Deep Learning Approaches to Detecting Safeguarding Concerns in Schoolchildren's Online Conversations

Emma Franklin*
Renato Software Ltd.
Nottingham, UK
e.franklin@senso.cloud

Tharindu Ranasinghe*
Aston University
Birmingham, UK
t.ranasinghe@aston.ac.uk

Abstract

For school teachers and Designated Safeguarding Leads (DSLs), computers and other school-owned communication devices are both indispensable and deeply worrisome. For their education, children require access to the internet, as well as a standard institutional ICT infrastructure, including e-mail and other forms of online communication technology. Given the sheer volume of data being generated and shared on a daily basis within schools, most teachers and DSLs can no longer monitor the safety and wellbeing of their students without the use of specialist safeguarding software. In this paper, we experiment with the use of state-of-the-art neural network models on the modelling of a dataset of almost 9,000 anonymised child-generated chat messages on the Microsoft Teams platform. The dataset was manually annotated into two binary classes: true positives (real safeguarding concerns) and false positives (false alarms) that a monitoring program would be interested in. These classes were then further annotated into eight fine-grained classes of safeguarding concerns (or false alarms). For the binary classification, we achieved a macro F1 score of 87.32, while for the fine-grained classification, our models achieved a macro F1 score of 73.56. This first experiment into the use of Deep Learning for detecting safeguarding concerns represents an important step towards achieving high-accuracy and reliable monitoring information for busy teachers and safeguarding leads.

1 Introduction

As our lives become ever more digital, traditionally "offline" activities are steadily moving online, and child safeguarding is no exception. In simpler times, it might have been enough for a schoolteacher to walk up and down a classroom to cast an eye over their pupils, or for a member of staff to oversee breaks in the playground to ensure that no bullying takes place. These days, however, children are often to be found online: when they aren't using school computers to do their work, they are reading news and social websites, watching videos, messaging one another, and sharing content. As a result, schools are now reliant on specific safeguarding technology to help monitor the online activities of their pupils.

So necessary is this technology that the UK's statutory guidance for schools and colleges on safeguarding children, *Keeping Children Safe in Education* (KCSIE)[1], heavily emphasises the dangers posed by the internet in schools and outlines the obligations of staff to ensure that appropriate web filtering and monitoring systems are in place. As a result, such systems are commonplace and are used in schools and colleges across the UK as well as abroad. KCSIE points to a range of online risks to which schools must be vigilant, ranging from harmful web content (e.g. pornography, fake news, extremism) to problematic forms of contact (e.g. online grooming, child exploitation), bad behaviour (e.g. cyberbullying, sharing of explicit images), and financial traps (e.g. online gambling, inappropriate advertising, phishing).

Given that no digital monitoring system can be perfect, and given the seriousness of child safety, human discernment is still required even for the most sophisticated risk-detecting algorithms. The output of a school's online monitoring system is typically reviewed by a Designated Safeguarding Lead (DSL) or other trusted member of staff before incidents can be triaged and acted upon. As such, it is a priority that such systems capture as many true positive cases as possible while minimising

WARNING: This paper contains offensive examples.
*The two authors contributed equally to this work.

[1]The guidance can be found online at https://www.gov.uk/government/publications/keeping-children-safe-in-education--2

Proceedings of Recent Advances in Natural Language Processing, pages 364–372
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_041

the number of false positives (i.e. noise). For a sensitively tuned safeguarding system that's geared more towards recall than precision, false positives are unavoidable, but they also represent a burden on the DSL in that they require time and energy to review and discard before real safeguarding concerns can be acted upon.

While much progress is being made in online safeguarding technology, most products are still bedevilled by the same NLP challenges faced in every other sector that utilises computational linguistics: word-sense disambiguation, parsing, coreference resolution, and sentiment analysis, just to name a few. Meanwhile, we have witnessed huge strides in NLP applications with the assistance of neural networks and other advanced machine learning techniques, the likes of which are only very recently becoming visible in the educational technology and child safeguarding sectors.

In this paper, we describe some initial experiments into applying Deep Learning (DL) techniques to the problem of online safeguarding for schoolchildren. We carry out these experiments in the hope of developing more useful and accurate safeguarding technology that will save schools time and effort and ultimately help to protect children better. In this particular case, we focus on messages sent between children on school-owned devices, specifically on the chat platform of Microsoft Teams, as captured by a keylogging cloud-based safeguarding tool, Senso.cloud. A safeguarding concern in such chat messages might be anything from bullying and discriminatory language to disclosures of self-harm and other indications of mental health risks.

The remainder of the paper is structured as follows. In Section 2, we explore some of the related work that has already been carried out, as well as the gap that we aim to address with our ongoing work. Section 3 describes the process of data collection and data annotation, followed by Section 4, which explains the use of machine learning models in our experiments. Section 5 reports on the results of our experiments, and in Section 6, we conclude the paper with a brief discussion and some comments on future work.

2 Related Work

While there is not, to our knowledge, a safeguarding study that is directly comparable with this one, we discuss in this section some examples of machine learning and deep learning in NLP generally, as well as the use of NLP for various safeguarding applications.

2.1 Machine Learning and Deep Learning in NLP

Over the years, machine learning has been widely used in NLP tasks including text classification, which we utilise in this study. Early approaches relied heavily on feature engineering combined with traditional machine learning classifiers such as Naive Bayes and support vector machines (Dadvar et al., 2013; Xu et al., 2012). More recently, neural networks such as LSTMs, bidirectional LSTMs, and GRUs combined with word embeddings have proved to outperform traditional machine learning methods in text classification (Aroyehun and Gelbukh, 2018; Modha et al., 2018).

With the recent introduction of transformer models such as BERT (Devlin et al., 2019), deep learning methods have been applied to various text classification tasks and achieved state-of-the-art results in many benchmarks. The transformer models have a transfer learning approach in which the model is pre-trained on a large number of documents and then fine-tuned to a downstream task such as text classification (Ranasinghe et al., 2019). This transfer learning strategy has provided excellent results and, consequently, the NLP community has successfully applied transformers to many tasks (Ranasinghe and Zampieri, 2020).

2.2 NLP for Safeguarding

Hatespeech, trolling, cyberaggression and cyberbullying have become the focal areas of regular shared tasks, conferences and special issues (Zampieri et al., 2020, 2019b; Satapara et al., 2023; Modha et al., 2022). There have also been recent works dedicated to the detection of mental health problems online, such as on social media (Bucur et al., 2021; Bannink et al., 2014). All of these represent useful and timely applications of machine learning methods to certain specific aspects of online safety.

Promising work has also been undertaken in automatic online grooming detection, such as Cano et al. (2014), Zuo et al. (2018) and Anderson et al. (2019); see also Borj et al. (2022). Building on this body of research, the DRAGON-S project at Swansea University seeks to utilise machine learning to identify the conversational stages that characterise an online grooming interaction and then develop an automatic groomer "spotter" tool

(Lorenzo-Dus et al., 2023). Meanwhile, the detection of online sexual predatory behaviour using DL has become the subject of an edited volume published this year (Kesavamoorthy et al., 2023).

SafeChat, a system developed by researchers at the University of Sunderland (MacFarlane and Holmes (2018); Seedall et al. (2019)), is a DL-driven chat moderation app for children that specifically seeks to prevent children from sharing inappropriate personal information (e.g. home address, or a meeting place) to mitigate threats to physical safety. Similarly, SafeToWatch is a visual threat detection solution for mobile phones, developed by SafeToNet and the Internet Watch Foundation, which utilises machine learning to recognise the generation of child sexual abuse material in real time and proactively prevent the material from being created or sent (IWF, 2023).

All of these represent important contributions to data-driven, intelligent child protection. However, each of these is focused on achieving one specific safeguarding goal, such as detecting depression or identifying conversations with online predators. Research and development that applies deep learning to generalised safeguarding, i.e. seeks to detect a range of safeguarding concerns for the benefit of teachers and DSLs, is thin on the ground. While there are commercial safeguarding systems that claim to utilise AI technology to this end, details of such systems are not (to our knowledge) made available in public-facing documents or publications.

3 Data

In this section we outline our data collection and annotation as well as ethical considerations.

3.1 Data Collection

Senso.cloud[2] is proprietary, cloud-based software used to help monitor and protect children using computers in schools. It primarily employs a key-logging approach to violation detection, which essentially matches a user's keystrokes against a set of *a priori* keyword "libraries", each one centred around a particular safeguarding concern. For example, the word *porn* will trigger a "violation" against the keyword library related to inappropriate adult content. The violation, along with its surrounding textual context, will then be logged within the Senso.cloud portal for manual review by the designated member of staff responsible for safeguarding the user who typed it.

Because Senso.cloud only logs typing activity when a violation is triggered, the only data that is available for research purposes is that which has been deemed a potential safeguarding threat. For this experiment, we drew on roughly one year's worth of historical Microsoft Teams violation log entries (student-generated messages containing one or more strings matching a Senso.cloud violation keyword), and from this secure repository took a random anonymised sample of 10,000 messages. Of these 10,000, it was found that 1,148 were not analysable as they contained only empty HTML tags (from e.g. redacted GIFs and other images); these were discarded. The remaining 8,852 were manually annotated by a safeguarding specialist according to eight fine-grained labels:

- **TP1**: an unambiguous true positive violation that requires the attention of a safeguarder, e.g. *I feel suicidal*
- **TP2**: a somewhat ambiguous true positive violation that may require the attention of a safeguarder, e.g. *I will beat u*
- **FP1**: a false positive in the sense that it is copy-pasted media rather than self-generated, e.g. explicit song lyrics, or an unfortunate news story
- **FP2**: a false positive generated by discussion of problematic or adult themes within schoolwork assignments, e.g. a debate on gun control laws
- **FP3**: a false positive as a result of sentiment polarity, e.g. *you're fucking awesome*
- **FP4**: a false positive as a result of polysemy, e.g. *I'm hardcore*
- **FP5**: a false positive as a result of foreign language interference, e.g. *je vais être en retard*
- **FP6**: a false positive as a result of violations within other words, e.g. *gunna*

A portion of the dataset underwent annotation by two annotators. We measured the inter-annotator agreement with Cohen's kappa, which was 0.83. The high inter-annotator agreement suggests that the labels are straightforward and the annotation guidelines are clear.

[2] https://senso.cloud/gb/

It should be noted that the violation data used in this paper was captured by an older version of Senso.cloud's safeguarding module, and that the figures in Table 1 do not reflect Senso.cloud's current performance on Microsoft Teams chat monitoring. This historical data was used for our machine-learning purposes only.

Binary classes	Fine-grained classes	Totals	
True Positive	TP1	3,071	4,258
	TP2	1,187	
False Positive	FP1	157	4,594
	FP2	409	
	FP3	992	
	FP4	1,252	
	FP5	194	
	FP6	1,590	

Table 1: Number of Instances in Each Class

As shown in Table 1, the eight fine-grained classes can be grouped into two binary classes: true positive and false positive. From a safeguarder's point of view, the binary classification is the one that matters the most, as it determines whether or not further action is required. The fine-grained classes are there to provide more detailed distinctions between different kinds of textual messages, so that a monitoring program might better understand the nature of a keyword violation. The classes were not predetermined, but emerged during the course of the annotation process. It is also worth noting that the classes do not each relate to a different kind of safeguarding concern (e.g. bullying, mental health), but rather the question of whether or not a safeguarding concern of any kind is suggested in the text (binary), and, further to that, the nature of the keyword violation as captured by the keylogging system (fine-grained).

In a safeguarding system, the emphasis is always on safety over precision and so it will inevitably be sensitive enough to capture false positives as well as true positives. In child protection, it is better to err on the side of caution and then filter – usually manually – the output of the software for genuine safeguarding concerns. For this reason, we expect a high number of false positives in any safeguarding system, and it is to this end that a machine-learning-assisted approach could potentially help to create a more streamlined process for teachers and DSLs.

3.2 Ethical Considerations

Any research involving input from children is inherently sensitive from an ethical standpoint. In our case, there is a considered and lawful basis, rooted in safety and the public interest, to capture only the online activities of schoolchildren that indicate a reasonable likelihood of a safeguarding risk. To protect those children's privacy, we do not analyse this data in the context of usernames, device names, or school locations.

For ethical and data protection reasons, we do not have full access to, nor can we share, the metadata of the messages in our dataset. The sensitivity of child safeguarding data is one of the key reasons that such research is difficult to conduct and to replicate, and could explain why so little of it exists in the literature for us to compare our work against.

4 Methodology

Our methodology mainly consists of two steps: data preprocessing and machine learning, which we describe in the following subsections.

4.1 Data Preprocessing

For data preprocessing, we performed data cleaning, in which we removed HTML tags related to text formatting as they do not contribute to the machine learning models. After this simple data cleaning step, we fed the data into different machine learning models, which we describe below.

4.2 Machine Learning Models

During our experimentation, we explored a range of machine learning models, spanning from simple to more sophisticated ones. For instance, we tested models like BiLSTM, which offer efficient solutions for the task at hand. We also examined complex models like transformers, which will deliver superior results but come with a trade-off in terms of computational efficiency.

SVC Our simplest machine learning model is a linear Support Vector Classifier (SVC) trained on word unigrams. Prior to the emergence of neural networks, SVCs achieved state-of-the-art results for many text classification tasks (Schwarm and Ostendorf, 2005; Goudjil et al., 2018) including offensive language identification (Zampieri et al., 2019a; Alakrot et al., 2018). Even in the neural network era, SVCs produce an efficient and effective baseline.

BiLSTM As the first embedding-based neural model, we experimented with a bidirectional Long Short-Term Memory (BiLSTM) model, which we adopted from a pre-existing model for Greek offensive language identification (Pitenis et al., 2020). The model consists of (i) an input embedding layer, (ii) two bidirectional LSTM layers, and (iii) two dense layers. The output of the final dense layer is ultimately passed through a softmax layer to produce the final prediction. The architecture diagram of the BiLSTM model is shown in Figure 1. Our BiLSTM layer has 64 units, while the first dense layer has 256 units.

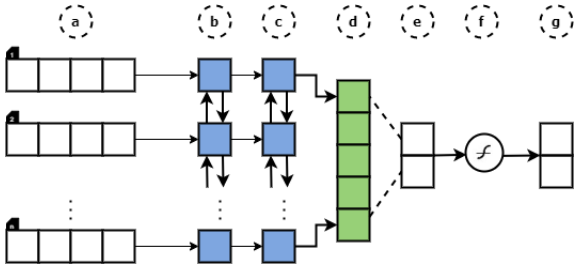

Figure 1: The BiLSTM model for sentence-level Sinhala offensive language identification. The labels are **(a)** input embeddings, **(b,c)** two BiLSTM layers, **(d, e)** fully-connected layers; **(f)** softmax activation, and **(g)** final probabilities (Ranasinghe and Zampieri, 2023)

CNN We also experimented with a convolutional neural network (CNN), which we adopted from a pre-existing model for English sentiment classification (Kim, 2014). The model consists of (i) an input embedding layer, (ii) 1 dimensional CNN layer (1DCNN), (iii) a max pooling layer and (iv) two dense layers. The output of the final dense layer is ultimately passed through a softmax layer to produce the final prediction.

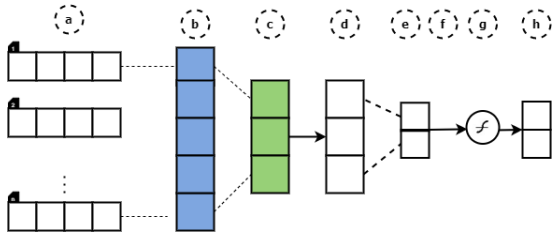

Figure 2: CNN model for sentence-level Sinhala offensive language identification. The labels are **(a)** input embeddings, **(b)** 1DCNN, **(c)** max pooling, **(d, e)** fully-connected layer; **(f)** with dropout, **(g)** softmax activation, and **(h)** final probabilities (Ranasinghe and Zampieri, 2023)

For the BiLSTM and CNN models presented above, we set three input channels for the input embedding layers: pre-trained word2vec embeddings, pre-trained fastText embeddings, and updatable embeddings learned by the model during training. For both models, we used the implementation provided in the *OffensiveNN* Python library[3].

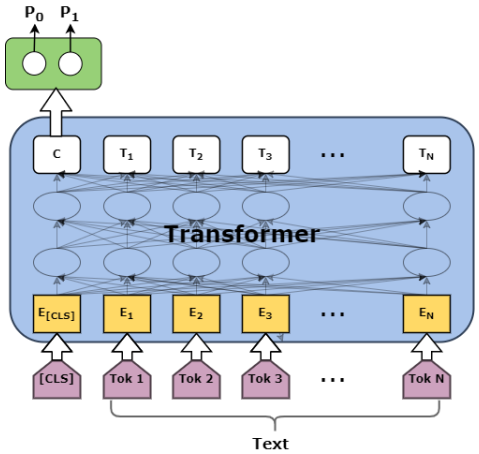

Figure 3: A schematic representation of the transformer models in classification (Uyangodage et al., 2021).

Transformers From an input sentence, transformers compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we build a classifier for the task. For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \text{softmax}(W\boldsymbol{h})$, where $W \in \mathbb{R}^{k \times d}$ is the softmax weight matrix, and k is the number of labels, which in our case is two. This architecture is depicted in Figure 3. We employed a batch size of 32, Adam optimiser with learning rate $2\text{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were evaluated while training using an evaluation set that had one-fifth of the rows in data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs. We experimented with BERT-BASE-CASED (Devlin et al., 2019), ROBERTA-BASE (Liu et al., 2019) and ELECTRA-BASE (Clark et al., 2020). All the pre-trained transformer models we used for the experiments are available in HuggingFace (Wolf et al., 2020).

[3]OffensiveNN is a pip package in https://pypi.org/project/offensivenn/

		TP			FP			Weighted			Macro F1
Type	**Model**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
SVC	-	0.65	0.46	0.55	0.70	0.81	0.73	0.64	0.65	0.65	0.63
BiLSTM	CBOW	0.71	0.76	0.74	0.80	0.76	0.81	0.75	0.74	0.73	0.76
	fastText	0.82	0.71	0.76	0.82	0.89	0.86	0.82	0.82	0.82	0.81
	Self-learned	0.66	0.34	0.45	0.66	0.88	0.76	0.66	0.66	0.63	0.60
CNN	CBOW	0.68	0.73	0.70	0.80	0.77	0.79	0.75	0.75	0.75	0.74
	fastText	0.82	0.73	0.77	0.83	0.89	0.86	0.83	0.83	0.82	0.82
	Self-learned	0.85	0.53	0.65	0.74	0.93	0.83	0.79	0.77	0.76	0.74
Transformers	BERT	0.84	0.79	0.81	0.85	0.87	0.85	0.82	0.84	0.85	0.86
	RoBERTa	0.83	0.80	0.80	0.87	0.87	0.87	0.84	0.86	0.86	**0.87**
	ELECTRA	0.81	0.83	0.79	0.85	0.86	0.85	0.83	0.83	0.83	0.82

Table 2: Results of the binary classification (Section 5.1). **Type** refers to the machine learning algorithm used, and **Model** refers to the embedding model used. We report Precision (P), Recall (R), and F1 for each model/baseline on all classes and weighted averages. Macro F1 is also listed (best in bold).

Type	**Model**	**Weighted F1**	**Macro F1**
SVC	-	0.55	0.48
BiLSTM	CBOW	0.73	0.64
	fastText	0.77	0.69
	Self-learned	0.69	0.56
CNN	CBOW	0.75	0.66
	fastText	0.77	0.68
	Self-learned	0.61	0.50
Transformers	BERT	0.79	0.72
	RoBERTa	0.81	**0.73**
	ELECTRA	0.78	0.70

Table 3: Results of the Fine-grained Classification (Section 5.2). **Type** refers to the machine learning algorithm used, and **Model** refers to the embedding model used. We report Weighted F1 and Macro F1 (best in bold).

5 Results

We show our results in two levels: binary classification in Section 5.1 and fine-grained classification in Section 5.2. For each level, we experiment with the machine learning models described in Section 4 to see how they perform. All the models were trained on the training set and then evaluated by predicting the labels for the held-out test set. As the label distribution is highly imbalanced, we evaluate and compare the performance of the different models using macro-averaged F1 score. We further report per-class Precision (P), Recall (R), and F1 score (F1) for the binary classification. We also experimented with several resampling methods to balance the classes, such as upsampling and downsampling. However, we did not see a significant improvement in the results. Therefore, we continued the experiments with the original training set distribution.

5.1 Level A - Binary Classification

As shown in Table 2, neural models outperform the traditional machine learning model, SVC. From the experimented word embedding models, fastText performed best, providing a macro F1 score of 0.82 with CNN architectures. The results suggests that the character embedding approach in fastText is effective at classifying user-generated content that contains unrecognised or improvised words, i.e. text-speak. The transformer models provided the best results. The best transformer model was RoBERTa which provided a macro F1 score of 0.87, closely followed by BERT, which provided a macro F1 score of 0.86.

The results clearly show that transformer models can successfully be used for a classification task such as this one.

5.2 Level B - Fine-grained Classification

The results for fine-grained classification are given in Table 3. Similar to the binary classification, neural models outperformed the traditional SVC model. Furthermore, transformer models produced the best macro F1 scores. As with the binary classification task, RoBERTa performed the best out of all models in the fine-grained classification.

The results of the fine-grained classification were not as good as those of the binary classification.

At the best of times, multi-class classification is a challenging task. Previous research (Zampieri et al., 2019a) has shown that multi-class classification usually performs worse than binary classification. Furthermore, in this sample, the number of instances available for some of the classes in the fine-grained classification was low, which can affect the machine learning models when predicting for that class. This can result in a low macro F1 score.

Considering both levels, we can conclude that deep learning architectures provided satisfactory results and they can be successfully utilised to detect generalised safeguarding concerns in schoolchildren's online conversations.

6 Conclusions

We have presented the first study using deep learning to detect generalised safeguarding concerns in schoolchildren's online conversations. We have developed and employed a new and highly relevant dataset consisting of more than 8,850 instances annotated on binary labels as well as fine-grained labels. We employed ten machine learning models, including state-of-the-art transformer models, on the two tasks. We showed that deep learning architectures provided the best results, and among them, the RoBERTa transformer model provided the best result. With this study, we show that machine learning and, particularly, deep-learning-based models can be employed to detect safeguarding concerns in schoolchildren's online conversations.

As for limitations, we acknowledge that the dataset is imperfect on a few fronts. For one, it is limited to the English language, and it is captured from just one app, which is Microsoft Teams chat. As a result of the data collection method via the Senso.cloud software, which nonetheless gains us access to a high volume of primary data generated by our target demographic, the data we receive is pre-filtered. That is to say, we only have access to messages that have been captured according to Senso.cloud's *a priori* safeguarding keyword libraries (an important limitation for personal data protection purposes), and as such we cannot comment on recall. It also means that there is an imbalance of data and this imbalance is reflected in the distribution across the classes. The classes themselves emerged in response to the nature of the data, and as such, they are fitted to our specific software and set of keywords. Finally, we acknowledge that the dataset is necessarily opaque, for sensitivity and proprietary reasons, as are the models developed during this industry research. At this very early stage in the work, we are not yet able to make these resources public or provide the level of detail that one would find with open-access resources. In future endeavours, we hope to find a safe and satisfactory way of doing so.

This initial study opens many exciting avenues in detecting safeguarding concerns in online conversation. In this research, we focused on English, and given that the dataset is anonymised, we cannot safely attribute each instance to a specific variety of English (e.g. British English, American English). However, the machine learning models that we explored are language-independent. In the future, we hope to evaluate these machine learning approaches in multilingual conversations. While the transformer models provided the best results, these models are large in size and computationally expensive. Therefore, it can be difficult to use them in real time. Recent work has shown that knowledge distillation can transfer knowledge from large models to computationally light models such as SVCs. In future work, we hope to build more practical models to detect safeguarding concerns in online conversations in real time.

In terms of direct, practical applications, the present research demonstrates the usefulness of pre-trained deep learning architectures in reliably identifying a concerning online message from a child, even without the wider context of the conversation. For teachers and DSLs, this translates to an intelligent system that can support them in processing the safeguarding alerts they receive daily via their school's safeguarding software. With more data and experiments, this vein of research promises to produce real-world benefits for those faced with high volumes of student safeguarding data in their day-to-day work.

Acknowledgements

We thank the anonymous RANLP reviewers who provided us with constructive feedback to improve the quality of this paper.

The computational experiments in this paper were conducted on an Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

Azalden Alakrot, Liam Murray, and Nikola S. Nikolov. 2018. Towards accurate detection of offensive language in online communication in arabic. *Procedia Computer Science*, 142:315–320. Arabic Computational Linguistics.

Philip Anderson, Zheming Zuo, Longzhi Yang, and Yanpeng Qu. 2019. An intelligent online grooming detection system using ai technologies. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rienke Bannink, Suzanne Broeren, Petra M. van de Looij – Jansen, Frouwkje G. de Waart, and Hein Raat. 2014. Cyber and traditional bullying victimization as a risk factor for mental health problems and suicidal ideation in adolescents. *PLOS ONE*, 9(4):1–7.

Parisa Rezaee Borj, Kiran Raja, and Patrick Bours. 2022. Online grooming detection: A comprehensive survey of child exploitation in chat logs. *Knowledge-Based Systems*, page 110039.

Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3600–3606, Online. Association for Computational Linguistics.

Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. 2014. Detecting child grooming behaviour patterns on social media. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings 6*, pages 412–427. Springer.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohamed Goudjil, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. 2018. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15(3):290–298.

IWF. 2023. Annual report 2022. Technical report, IWF.

R Kesavamoorthy, SP Anandaraj, TR Mahesh, V Rajesh Kumar, and Asadi Srinivasulu. 2023. Detection of online sexual predatory chats using deep learning. In *Artificial Intelligence and Blockchain in Digital Forensics*, pages 69–80. River Publishers.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Nuria Lorenzo-Dus, Craig Evans, and Ruth Mullineux-Morgan. 2023. *Online Child Sexual Grooming Discourse*. Elements in Forensic Linguistics. Cambridge University Press.

Kate MacFarlane and Violeta Holmes. 2018. Multi-agent system for safeguarding children online. In *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016: Volume 2*, pages 228–242. Springer.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2022. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 1–3, New York, NY, USA. Association for Computing Machinery.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2023. Teacher and student models of offensive language in social media. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3910–3922, Toronto, Canada. Association for Computational Linguistics.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *Working Notes of FIRE 2019-Forum for Information Retrieval Evaluation*, pages 199–207.

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2023. Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 4–7, New York, NY, USA. Association for Computing Machinery.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

Michael Seedall, Kate MacFarlane, and Violeta Holmes. 2019. Safechat system with natural language processing and deep neural networks. In *Proceedings of the 2019 Emerging Technology Conference*.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021. Can multilingual transformers fight the COVID-19 infodemic? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1432–1437, Held Online. INCOMA Ltd.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Zheming Zuo, Jie Li, Philip Anderson, Longzhi Yang, and Nitin Naik. 2018. Grooming detection using fuzzy-rough feature selection and text classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.