Cyber Security and Online Safety Education for Schools in the UK: Looking through the Lens of Twitter Data[*]

Jamie Knott, Haiyue Yuan, Matthew Boakes and Shujun Li
Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, Canterbury, UK

ABSTRACT

In recent years, digital technologies have grown in many ways. As a result, many school-aged children have been exposed to the digital world a lot. Children are using more digital technologies, so schools need to teach kids more about cyber security and online safety. Because of this, there are now more school programmes and projects that teach students about cyber security and online safety and help them learn and improve their skills. Still, despite many programmes and projects, there is not much proof of how many schools have taken part and helped spread the word about them. This work shows how we can learn about the size and scope of cyber security and online safety education in schools in the UK, a country with a very active and advanced cyber security education profile, using nearly 200k public tweets from over 15k schools. By using simple techniques like descriptive statistics and visualisation as well as advanced natural language processing (NLP) techniques like sentiment analysis and topic modelling, we show some new findings and insights about how UK schools as a sector have been doing on Twitter with their cyber security and online safety education activities. Our work has led to a range of large-scale and real-world evidence that can help inform people and organisations interested in cyber security and teaching online safety in schools.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Security and privacy** → **Social aspects of security and privacy**; *Human and societal aspects of security and privacy*; • **Social and professional topics** → **K-12 education**; **Informal education**;

KEYWORDS

Cyber Security, Education, Information Retrieval, Data Analysis, Data Visualisation

ACM Reference Format:
Jamie Knott, Haiyue Yuan, Matthew Boakes and Shujun Li. 2023. Cyber Security and Online Safety Education for Schools in the UK: Looking through the Lens of Twitter Data. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27–31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3555776.3577805

1 INTRODUCTION

In 2019, the United Nations International Children's Emergency Fund (UNICEF) reported results from a survey involving more than 14,000 internet-using children across 11 countries in four continents (Europe, South America, Africa, and Asia). According to the survey results, the average internet usage was two hours per day during the week and roughly doubled the time on a weekend day [12]. Even though most people agree that technology and the internet are suitable for kids and teens, there are also risks to their safety when they use them more. There have already been some worrying statistics published [8], e.g., around 1 in 6 children aged 10 to 15 years had spoken to a stranger online in 2020, and around 1 in 10 children aged 13 to 15 years had experienced receiving sexual messages. Therefore, to protect children and young people who know how to protect themselves online, equipping them with relevant knowledge and skills in cyber security and online safety becomes very important.

Many cyber security and online safety education programmes and initiatives targeting children and young people have been launched to meet the above mentioned needs. Some nations have started including relevant content in their national curricula or guidelines. Despite all the educational activities on cyber security and online safety education, there is still little effort on how such activities are received, and a particular area with even less evidence is *to what extent schools (i.e., pre-university educational institutions) have been actively engaging and publicly promoting such activities.* This paper tries to fill this gap by exploring if and how Twitter data can be used to infer insights about schools' engagement in cyber security and online safety education, using the UK as an example country and nearly 200k public tweets from over 15k schools. Our work provides positive evidence about the data-driven approach to the research question. It produces valuable insights for researchers, practitioners, and policymakers interested in cyber security and online safety education for schools.

The rest of the paper is organised as follows: The next two sections cover related work and background, respectively. Sections 3, 4 and 5 describe our methodology, the data collection process, and the results. The last section concludes the paper.

2 RELATED WORK

Children and teens spend much time on the internet, which can expose them to online risks. This exposure has made it more critical for schools to teach cyber security and online safety. Rahman et al. [10] conducted a systematic review to state that the critical reason for having cyber security education in schools is to educate children to become aware of the associated risks of using online services such as social media, chatting, and gaming. Furthermore, Macaulay et al. [6] was surveyed to evaluate the impact of children's subjective and objective knowledge on their perception and attitudes towards

[*] The full edition of this short poster paper can be found on arXiv.org as a preprint at https://doi.org/10.48550/arXiv.2212.13742.

SAC '23, March 27–31, 2023, Tallinn, Estonia

ACM ISBN 978-1-4503-9517-5/23/03.
https://doi.org/10.1145/3555776.3577805

online safety education, concluding that online safety education is essential, especially for children lacking awareness and knowledge.

In addition, much research has been conducted to understand the current practice and status of introducing cyber security and online safety education in schools and investigate the effectiveness of related programmes and initiatives for pupils regarding their perceptions and attitudes. Regarding cyber security and online safety education, a 2022 report [13] gives a very recent and comprehensive summary based on findings from 13 countries on five continents. It finds that the current practice of embedding cyber security and online safety content into the pre-university education curriculum is either by adding the educational content to technical subjects such as computing and computer science or by adding the content to a range of non-technological subjects. Much research has not been done on how well real-world cyber security and online safety education programmes work for kids. However, much of the research looked at the effectiveness of various tools for cyber security education at the pre-university level [15].

One thing that stands out is that most of the previous studies were small-scale empirical studies that relied on the self-reported perceptions of recruited human participants. We have not seen as much research that looks at large amounts of data in the real world to study cyber security and online safety education in the real world. To the best of our knowledge, the only work similar to ours reported in this paper was done by Zenebe and Yorkman [14]. Their research aimed to find patterns and insights about how people in the USA think about cyber security education in general. While comparing our work to theirs, we focused on a different research question: how much have schools in the UK promoted cyber security and online safety education to the public? Our analysis is much more advanced regarding the size and quality of data. We used nearly 200,000 tweets from more than 15,000 verified school accounts, which were chosen from lists of all UK schools kept by the government.

3 BACKGROUND AND METHODOLOGY

To study schools' engagement and public promotion of cyber security and online safety education, we decided to use public tweets of verified schools' accounts because we observed that many schools have an active presence on Twitter and the Twitter API allowed us to gather timelines of a Twitter account quickly. In order to get a list of verified school accounts, we felt the need to focus on a single country so that we could rely on that country's education authorities to obtain official information about recognised schools.

Out of all the countries, we decided to choose the UK because it has the most active cyber security and online safety educational activities, according to some recent reports [3]. Note that our methodology is general, so it does not depend on this specific choice of country.

With the UK chosen as the country of interest, we used a five-step process to do our work: 1) gathering official information of all schools in the UK; 2) using the school information to automatically gather several Twitter accounts that may belong to a school; 3) using a semi-automatic process to identify Twitter accounts that belong to a school; 4) collecting timelines of all verified school Twitter accounts and using a data cleansing process to get tweets related to cyber security and online safety education; and 5) applying different data analytic techniques to investigate how much schools have publicly engaged and publicly promoted cyber security and online safety education on Twitter to discover valuable insights for relevant stakeholders. It is worth noting that the first four stages are all about data collection, and the last one is about the analysis of collected data. In the following section, we give details of the four data collection stages, and Section 5 shows how we conducted our data analysis and critical findings.

4 DATA COLLECTION

To collect data, we used the following steps:

1. Collecting Information about Schools: we used Google search and checked the websites of educational authorities in the UK and the four countries within the UK to identify official lists of schools in the UK from a number of sources: England (https://www.compare-school-performance.service.gov.uk/), Wales (https://gov.wales/address-list-schools), Scotland (https://www.webarchive.org.uk/wayback/archive/20150221112355/http://www.gov.scot/Topics/Statistics/Browse/School-Education/Datasets/contactdetails#), and Northern Ireland (http://apps.education-ni.gov.uk/appinstitutes/).

2. Collecting Candidate Twitter Accounts: given the name and location of a school, a two-step search algorithm was made to find the Twitter account handles automatically, resulting 19,811 unique candidates to check further.

3. Verifying Twitter Accounts of Schools: All candidate Twitter accounts were grouped into four groups according to the country the school belongs to (i.e., England, Wales, Scotland, and Northern Ireland). Semi-automatic validation was used to eliminate Twitter accounts that are not officially linked to a school. As a result, all 19,811 Twitter accounts were checked, and 12,249 were confirmed as official school Twitter accounts. The rest of the unverified Twitter accounts were checked manually, resulting in an additional 2,761 verified Twitter accounts. In total, we got 15,010 verified Twitter accounts.

4. Collecting and Cleansing Twitter Data: after obtaining the verified Twitter accounts, the Python library snscrape [4] was used to retrieve all tweets posted by these accounts from 2009 to March 2022, resulting in 20,617,709 tweets. A further data cleansing step was performed to remove images, videos and URLS, resulting 193,424 tweets ready for the last stage of our data analysis work. Finally, to clean the data for follow-up NLP operations, we used the Python-based NLP library NLTK [2] to clean the data by removing stop-words, URLs, punctuation marks, email addresses, Twitter screen names, and other non-textual data such as symbols, emoticons, and icon emojis.

5 DATA ANALYSIS AND RESULTS

Different techniques were utilised for analysing the cleaned data. Firstly, an exploratory data analysis (EDA) approach was applied to explore and visualise the data to get some initial insights. Then, NLP-based techniques such as sentiment analysis and topic modelling were applied to analyse the data further for more in-depth insights regarding how the schools covered have engaged with and publicly promoted cyber security and online safety education on Twitter.

Table 1: Some example organisations and associated initiatives with keywords we used

Organisation	Keywords	Associated Initiatives	Keywords
NCSC	NCSCgov, #NSCS, @NCSC	Cyber Aware	CyberAware, cyberawaregov
DCMS	#DCMS, @DCMS Cyber Discovery	CyberDiscovery	CyberDiscUK
BCS	BCS, @BCS, #BCS	Computing at School	CAS, CASInspire, CompAtSch, CASPLYM22

5.1 Exploratory Data Analysis

5.1.1 Twitter Activity. By observing the Twitter activity amongst schools from 2019 to 2021, we discovered the three most dominating peaks to be Tuesday in the second week of February of every year, which were the Safer Internet Days (SIDs, https://www.saferinternetday.org/) in the three years. The SID is an initiative launched by the EU SafeBorders project in 2005 and has grown significantly worldwide with the participation of approximately 200 countries and regions. While it is not surprising that SID is the most successful cyber security and online safety education initiative across UK schools, this is the first time that such large-scale evidence has been produced for this fact. Therefore, extending our work to see how schools in other countries and regions have been engaged with and publicly promoting this initiative could potentially offer valuable insights in the global context.

5.1.2 Mentions of Organisations & Initiatives. As mentioned in Section 3, many organisations in the UK have been actively running other cyber security and online safety education initiatives. Therefore, we are interested in seeing how much schools in the UK have engaged with and publicly promoted a more comprehensive range of organisations and initiatives on Twitter. To this end, we manually populated a list of keywords (e.g., organisations’ or initiatives’ names, acronyms, alternative names, Twitter handles, and hashtags) associated with key organisations and initiatives in the UK (see Table 1 for some examples). In addition, we count the number of schools whose tweets mention each keyword at least once.

By searching all tweets using the generated keywords, we summarised the mentioned frequency of each keyword. The organisations with mentions by at least 700 schools include the National Society for the Prevention of Cruelty to Children (NSPCC); Personal, Social, Health and Economic (PSHE) Association; Childline; Young Minds; National Online Safety (NOS); and UK Government’s Department for Education (DfE). Among all schools, the organisation with the most mentions is the NSPCC (19.7%), closely followed by the PSHE Association (14.8%) and the DfE (13.6%). In addition, we conducted a similar analysis using the keywords associated with selected initiatives and compared initiatives mentioned in at least 100 schools. UK schools mainly mentioned the Safer Internet Day initiative, echoing the observation of Twitter activity shown in Section 5.1.1.

Furthermore, the second most-mentioned initiative is “ThinkUKnow”, a programme organised by the Child Exploitation and Online Protection Command (CEOP) education team within the NCA. The third most-mentioned initiative among all schools is “Wake Up Wednesday”, an online safety programme run by the NOS. Moreover, it is worth noting that among all the initiatives mentioned by fewer than 100 schools, the dominating initiative is Cyber Discovery, a programme of the UK Government’s Department for Digital, Culture, Media & Sport (DCMS).

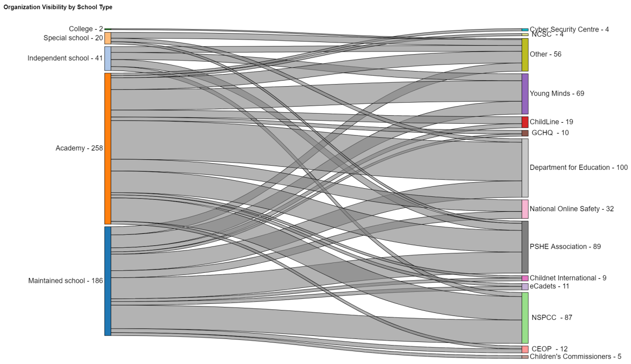

Figure 1: Organisations mentioned by different school types

5.1.3 Engagement Diversity by School Type. We also wanted to know if cyber security and online safety education have been taught and promoted differently by different types of schools. In the four UK countries, there are different types of schools. Using England as an example, there are five main types of schools according to the dataset of English schools we used: (state-)maintained schools, independent (private) schools, academies, special schools, and colleges (https://www.gov.uk/types-of-school). A sample of 1,000 verified Twitter accounts of English schools were selected randomly, and a Sankey diagram was produced to show how the different types of schools mentioned different organisations, as illustrated in Figure 1.

The results revealed significantly different engagement patterns. Both academies and maintained schools have a diverse engagement profile: they have been actively engaging with 13 and 12 organisations, respectively. However, independent schools have a much less active profile by engaging with only four organisations. One noticeable pattern is that independent schools did not post any tweets that mentioned the DfE, which may be explained by the fact that independent schools do not rely on the resources of the DfE, nor do they have to follow the national curriculum. Given the richer resources at independent schools, we found their lack of engagement worrying, and more work should be done to motivate them to do more. The remaining two types of schools have an even less active profile: special schools only engaged with three organisations, and colleges did not engage with any leading organisations shown in Figure 1. Such a noticeable discrepancy may be rooted in the different goals and interests of different types of schools.

5.2 NLP-based Content Analysis

In addition to what has been described above, we also applied two NLP-based techniques, sentiment analysis and topic modelling, to analyse the Twitter data to understand more about the content of the school's Twitter accounts.

5.2.1 Sentiment Analysis. The essential task of sentiment analysis is to classify if the sentiment status of a given text is positive, negative or neutral. Such an analysis helps analyse the emotional status of the author of a given text. It, therefore, has been widely used for analysing user-generated textual data such as tweets in different applications [1, 5]. In this study, an off-the-shelf sentiment analysis package provided by NLTK (https://www.nltk.org/api/nltk.sentiment.html) was used to conduct sentiment analysis to quantify the sentiments of tweets data to infer the schools' opinions and attitudes towards cyber security education. By analysing all tweets posted by schools, 82.2% of all tweets are classified as positive, whereas 12.8% and 5% of tweets are classified as negative and neutral, respectively.

It is not surprising to see a majority of the posted tweets with positive sentiment, as many schools' Twitter accounts are professionally managed and maintained, so it is less likely for them to publicly express negative opinions on other organisations and initiatives with a good purpose. We manually examined some negative tweets and noticed that some are misclassified because of the use of some negative-meaning words such as 'cyber bullying'. This misclassification suggests that the rate of negative sentiment is likely over-estimated, so even more tweets should be considered positive or neutral.

5.2.2 Topic Modelling. Topic modelling is an NLP technique for identifying the main topics in a collection of texts, which could help identify hidden semantic structures in a text body. It has been used frequently for analysing Twitter data in different contexts [7, 9]. In this study, we used the latent Dirichlet allocation (LDA) provided by Gensim [11], a popular topic modelling and NLP Python library, to analyse our Twitter data, aiming to identify different popular topics UK schools discussed regarding cyber security and online safety education on Twitter.

One of the critical parameters of using the LDA is to specify the number of topics. To determine this parameter, we manually examined and compared the results of topics and their associated terms generated by the LDA model using different numbers of topics. Then we decided to use six as the best number because it gives the most meaningful set of topics. Six distinct topics are extracted with manually added labels. These identified topics are closely interconnected and reflect the core concepts and the general scope of cyber security and online safety education. *Topic 1: Safety* and *Topic 2: Cyber security* are important to protect *Topic 3: Personal data* including identify, data, and health-related information. *Topic 4: Delivery methods* mainly represents how to deliver cyber security and online safety education programmes/initiatives to schools, where workshops, assemblies, and talks are amongst the most popular physical delivery methods. Many tweets aim to market or showcase cyber security and online safety events organised at school, especially those hosted by external organisations and partners such as the NOS or the NCA. *Topic 6: Delivery formats* is more related to the different formats of content delivery, often in the form of online safety posters, websites/links, guides, and newsletters. Numerous tweets aimed to share these resources with the relevant stakeholders and followers (such as parents and pupils).

6 CONCLUSION

Via a data-driven analysis of nearly 200k tweets from over 15k Twitter accounts of schools in the UK, this paper provides the first set of large-scale and real-world evidence about how schools have been engaging with and publicly promoting cyber security and online safety education. The findings can not only help evaluate how various cyber security and online safety education initiatives have been perceived by schools but also reveal areas for improvement, e.g., it seems that independent schools have not been sufficiently engaged with cyber security and online safety education initiatives despite their access to more resources.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J. Passonneau. 2011. Sentiment Analysis of Twitter Data. In *Proc. LSM 2011*. ACL, 30–38. https://aclanthology.org/W11-0705.pdf

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

[3] DCMS. 2022. *Mapping informal cyber security initiatives for young people aged 5-19 — gov.uk*. Technical Report. https://www.gov.uk/government/publications/mapping-informal-cyber-security-initiatives-for-young-people-aged-5-19/mapping-informal-cyber-security-initiatives-for-young-people-aged-5-19 [Accessed 16-Aug-2022].

[4] JustAnotherArchivist. 2022. A social networking service scraper in Python. https://github.com/JustAnotherArchivist/snscrape [Accessed 10-Aug-2022].

[5] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2021. Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (2021), 538–541. https://doi.org/10.1609/icwsm.v5i1.14185

[6] Peter J. R. Macaulay, Michael J. Boulton, Lucy R. Betts, Louise Boulton, Eleonora Camerone, James Down, Joanna Hughes, Chloe Kirkbride, and Rachel Kirkham. 2020. Subjective versus objective knowledge of online safety/dangers as predictors of children's perceived online safety and attitudes towards e-safety education in the United Kingdom. *Journal of Children and Media* 14, 3 (2020), 376–395. https://doi.org/10.1080/17482798.2019.1697716

[7] Edi Surya Negara, Dendi Triadi, and Ria Andryani. 2019. Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. In *Prof. ICECOS 2019*. IEEE, 386–390. https://doi.org/10.1109/ICECOS47637.2019.8984523

[8] Office for National Statistics, UK. 2021. *Children's online behaviour in England and Wales: year ending March 2020*. Technical report. https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/childrensonlinebehaviourinenglandandwales/yearendingmarch2020

[9] David Alfred Ostrowski. 2015. Using Latent Dirichlet Allocation for Topic Modelling in Twitter. In *Proc. ICOSC 2015*. IEEE, 493–497. https://doi.org/10.1109/ICOSC.2015.7050858

[10] N Rahman, I Sairi, NAM Zizi, and Fariza Khalid. 2020. The importance of cybersecurity education in school. *International Journal of Information and Education Technology* 10, 5 (2020), 378–382.

[11] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. LREC 2010*. ELRA, 46–50. https://radimrehurek.com/lrec2010_final.pdf

[12] Peter Stalker, Sonia Livingstone, Maria Eugenia Sozio, Kjartan Ólafsson, Petar Kanchev, and Marium Saeed. 2019. *Growing Up in a Connected World: Understanding Children's Risks and Opportunities in a Digital Age*. Technical report. UNICEF Office of Research. https://www.unicef-irc.org/growing-up-connected

[13] Krysia Emily Waldock, Vince Miller, Shujun Li, and Virginia N. L. Franqueira. 2022. *Pre-University Cyber Security Education: A report on developing cyber skills amongst children and young people*. Technical report. GFCE. https://cybilportal.org/wp-content/uploads/2022/08/GFCE-report-20220731.pdf

[14] Azene Zenebe and Tony Yorkman. 2018. Discovery of Insights on Cybersecurity Education from Twitter Using Analytics. *Journal of The Colloquium for Information Systems Security Education* 5, 2 (2018), 19. https://cisse.info/journal/index.php/cisse/article/view/81

[15] Leah Zhang-Kennedy and Sonia Chiasson. 2021. A Systematic Review of Multimedia Tools for Cybersecurity Awareness and Education. *ACM Computing Surveys* 54, 1, Article 12 (2021), 39 pages. https://doi.org/10.1145/3427920