Choukas-Bradley, and M. J. Prinstein, "Transformation of adolescent peer relations in the social media context: Part 1—a theoretical framework and application to dyadic peer relationships," *Clinical Child and Family Psychology Review*, vol. 21, no. 3, pp. 267–294, 2018.

[94] K. D. Niwa and M. N. Mandrusiak, "Self-injury groups on facebook," *Canadian Journal of Counselling and Psychotherapy*, vol. 46, no. 1, 2012.

[95] T. G. Baker and S. P. Lewis, “Responses to online photographs of non-suicidal self-injury: A thematic analysis,” *Archives of Suicide Research*, vol. 17, no. 3, pp. 223–235, 2013.

[96] H. Smith and W. Cipolli, “The instagram/facebook ban on graphic self-harm imagery: A sentiment analysis and topic modeling approach,” *Policy & Internet*, 2021.

[97] T. Lancet, “Social media, screen time, and young people’s mental health,” p. 611, 2019.

[98] R. C. Brown, T. Fischer, A. D. Goldwich, F. Keller, R. Young, and P. L. Plener, “# cutting: Non-suicidal self-injury (nssi) on instagram,” *Psychological medicine*, vol. 48, no. 2, pp. 337–346, 2018.

[99] R. M. Shafi, P. A. Nakonezny, M. Romanowicz, A. L. Nandakumar, L. Suarez, and P. E. Croarkin, “Suicidality and self-injurious behavior among adolescent social media users at psychiatric hospitalization,” *CNS spectrums*, vol. 26, no. 3, pp. 275–281, 2021.

[100] J. Smithson, S. Sharkey, E. Hewis, R. B. Jones, T. Emmens, T. Ford, and C. Owens, “Membership and boundary maintenance on an online self-harm forum,” *Qualitative Health Research*, vol. 21, no. 11, pp. 1567–1575, 2011.

[101] A. Khasawneh, K. C. Madathil, E. Dixon, P. Wiśniewski, H. Zinzow, R. Roth *et al.*, “Examining the self-harm and suicide contagion effects of the blue whale challenge on youtube and twitter: qualitative study,” *JMIR mental health*, vol. 7, no. 6, p. e15973, 2020.

[102] K. Rodham, J. Gavin, S. P. Lewis, J. M. St. Denis, and P. Bandalli, “An investigation of the motivations driving the online representation of self-injury: A thematic analysis,” *Archives of suicide research*, vol. 17, no. 2, pp. 173–183, 2013.

[103] A. Marchant, K. Hawton, L. Burns, A. Stewart, A. John *et al.*, “Impact of web-based sharing and viewing of self-harm–related videos and photographs on young

people: Systematic review," *Journal of medical Internet research*, vol. 23, no. 3, p. e18048, 2021.

[104] Y. Seko, S. A. Kidd, D. Wiljer, and K. J. McKenzie, "On the creative edge: exploring motivations for creating non-suicidal self-injury content online," *Qualitative health research*, vol. 25, no. 10, pp. 1334–1346, 2015.

[105] S. P. Lewis and N. J. Michal, "Start, stop, and continue: Preliminary insight into the appeal of self-injury e-communities," *Journal of health psychology*, vol. 21, no. 2, pp. 250–260, 2016.

[106] J. Picardo, S. K. McKenzie, S. Collings, and G. Jenkin, "Suicide and self-harm content on instagram: A systematic scoping review," *PloS one*, vol. 15, no. 9, p. e0238603, 2020.

[107] S. Garcia and K. Barclay, "Adapting research methodologies in the covid-19 pandemic: Resources for researchers," *Nippon Foundation Ocean Nexus, EarthLab, University of Washington*, 2020.

[108] V. Braun and V. Clarke, "What can "thematic analysis" offer health and wellbeing researchers?" 2014.

[109] N. Jacob, R. Evans, and J. Scourfield, "The influence of online images on self-harm: A qualitative study of young people aged 16–24," *Journal of adolescence*, vol. 60, pp. 140–147, 2017.

[110] M. Weisser, *Practical corpus linguistics: An introduction to corpus-based language analysis.* John Wiley & Sons, 2016, vol. 43.

[111] A. Lavis and R. Winter, "# online harms or benefits? an ethnographic analysis of the positives and negatives of peer-support around self-harm on social media," *Journal of child psychology and psychiatry*, vol. 61, no. 8, pp. 842–854, 2020.

[112] H. Lambert and C. McKevitt, "Anthropology in health research: from qualitative methods to multidisciplinarity," *Bmj*, vol. 325, no. 7357, pp. 210–213, 2002.

[113] C. Hine, *Virtual ethnography.* Sage, 2000.

[114] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[115] D. A. Pisner and D. M. Schnyer, “Support vector machine,” in *Machine Learning.* Elsevier, 2020, pp. 101–121.

[116] S. Scherr, F. Arendt, T. Frissen, and J. Oramas M, “Detecting intentional self-harm on instagram: development, testing, and validation of an automatic image-recognition algorithm to discover cutting-related posts,” *Social Science Computer Review*, vol. 38, no. 6, pp. 673–685, 2020.

[117] M. E. Aragón, A. P. López-Monroy, L. C. González, and M. Montes-y Gómez, “Detecting traces of self-harm on reddit through emotional patterns,” in *Early Detection of Mental Health Disorders by Social Media Monitoring.* Springer, 2022, pp. 207–234.

[118] L. Xian, S. D. Vickers, A. L. Giordano, J. Lee, I. K. Kim, and L. Ramaswamy, “# selfharm on instagram: Quantitative analysis and classification of non-suicidal self-injury,” in *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI).* IEEE, 2019, pp. 61–70.

[119] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, “Adversarial complementary learning for weakly supervised object localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1325–1334.

[120] T. Highfield and T. Leaver, “A methodology for mapping instagram hashtags,” *First Monday*, vol. 20, no. 1, pp. 1–11, 2015.

[121] I. Mergel, “A framework for interpreting social media interactions in the public sector,” *Government information quarterly*, vol. 30, no. 4, pp. 327–334, 2013.

[122] A. Bryman, *Social research methods.* Oxford university press, 2016.

[123] R. Mather, "A mixed-methods exploration of an environment for learning computer programming," *Research in Learning Technology*, vol. 23, 2015.

[124] J. W. Creswell and V. L. P. Clark, *Designing and conducting mixed methods research.* Sage publications, 2017.

[125] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction.* Morgan Kaufmann, 2017.

[126] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[127] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014.

[128] M. Daquino, I. Heibi, S. Peroni, and D. Shotton, "Creating restful apis over sparql endpoints using ramose," *Semantic Web*, no. Preprint, pp. 1–19, 2022.

[129] S. McRoberts, E. Bonsignore, T. Peyton, and S. Yarosh, "Do it for the viewers! audience engagement behaviors of young youtubers," in *Proceedings of the The 15th International Conference on Interaction Design and Children*, 2016, pp. 334–343.

[130] D. Gaffney and C. Puschmann, "Data collection on twitter," *Twitter and society*, pp. 55–67, 2014.

[131] K. Kumar and G. Geethakumari, "Detecting misinformation in online social networks using cognitive psychology," *Human-centric Computing and Information Sciences*, vol. 4, no. 1, pp. 1–22, 2014.

[132] S. Santoveña-Casal and C. Bernal-Bravo, "Exploring the influence of the teacher: Social participation on twitter and academic perception," *Comunicar. Media Education Research Journal*, vol. 27, no. 1, 2019.

[133] A. K. Bansal, J. I. Khan, and S. K. Alam, *Introduction to computational health informatics.* CRC Press, 2019.

[134] P. J. Lillie, A. Samson, A. Li, K. Adams, R. Capstick, G. D. Barlow, N. Easom, E. Hamilton, P. J. Moss, A. Evans *et al.*, "Novel coronavirus disease (covid-19): the first two patients in the uk with person to person transmission," *The Journal of infection*, vol. 80, no. 5, p. 578, 2020.

[135] L. Townsend and C. Wallace, "Social media research: A guide to ethics," *University of Aberdeen*, vol. 1, p. 16, 2016.

[136] D. Berrar, "Cross-validation." 2019.

[137] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2019.

[138] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence).* IEEE, 2019, pp. 263–268.

[139] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "Knn-diffusion: Image generation via large-scale retrieval," *arXiv preprint arXiv:2204.02849*, 2022.

[140] E. J. Chikofsky and J. H. Cross, "Reverse engineering and design recovery: A taxonomy," *IEEE software*, vol. 7, no. 1, pp. 13–17, 1990.

[141] B. S. Bello and R. Heckel, "Analyzing the behaviour of twitter bots in post brexit politics," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).* IEEE, 2019, pp. 61–66.

[142] L. Stracqualursi and P. Agati, "Tweet topics and sentiments relating to distance learning among italian twitter users," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[143] M. B. Mutanga and A. Abayomi, “Tweeting on covid-19 pandemic in south africa: Lda-based topic modelling approach,” *African Journal of Science, Technology, Innovation and Development*, vol. 14, no. 1, pp. 163–172, 2022.

[144] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[145] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, “Detecting changes in suicide content manifested in social media following celebrity suicides,” in *Proceedings of the 26th ACM conference on Hypertext & Social Media*, 2015, pp. 85–94.

[146] S. Moghaddam and M. Ester, “Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 665–674.

[147] H. Wang, Y. Ding, J. Tang, X. Dong, B. He, J. Qiu, and D. J. Wild, “Finding complex biological relationships in recent pubmed articles using bio-lda,” *PloS one*, vol. 6, no. 3, p. e17243, 2011.

[148] A. E. Barry, D. Valdez, A. A. Padon, and A. M. Russell, “Alcohol advertising on twitter—a topic model,” *American Journal of Health Education*, vol. 49, no. 4, pp. 256–263, 2018.

[149] F. Alhayan and D. Pennington, “Twitter as health information source: exploring the parameters affecting dementia-related tweets,” in *International Conference on Social Media and Society*, 2020, pp. 277–290.

[150] P. Resnik, A. Garron, and R. Resnik, “Using topic modeling to improve prediction of neuroticism and depression in college students,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1348–1353.

[151] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using twitter sentiment analysis," in *2016 international conference on inventive computation technologies (ICICT)*, vol. 1. IEEE, 2016, pp. 1–5.

[152] Y. B. Kim, J. G. Kim, W. Kim, J. H. Im, T. H. Kim, S. J. Kang, and C. H. Kim, "Predicting fluctuations in cryptocurrency transactions based on user comments and replies," *PloS one*, vol. 11, no. 8, p. e0161197, 2016.

[153] K. Loveys, P. Crutchley, E. Wyatt, and G. Coppersmith, "Small but mighty: affective micropatterns for quantifying mental health from social media language," in *Proceedings of the fourth workshop on computational linguistics and clinical Psychology—From linguistic signal to clinical reality*, 2017, pp. 85–95.

[154] S. P. Lewis, S. A. Rosenrot, and M. A. Messner, "Seeking validation in unlikely places: the nature of online questions about non-suicidal self-injury," *Archives of Suicide Research*, vol. 16, no. 3, pp. 263–272, 2012.

[155] C. Pornpitakpan, "The persuasiveness of source credibility: A critical review of five decades' evidence," *Journal of applied social psychology*, vol. 34, no. 2, pp. 243–281, 2004.

[156] M. A. Alhassan and D. Pennington, "Detecting critical responses from deliberate self-harm videos on youtube," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 2020, pp. 383–386.

[157] A. McLellan, K. Schmidt-Waselenchuk, K. Duerksen, and E. Woodin, "Talking back to mental health stigma: An exploration of youtube comments on anti-stigma videos," *Computers in Human Behavior*, vol. 131, p. 107214, 2022.

[158] C. Brew, "Classifying reachout posts with a radial basis function svm," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 138–142.

[159] D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo, "Clpsych 2016 shared task: Triaging content in online peer-support forums," in *Proceedings of the Third*

Workshop on Computational Linguistics and Clinical Psychology, 2016, pp. 118–127.

[160] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[161] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[162] S. W. Stirman and J. W. Pennebaker, “Word use in the poetry of suicidal and nonsuicidal poets,” *Psychosomatic medicine*, vol. 63, no. 4, pp. 517–522, 2001.

[163] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[164] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew, “Sentiment analysis of suicide notes: A shared task,” *Biomedical informatics insights*, vol. 5, pp. BII–S9042, 2012.

[165] W. Lian, P. Rai, E. Salazar, and L. Carin, “Integrating features and similarities: Flexible models for heterogeneous multiview data,” in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[166] S. P. Lewis, N. L. Heath, N. J. Michal, and J. M. Duggan, “Non-suicidal self-injury, youth, and the internet: What mental health professionals need to know,” *Child and adolescent psychiatry and mental health*, vol. 6, no. 1, pp. 1–9, 2012.

[167] J. L. Whitlock, J. L. Powers, and J. Eckenrode, “The virtual cutting edge: the internet and adolescent self-injury.” *Developmental psychology*, vol. 42, no. 3, p. 407, 2006.

[168] G. Gerbner, L. Gross, M. Morgan, N. Signorielli, J. Shanahan *et al.*, “Growing up with television: Cultivation processes,” *Media effects: Advances in theory and research*, vol. 2, no. 1, pp. 43–67, 2002.

[169] R. Arora and B. Ravindran, "Latent dirichlet allocation based multi-document summarization," in *Proceedings of the second workshop on Analytics for noisy unstructured text data*, 2008, pp. 91–97.

[170] M. A. Alhassan and D. R. Pennington, "Youtube as a helpful and dangerous information source for deliberate self-harming behaviours," in *International Conference on Information.* Springer, 2022, pp. 347–362.

[171] P. Alasuutari, L. Bickman, and J. Brannen, *The SAGE handbook of social research methods.* Sage, 2008.