SafeChat System with Natural Language Processing and Deep Neural Networks

Michael Seedall
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
michael.seedall@blackburn.ac.uk

Kate MacFarlane
Faculty of Technology
University of Sunderland
Sunderland, UK
kate.macfarlane@sunderland.ac.uk

Violeta Holmes
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
v.holmes@hud.ac.uk

***Abstract*—The internet plays an ever-increasing part in the day-to-day lives of many people. Ubiquitous computing has given rise to sophisticated, streamlined and faster connections across a range of devices. Mobile smart phones are in the hands of children as young as five years old, and whilst this allows them to interact with educational applications and the wealth of information available on-line, it can put them in danger.**

There has been a consistent stream of stories involving children and adolescents being at risk because of unsafe on-line behaviour. Predators can prey on the vulnerable, by pretending to be a peer and convincing them, by charm or threats, to compromise their safety. Governments across the globe have initiatives to combat this threat, there are working groups and police task forces in place to respond to both the growing number, and impact of these incidents on children, young people, families and communities. In order to monitor on-line conversation and identify different levels of threats, the SafeChat system was designed and implemented using an ontology-based system and Natural Language Processing (NLP) techniques.

Keywords—artificial intelligence, online safety, natural language processing, deep neural networks, autonomous systems, internet security.

I. Introduction

Global governmental efforts to address the issue of child safety in an online setting continue. The Internet Taskforce on Child Protection was established in the United Kingdom in March 2001. The task force went on to release a comprehensive set of guidelines for safe practice on the internet aimed at parents and children in 2010. Whilst this was well publicised at the time, it failed to address incidents of children compromising their safety.

To check the engagement with government guidance we have carried out several surveys at the outset of our project, in 2007, research was carried out amongst 437 school children, and 37 of children surveyed said that they had arranged to meet someone they had met online [1, 6]. Subsequently, from December 2015 to March 2016 a focus group of 29 parents were asked to complete an on-line survey into online access, supervision, application usage and privacy for their children.

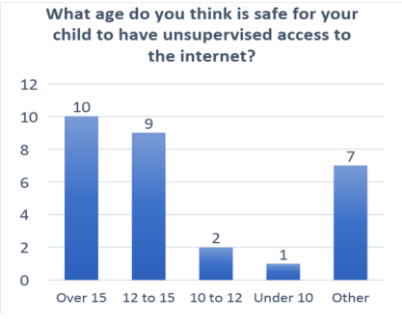


Fig. 1. Results of parent questionaire (unsupervised access)

Whilst most parents stated that they did worry about their child's safety in an online setting, as seen in figure 1, they went on to confirm that they would let their children access applications and the internet unsupervised once they reached a certain age.

The UK Government Department for Education (DFE) outlines in their 2017 guidance on child sexual exploitation that "Child sexual exploitation is a crime with devastating and long-lasting consequences for its victims and their families". In 2016/17 there had been increases in police recorded child sexual offences and indecent image offences across the UK [3]. Office of Communications (Ofcom) found that one in five 8 to 11-year old's and seven in ten 12 to 15-year old's have a social media profile. The same study also observed that in the age group of 5 to 15-year old's surveyed, 48% of children owned or used a smartphone device [3].

Whilst the number of child grooming and child online grooming cases have increased year on year [4], it could be argued that there is some correlation between the number of crimes against children online versus the continued uptake and ownership of digital devices enabling a growing online child presence.

In response to the growing trend of online child sexual exploitation the UK government introduced new legislation which brought into force section 67 of the Serious Crime Act 2015 in April of 2017. The legislation states that "It is now a criminal offence for anyone aged 18 or over to intentionally communicate with a child under 16, where the person acts for a sexual purpose and the communication is sexual or intended to elicit a sexual response. The offence applies to online and offline communication, including social media, e-mail, texts, letters." [5].

The United Nations published a revised *Convention on the Rights of the Child* [6], Article 16 defines a child's right to privacy and article 17 stipulates that children must have access to information from mass media. Governments are charged with protecting children from sexual exploitation and abduction in articles 34 and 35 respectively.

This paper presents the latest work on the SafeChat system. Recognising the additional overheads of an ontology based multi-agent system [6], coupled with the latest advances in natural language processing and machine learning techniques, current efforts focus on developing a solution using deep neural networks to recognise predator activities and identify risk behaviours to enable real time autonomous intervention in online communication mediums.

The rest of this paper is organised as follows: Section 2 outlines the latest developments in NLP frameworks, section 3 details data gathering and preparation from a variety of sources, which can be used to gather information on behaviours of both victim and perpetrators of online abuse;

Section 4 will present an analysis of initial findings of the data analysis using the latest language processing techniques and tool kits; and finally, Section 5 will present conclusions and discuss future directions for this work.

II. Natural Language Processing frameworks

Natural Language Processing (NLP) is a set of techniques and algorithms that use computers for analyzing natural human language. NLP can be used to solve a variety of problems. Some of the goals of NLP are analysis of (free) text, knowledge and abstract concept extraction from textual data (e.g. text understanding), generative models (e.g. chat bots, virtual assistants, etc.), similarity and classification of words and paragraphs, and sentiment analysis.

Early NLP systems used rules manually designed by domain experts. As the field advanced, the use of machine learning enabled the application of more powerful models that took advantage of ever-growing amounts of data. Today we are taking advantage of Deep Learning and the immense computational power of GPUs and TPUs to tackle ever more complex NLP tasks. Many different deep models have been used since their initial inception in 2000;
Deep Learning (DL):

- learns from the data,
- enables more complex reasoning and unsupervised learning,
- learns multiple levels of representation

Word embeddings is using word represented by means of its neighbors.

- Word2Vec is group of efficient predictive models (input, projection and output layers)
- Skip-Gram model and Continues Bag of Words (BoW) model.

Convolutional Neural Networks (CNN's) can be used for feature extraction of the textual data. In their paper, Shin et al [7] recognize that CNN have given state of the art performance completing sentence classification tasks. They go on to say that this is mainly due to the CNN's ability to extract local features from the data by employing convolution. Recurrent Neural Networks (RNN's) are used for time-series modelling, requiring the 'short memory' of the past, whilst Long Short-Term Memory (LSTM) networks are an extension to RNN that encapsulate long-term memory.

Often the programming language of choice for machine learning is Python. Some popular frameworks being used in NLP solutions are; Torch [8], SpaCy [9], TensorFlow [10], and Caffe2 [11].

- Torch is a scientific computing framework with wide support for machine learning algorithms that puts GPUs first.
- SpaCy is considered to be the fastest NLP (and NLP only) framework. It comes with a lot of pre-trained models to solve many problems straight out of the box.
- TensorFlow is an open-source distributed numerical computational framework released by Google, supporting efficient NLP computations on CPUs and GPUs.
- CAFFE (Convolutional Architecture for Fast Feature Embedding) is a deep learning framework that supports GPU- and CPU-based acceleration computational kernel libraries such as NVIDIA cuDNN and Intel MKL.

One of the most important issues that data scientists encounter is how to represent their data to an algorithm. This is especially relevant in NLP where inputs often differ in lengths, taking the form of sentences or even entire documents. Regardless of input length it is important to develop a representation that can capture similar themes and/or uses of domain-specific terms and vocabulary.

III. Data preparation

In order to analyse and classify on-line conversations and identify potential predatory attempt, we have acquired over 30,000 lines of predator data which has the potential to extend to over 800,000 lines of discourse once all of the predatory data has been parsed and imported. The typical raw data format is shown in Figure 2. It has to be pre-processed using parsing, to identify component parts including adding the Case Number and inserting a flag in the data to identify the predator/victim. An example of the parsed data is shown in Figure 3.

tblDataIn
RawData
jtwant2play (02/04/07 7:25:28 PM): hi
shelly_belly_93 (02/04/07 7:26:01 PM): hi

Fig. 2. Predatory data

Other digital discourse has been acquired via Twitter (104m lines), Reddit (491m lines) and the Westbury Chat Corpus (180m lines). These other sources of discourse will aid in the identification of general chat behaviour, typical acronyms/types of interaction used in digital discourse and will also aid in testing of predatory behaviour detection when predatory discourse in embedded within a general chat corpus.

From the predator discourse data, we will identify the number of questions being posed by the predator and the victim, then compare this to a comparative sized dataset from the other discourse sources / corpus. We will analyse the data to find the typical linguistics, grammar and phrases that would indicate a question being asked and whether this would aid in the detection and if such questioning prevails throughout all stages of the discourse. In order to grasp the various types of word(s) outside of a standard discourse we will use Apache Spark cluster to perform a word count on the various data sources to aid in the accuracy of question detection and possibly some typical/key indicators to consider when training the intended solution.

tblDataIn						
UserName	ChatText	CaseNumber	Type	Date	Time	CountOfDays
raidersdawg5	u never called me that night	#C2000623	P	24/09/2010	19:50:52	6
danc1njazz	yea i no im sorry	#C2000623	V	24/09/2010	19:51:29	6

Fig. 3. Parsed predatory data

Our initial findings in predator data analysis showed a distinct bias towards the predator interrogation of the victim. A comparison between a predator conversation (1800 lines)

and general chat corpus conversation (1800 lines) displayed the predator conversation had a 6% higher count of question type discourse. However, as further predator data was acquired there was a shift toward the victim interrogation of the predator. Typical lines of questioning can be the victim seeking reassurance from the predator about a type of sexual act or activity.

The predator data analysis has also revealed that some of the predators very quickly suggest a migration to different digital platform / medium to continue their discourse ie: exchange telephone numbers and text message or move from one chat platform to another, with the intention of avoiding detection or looking to increase their interaction with the victim ie: an easier exchange of Sexually Explicit Images / Video.

IV. Initial Analysis

Once the data is prepared it can then be processed using tools and techniques for feature extraction, classification and analysis. The Natural Language Tool Kit (NLTK) [12] is a fully developed platform that allows the interpretation, analysis and modelling of human language data in a Python programming environment. Given the explicit nature of the collected data, we will use other examples of online discourse to illustrate the way we intend to work with the data.

```
In [2]: from nltk.book import *

*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Fig. 4. NLTK corpus import

Using NLTK, we can store the data in logical ways and then sort them into a corpus that can be processed as a whole, or in part, depending on the results of initial testing. For example, figure 4 shows imported test corpuses which contains a Chat corpus (text5) and a Personals corpus (text8).

```
In [8]: text5.concordance("meet")

Displaying 5 of 5 matches:
ly back lol JOIN lmao U16 pleased to meet you , hope you guess my name howdy U
lol Lets make babiess !!! ok nice to meet you U64 U107 !!!! PART JOIN ((((((((
in bout willis " PART hi U30 nice to meet you lol ... U18 ahhh . U20 ! <<<<, s
ighs happily . U3 did you physically meet neysa ? a bag full o beans girl take
If you 're single , you 're going to meet the person of your dreams . If your

In [9]: text5.similar("meet")

tell me you and what take have let get see shut put be pick check ask
stop help use write
```

Fig. 5. Example of the *concordance* function in NLTK

Once the corpus has been imported there are a number of functions that we can run on the data to quickly establish initial patterns in the discourse. These are:

- Concordance: this function lists each instance of a word in the text and displays a list of sentences where it is present, see figure 5.
- Similar: is a particularly useful function that lists words used in a similar way to others, which could be key to finding patterns in discourse where someone is trying to avoid detection. It is also useful to see how one user uses language components compared to another user.
- Collocations: as seen in figure 4, this function detects the habitual juxtaposition of a particular word with another word (or words) with any regular frequency
- Lexical Dispersion Plots: these are a graphical representation of words or lists of words as they appear in the whole corpus, see figure 7.

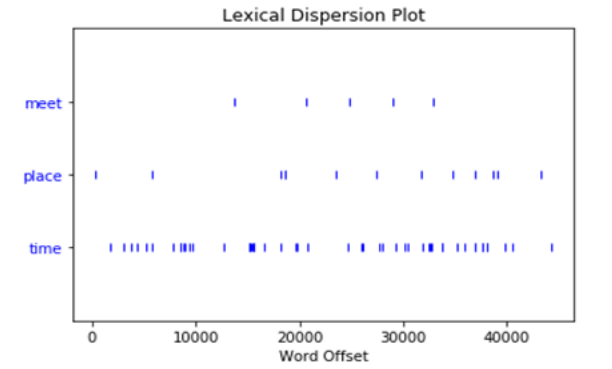
```
In [23]: text5.collocations()

wanna chat; PART JOIN; MODE #14-19teens; JOIN PART; PART PART;
cute.-ass MP3; MP3 player; JOIN JOIN; times .. .; ACTION watches; guys
wanna; song lasts; last night; ACTION sits; -...)...- S.M.R.; Lime
Player; Player 12%; dont know; lez gurls; long time

In [25]: text8.collocations()

would like; medium build; social drinker; quiet nights; non smoker;
long term; age open; Would like; easy going; financially secure; fun
times; similar interests; Age open; weekends away; poss rship; well
presented; never married; single mum; permanent relationship; slim
build
```

Fig.6. Example of the *collocations* function in NLTK

While these functions alone do not reveal rich information about the nature of the discourse, they do help to create a picture of the nature and sentiment of some of the data. Used in combination they help to build a clearer picture for possible feature extraction and classification.

Fig.7. Example of the *lexical dispersion* function in NLTK

```
text5.dispersion_plot(["meet","place","time"])
```



V. Conclusion and Future Work

Processing the data effectively is perhaps the most important factor of success in training intelligent systems. Initial findings are promising, and the next steps will be focusing on classifying and testing the data using neural networks.

Convolutional Neural Networks (CNN's) can be used for feature extraction of the textual data. Recurrent Neural Networks (RNN) have been used to good effect when preservation of context is an important factor. In the case of

grooming, context is key, so performance will be measured using both CNN and RNN, and the better system will be adopted to address other online threats, such as, cyber bullying, radicalization and fraud.

Further work will include the development and testing of natural discourse, through laboratory simulations. Multiple chat scenarios can then be tested in real time across bespoke simulated networks to test speed of response, network load and overhead wait times. This will dictate measures needed at a SafeChat system level to secure and maintain transparency of use.

The way humans interact with computers is ever changing and any long-term solutions must take these changes into consideration, potential expansions must include the development of a similar system to work with voice recognition systems. Image and video recognition will also require a similar solution, developing transparent systems to provide protection across these applications areas will present a serious challenge. Combining these systems will facilitate creation of a multi-facet tool for monitoring and detection of potentially predatory behavior in on-line conversations.

REFERENCES

[1] MacFarlane, K., and Holmes, V. (2009) Agent-Mediated Information Exchange: Child Safety On-line. In: 2009 International Conference on Management and Service Science. IEEE, pp. 1-5

[2] Ofcom, (2016) Children and Parents: Media Use and Attitudes Report. United Kingdom

[3] Bentley, H. et al (2017) How safe are our children? The most comprehensive overview of child protection in the UK 2017. London: NSPCC

[4] Gov.uk. (2017). New crackdown on child groomers comes into force - GOV.UK. [online] Available at: https://www.gov.uk/government/news/new-crackdown-on-child-groomers-comes-into-force [Accessed 15 Dec. 2017].

[5] UNICEF (2016) What is the UNCRC? | children's rights | UNICEF UK. Available at:http://www.unicef.org.uk/UNICEFs-Work/UN-Convention/ (Accessed: 12 May 2016

[6] MacFarlane, K., and Holmes, V. (2017) Multi-agent System for Safeguarding Children Online, In: Lecture Notes in Networks and Systems, Springer International, ISSN 2367-3370, Volume 16, pp. 228-2

[7] J. Shin, Y. Kim, S. Yoon and K. Jung, (2018)Contextual-CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification, *8 IEEE International Conference on Big Data and*

[8] Torch.ch. (2019). *Torch | Scientific computing for LuaJIT.*. [online] Available at: http://torch.ch/ [Accessed 10 Mar. 2019].

[9] Anon, (2019). *spaCy · Industrial-strength Natural Language Processing in Python*. [online] Available at: https://spacy.io/ [Accessed 10 Mar. 2019].

[10] TensorFlow. (2019). *TensorFlow*. [online] Available at: https://www.tensorflow.org/ [Accessed 10 Mar. 2019].

[11] Facebook Research. (2019). *Caffe2 - Facebook Research*. [online] Available at: https://research.fb.com/downloads/caffe2/ [Accessed 10 Mar. 2019]

[12] Nltk.org. (2019). *Natural Language Toolkit — NLTK 3.4 documentation*. [online] Available at: https://www.nltk.org/ [Accessed 18 Mar. 2019].