Detecting child grooming behaviour patterns on social media

Conference or Workshop Item

Detecting Child Grooming Behaviour Patterns on Social Media

Amparo Elizabeth Cano,[1] Miriam Fernandez,[1] and Harith Alani[1]

Knowledge Media Institute, Open University, UK
amparo.cano@open.ac.uk, m.fernandez@open.ac.uk,
h.alani@open.ac.uk

Abstract. Online paedophile activity in social media has become a major concern in society as Internet access is easily available to a broader younger population. One common form of online child exploitation is *child grooming*, where adults and minors exchange sexual text and media via social media platforms. Such behaviour involves a number of stages performed by a predator (adult) with the final goal of approaching a victim (minor) in person. This paper presents a study of such online grooming stages from a machine learning perspective. We propose to characterise such stages by a series of features covering sentiment polarity, content, and psycho-linguistic and discourse patterns. Our experiments with online chatroom conversations show good results in automatically classifying chatlines into various grooming stages. Such a deeper understanding and tracking of predatory behaviour is vital for building robust systems for detecting grooming conversations and potential predators on social media.

Keywords: children protection, online grooming, behavioural patterns

1 Introduction

The online exposure of children to paedophiles is one of the fastest growing issues on social media. As of March 2014, the National Society for the Prevention of Cruelty to Children (NSPCC), reported that i) 12% of 11-16 year olds in the UK have received unwanted sexual messages; and ii) 8% of 11-16 year olds in the UK have received requests to send or respond to a sexual message [16]. The detection of children cyber-sexual-offenders is therefore a critical issue which needs to be addressed.

Children in their teens have started to use social media as their main means of communication [20]. Moreover a recent study of cognition, adolescents and mobile phones (SCAMP) has revealed that 70% of 11-12 year olds in the UK now own a mobile phone rising to 90% by age 14 [28]. While social media outlets (e.g., chat-rooms, images and video sharing sites, microblogs) serve as contact points for paedophile (predators) to potentially exploit children (victims), the automatic detection of children abuse on the Web is still an open question. A common attack from paedophiles is the so-called online child grooming, where adults engage with minors via social media outlets to eventually exchange sexually explicit content. Such grooming consists of building a trust-relationship with a minor, which finally leads into convincing a child to meet them in person [19].

Previous research on detecting cyberpaedophilia online, including the efforts of the first international sexual predator identification competition (PAN'12)[11], has focused on the automatic identification of predators in chat-room logs. However little has been done on understanding predators behaviour patterns at the various stages of online child grooming, which include Deceptive Trust Development, Grooming, and Seeking for Physical Approach (Section 2). Characterising such stages is a critical issue since most of the sexually abused children have been driven to voluntarily agree to physically approach the predator [36]. This suggests that understanding the different strategies a predator uses to manipulate children behaviour could help in educating children on how to react when expose to such situations.

Moreover the early detection of such stages could facilitate the detection of malicious conversations on the Web. We believe that a deeper characterisation of predator behaviour patterns in such stages could aid in the development of more robust surveillance systems which could potentially reduce the number of abused children. This paper advances the state of the art on predator detection by proposing a more fine-grained characterisation of predators' behaviour in each of the online child grooming stages [21]. The main contributions of this paper can be summarised as follows:

(1) We propose an approach to automatically identify grooming stages in an online conversation based on multiple features: i) lexical; ii) syntactical; iii) sentiment; iv) content; v) psycho-linguistic; and vi) discourse patterns.
(2) We generate classification models for each stage, using single and multiple features. Our findings demonstrate that the use of Label discourse pattern features alone can achieve on average a gain in precision (P) of 4.63% over lexical features. While the use of combined features in classifiers consistently boost performance in P with a gain of 7.6% in all grooming stages.
(3) We present a feature analysis to identify the most discriminative features that characterise each online grooming stage.

The rest of the paper is organised as follows: Section 2 introduces Olson's theory of luring communication which characterises predator's child grooming stages. Section 3 presents related work regarding detection of online predator-victim conversations as well as previous work in online child grooming. Section 4 presents the set of features selected to characterised the language used by predators. Section 5 introduces our methodology for characterising and identifying grooming stages. Results and discussion are presented in sections 6 and 7. Conclusions are presented in Section 8.

2 Online Child Grooming Stages

Child grooming is a premeditated behaviour intending to secure the trust of a minor as a first step towards future engagement in sexual conduct [19]. One of the psychological theories which explains the different child grooming stages in the physical world is Olson's theory of luring communication (LCT) [21]. Previous research has shown that such grooming stages resemble those used by predators in online child grooming [15][9]. According to LCT, once a predator has gained access to a child, the first stage is the *Deceptive Trust Development* which consists of building a trust relationship with the minor. In this first stage a predator exchanges personal information including age,

likes, dislikes, former romances, etc. This stage allows the predator to build a common ground with the victim. In this way the predator gets information regarding the victim's support system. Once a trust relationship is established, the predator proceeds to the *Grooming stage*. In this stage the predator triggers the victim's sexual curiosity. This stage involves the use of sexual terms. In such a stage a predator is able to communicatively groom and entrap a child into online sexual conduct. Once the victim has been engaged in this stage, the so-called cycle of entrapment begins. In this cycle, the victim begins to entrust the predator. As the grooming process intensifies, the victim becomes isolated from friends and family, which promotes the predator-victim trust relationship.

In the final stage, the predator seeks to *Physically Approach* the minor. In this stage the predator requests information regarding, for example, the minor's and parent's schedules, and the minor's location. Table 1 presents extracts from the logs dataset provided by the Perverted Justice (PJ) foundation [14]. Here we can see how the different stages are represented in different sentences of the conversation. For example the sentence "I'm sorry your parents are at home all the time" indicates an intention of the predator to seek physical approach.

In the following section we present an overview of the different existing works targeting the detection of online predator-victim conversations as well as online child grooming.

STAGE	PREDATOR	VICTIM
Deceptive Trust Development	`where are you from?`	`Whats your asl?`
Grooming	`So do u masturbate?`	`not really that borin`
Seek Physical Approach	`Im sorry your parents home all the time`	`no`

Table 1. Conversation lines extracted from PJ conversations characterising the LCT child grooming stages.

3 Related Work

Online groooming detection has been widely researched in the past from both social [7] [32] and psychological perspectives [23][18] [35]. More recently the problem of predicting child-sex related solicitation conversations has started to be researched by applying data mining techniques. Simple text mining approaches have been applied to analyse paedophile activity in chat-rooms [24] [14] [15]. One of the major data sources for the automatic detection of paedophiles is the chat logs dataset provided by the Perverted Justice (PJ) foundation. In this foundation, adults volunteer to enter to chat rooms acting like minors. When a conversation involves sexual solicitation, the volunteers share the chat log with the foundation and authorities to prosecute the offenders. Those conversations that result in a predator's conviction are made available at this website.[1] Research involving the use of the PJ dataset for the detection of predators in chat-rooms includes the work of Pendar [24]. In his work he splits conversations into those of predators and those of pseudo-victims. He characterises this dataset by applying supervised (SVM) and non-parametric (kNN) classification models based on n-grams.

[1] Perverted Justice, http://www.perverted-justice.com

Kontostathis et al. [14] generate a tool which enables human annotators to tag conversation lines with child grooming stages. They consider the following four categories from Olson's theory of luring communication (LCT) [21]: Deceptive Trust Development, Grooming, Isolation, and Approach. Later on in [15] they apply a phrase-matching and rules-based approach to classify a sentence in a conversation as being related to grooming stages or not. Their results show that they can characterise non-grooming sentences with an accuracy of 75.13%. However, their work did not focus on finding out how accurately they can classify phrases to specific grooming stages.

Another study which focuses on child grooming stages is the one by Michalopoulos et al. [17]. They use a bag of words approach to characterise the stages proposed by [15], however, their goal is to detect a grooming attack rather than to characterise the particular stages within the grooming process. Their results are promising in the use of such stages as a discriminator of predator/non-predator behaviour in chat-room conversations. In [6], Escalante et al. propose a chain-based approach where the prediction of local classifiers are used as input to subsequent local classifiers with the aim of generating a predator-detection system. In their work, they use three classifiers which are applied in different segments of the conversations. Such classifiers are hypothesised to correspond to grooming stages. Based on such neural-network-based classifiers they generate a final classifier which characterises conversations as being from a predator or otherwise.

In [2], Bogdanova et al., approach the problem of discriminating cyber-sex conversations from child grooming conversations by characterising them using n-grams and high-level features. Such features include emotion, neurotism, and those proposed by Michalopoulos et al. [17]. In their task, emotion features appeared to be particularly helpful.

Our work differs from previous approaches in that, rather than characterising the predator-victim roles, we focus on characterising predators' behaviour in each of the child grooming stages. The study of grooming stages have been previously addressed by Gupta et al. [9]. They present an empirical analysis of chat-room conversations focusing on the six stages of online grooming introduced by O'Connell [22]. Their findings suggest that the relation-forming stage is more prominent than the sexual stage. However, while their study focuses on analysing online grooming stages, they do not provide an automatic classification of conversation lines into such stages. To provide such classification, our work introduces a novel set of features which pay particular attention on characterising the pyscho-linguistic and discourse patterns of the predator conversations. The complete set of features used in this work is presented in the following section.

4 Feature Engineering

In this work we use a collection of features which aim to characterise predator conversations in online grooming stages by profiling a predator based on the characterisation of: 1) bag of words (BoW); 2) syntactical; 3) sentiment polarity; 4) content; 5) psyhco-linguistic; and 6) discourse patterns.

The complete set of features is summarised in Table 2. As we can see in this table, the BoW patterns are represented using different sets of n-grams. To characterise the

Feature	Description
	Bag of Words (BoW) Patterns
N-grams	n-grams (n=1,2,3) BoW extracted from a sentence.
	Syntactical Patterns
Part-of-Speech tagging	POS tags extracted from a sentence.
	Sentiment Patterns
Sentiment Polarity	Indicates the average sentiment polarity of the terms contained in a sentence.
	Content Patterns
Complexity	Indicates the lexical complexity of a sentence. This is computed based on the cumulative entropy of the terms in a sentence (Section 4.1).
Readability	Computed following the Gunning fox index [8].
Length	Number of terms contained in a sentence.
	Psycho-linghitic Patterns
LIWC dimensions	62 dimensions caracterising psycho-linguistic patterns in English. Each dimension is composed of a collection of terms (Section 4.2).
	Discourse Patterns
Semantic Frames	Consists of a collection of over 10K words senses. This collection describes the lexical use of English in actual texts. A semantic frame can be understood as a description of a type of event, relation or entity and the participants in it (Section 4.3).

Table 2. Description of features used for characterising patterns in predator conversation lines.

syntactical patterns we extract the part of speech (POS) tags of each sentence using the Stanford POS tagger [33]. Sentiment patterns are characterised by computing the sentiment polarity of the sentences. Since peadophiles are known to suffer from emotional instability and psychological problems, [18], we include the use of sentiment polarity as a feature which could describe those changes in a predator's discourse. To compute the sentiment polarity of a sentence we use Sentistrength.[2]

The features used to characterise content, psycho-linguistic and discourse patterns are a bit more complex and will therefore be explained in more detail in the following subsections.

4.1 Content Patterns

To derive content patterns we make use of a set of features which have been successfully used in the past for modelling engagement in social media [34][29]. These features include:

[2] Sentistrength http://sentistrength.wlv.ac.uk/

– **Complexity** captures the word diversity of a sentence. The complexity C of a sentence s is defined as:

$$C(s) = \frac{1}{|W|} \sum_{w=1}^{W} f_w(log|W| - logf_w) \quad (1)$$

where W is the total number of words in the sentence and f_w is the frequency of the word w in the sentence s.
– **Readability** gauges how hard a text is to parse by humans. The readability R of a sentence s is computed based on the Gunning Fox index [8] as follows:

$$R(s) = 0.4 \left(\frac{words}{sentence} + 100 * \left(\frac{complexwords}{words} \right) \right) \quad (2)$$

– **Length** indicates the number of words in a sentence.

4.2 Psycho-linguistic Patterns

Previous work on authorship profiling [12] has shown that different groups of people writing about a particular genre use language differently. Such variations include the frequency in the use of certain words as well as the use of syntactic constructions. Authorship profiling based on such variations has been successfully used before for detecting personality features including for example neuroticism, and extraversion [12][30]. In this work we profile predator changes in the different grooming stages based on the variation of the use of different psycho-linguistic dimensions. Here we use the LIWC2007 dataset [26][25], which covers over 60 dimensions of language. These dimensions include style features like, for example, prepositions (e.g., for, beside), conjunctions (e.g., however, whereas), and cause (e.g., cuz, hence) as well as other type of dimensions relevant to psychological patterns like, for example: swearing (e.g., damn, bloody), affect (e.g.,agree, dislike), sexual(e.g., naked, porn). Each dimension is composed of a dictionary of terms. To compute the psycho-linguistic patterns appearing in a sentence we made use of the 62 dictionaries provided in LIWC [25]. To provide a representation of a sentence in these dictionaries, we propose the following approach:

LIWC

Let $LIWC_k$ be the vector representation of the k dictionary in LIWC. To calculate how close is a sentence s to this dictionary we compute the cosine similarity between the word-frequency vector representation of s and the vector $LIWC_k$. Therefore the representation of a sentence in LIWC, is a vector where each entry k corresponds to the cosine similarity of the sentence to the corresponding dictionary $LIWC_k$.

4.3 Discourse Patterns

Previous qualitative analysis [5] of PJ's predators transcripts revealed the frequent use of fixated discourse, showing the predator unwillingness to change a topic. Based on that, we believed that the use of features, which characterise the type of discourse in

a conversation could be helpful to discriminate each online grooming stage. In this work we propose to make use of the the FrameNet semantic frames [1], which incorporate semantic generalisations of a discourse. A semantic frame is a description of context in which a word sense is used. These frames consists of over 1000 patterns used in English. Such patterns include: Intentionally_Act, Causality, Grant_Permission, and Emotion_Directed.

To obtain the semantic frames of the sentences produced by a predator in a conversation we apply SEMAFOR[4]. To understand this feature type consider the semantic frame extracted from the sentence "Your mom will let you stay home?, I'm happy" in Table 3. In this sentence two semantic frames (Grant_Permission, and Emotion_Directed) are detected and for each frame different semantic roles and labels can be extracted.

Sentence A: Your mom will let you stay home?, I'm happy			Sentence B: would you sleep with a guy like that		
FRAME	SEMANTICROLE	LABEL	FRAME	SEMANTICROLE	LABEL
Grant_Permission	Target	you	Capacity	Target	sleep
	Action	stay home		Theme	with a guy like that
	Grantee	you		Entity	you
	Grantor	your mom			
	Action	stay home			
Emotion_Directed	Target	happy	People	Target	a guy
	Experiencer	I			

Table 3. Semantic frames parsed for two predator conversation sentences.

From each parsed frame we generate three types of frame-semantic derived features. In this work we propose to use this information by incorporating them as features encoded in the following way:

Frame
The frame representation of a sentence is the bag of words (BoW) of frames parsed from the sentence. The frame feature representation for sentence A is therefore, {Grant_Permission, Emotion_Directed}.

Semantic Label
The Semantic Label representation of a sentence is the BoW of Labels extracted from the Semantic Frames parsed from the sentence. The Semantic Label feature representation for sentence A is therefore:
{you, stay home, your mom, happy, I}.

FRL
This feature combines Frames, Semantic Roles, and Labels extracted from the Semantic Frames parsed from a sentence. For the cases in which a Label is composed of two or more words we include the merged separated cases. Therefore the FRL feature representation of sentence A is: {Grant_Permission-Action-stay, -Grant_Permission-Action-home, Grant_Permission-Grantee-you, ..},where
Grant_Permission-Action-stay is composed of the Frame *Grant_Permission*, the Semantic Role $Action$ and the first part of the Label *stay home*.

In Section 5, we present how the set of features introduced in this section have been used to characterise and identify online child grooming stages.

5 Characterising and Identifying Child Grooming Stages

In this work we focus on the automatic identification of the three online grooming stages described in Section 2: *Trust Development, Grooming and Approach*. Since changes on predator's discourse are stage-dependent, we propose to characterise the language model used by predators per grooming stage. In this paper, we aim to understand which are the most discriminative features in each stage. To this end, we follow a binary classification approach. We trained three different classifiers, one per stage. Each classifier assigns a stage label to a conversation sentence.

Figure 1 presents a summary of the architecture used in our proposed framework. The first step consists of extracting predator lines from the PJ chat-log conversations, described in section 5.1. Each of these lines is then preprocessed as described in subsection 5.2. Each sentence is then represented into the feature space described in Section 4. To perform feature selection we followed an information gain approach. To build the classifiers we employed a supervised discriminative model (Support Vector Machine [3]) for our experiments.

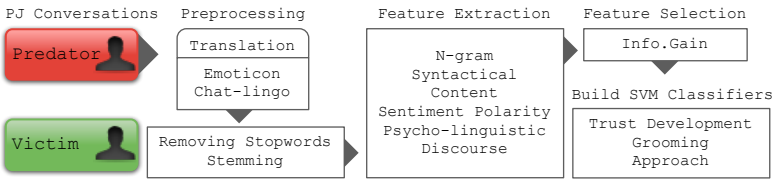

Fig. 1. Architecture for the characterisation and identification of child grooming stages.

The following subsections describe the experimental set up used in this work including: i) the description of the selected dataset, Section 5.1, ii) the data preprocessing and feature extraction phases, Section 5.2 and, iii) the construction of the different classifiers to identify grooming stages, Section 5.3.

5.1 Dataset

In this work we make use of the dataset introduced by [15]. This dataset is based on chat conversation transcripts extracted from the PJ website. The provided dataset consists of 50 transcripts corresponding to conversations between convicted predators and volunteers who posed as minors. The length of these conversations varies from 83 to over 12K lines. During the annotation process each line produced by a predator was manually labelled by two trained analysts (Media and Comunication students). Only overlapping annotations were kept as final annotations. These annotations cover four labels: 1) Trust Development; 2) Grooming; 3) Seek for physical approach (Approach) and; 4) Other. The first three describing grooming stages presented in Section 2 and the latter describing the "Other" label for sentences belonging to none of the grooming stages.[3] General

[3] Criteria provided to the annotators during the labelling process is further explained in [15]

statistics of the number of sentences labelled for each stage are presented in Table 4. There were 10,871 sentences labelled as "Other". However, since we aim to classify the language model of grooming stages, we need to have a more balanced dataset to reduce potential bias in our experiments. Therefore we randomly picked a fixed set of 3,304 sentences (highest number of sentences per grooming stages) to represent the "Other" dataset.

	Dataset			
	Trust Dev.	Grooming	Approach	Other
Sentences	1,225	3,304	2,700	3,304
After Processing	1,102	3,065	2,531	3,100

Table 4. Statistics of the datasets used for generating the classifier of each grooming stage extracted from 50 predator-victim conversations.

5.2 Data Preprocessing and Feature Extraction

One of the challenges of processing chat-room conversations is the appearance of non-standard English terms. It is common to find ill-formed words as well as chat and teen-age lingo. To overcome this issue we first generated a list of over 1,000 terms (including emoticons), which we then translated into standard English. Table 5 presents an extract of this list.

CHAT-ROOM TERM	TERM-TRANSLATION	EMOTICON	EMO-TRANSLATION
ASLP	age, sex, location, picture	:'-(	I'm crying
AWGTHTGTTA	are we going to have to go through this again?	o /\o	High Five
BRB	be right back	@_@	I'm tired, trying to stay awake
CWOT	complete waste of time	('}{')	kiss

Table 5. Extract of the over 1K terms translated into standard English.

This first stage of preprocessing resulted in our base dataset. From the base dataset we computed syntactical, psycholinguistic, and frame features. Before computing n-grams, polarity, and content features we performed the following preprocessing: i) stop-words were removed and ii) remaining words were stemmed using Porter stemmer [27].

5.3 Generation and Assessement of Grooming Stage Classifiers

For each child grooming stage we built supervised stage classifiers using the independent feature types (i.e. n-gram, syntactical, polarity, content, psycholinguistic, and semantic frames) and the merged features (All). To generate binary classifiers for each stage, the Stage-labelled sentences (i.e., sentences labeled as belonging to Trust Development, or Grooming, or Approach stages- Section 5.1) were considered as the 'positive' set, while the sentences labelled as "Other" where considered as the 'negative' set.

To assess the classification impact of features in each of the stages, we use as a baseline the performance of a stage classifier using the unigram bag of words approach

(1-gram). All the experiments reported in this paper where conducted using a 10 fold cross-validation 5 trial setting [31][13].

6 Results

In this study we report results for the performance of the supervised classifiers generated for the three online grooming stages. We also perform a feature analysis to identify the features that better characterise/discriminate the three child grooming stages.

6.1 Performance Analysis

Performance results are presented in Table 6. In all three stages the results obtained with the unigram baseline feature achieve a 100% recall while providing a precision of over 70%, and an F measure of over 80%. However although high recall values ensure good coverage of the stage-detected sentences, in this task we aim to also obtain high precision values in order to minimise the number of false positives.

When analysing the bigram and trigram features, we observe that in all three stages the use of n-gram feature representation did not improve upon the baseline in any of the performance metrics. The same trend follows for the syntactical features, which alone do not provide good classification performance. Moreover, the classification performance on all three stages drops particularly when using sentiment polarity and content features independently. This is surprising since we expected to find more verbose or complex patterns used by predators when trying to engage with minors, however this is not the case.

Figure 2 presents the distributions of such features per online grooming stage using box plots. Each box plot represents the distribution of positive and negative instances on the scale of values of a feature. For example in the case of sentiment polarity this scale goes from -1 (more negative) to 1 (more positive). The dark line within each green or red boxes represents the median, marking the mid-point of the data. We can see that in the *Trust Development* stage, the levels of complexity and length used within this stage (green box) and other-stage (red box) related conversations are very similar. While such levels slightly increase during the *Grooming* and *Approach* stages (i.e., sentences are longer and more complex).

Our results also show that sentiment polarity features alone are not good discriminators for characterising the online grooming stages. Based on Figure 2, sentiment levels are similar between positive and negative instances in the Trust Development and Grooming stages, while they present a slightly more negative polarity for the Approach stage.

Moving on to the psycho-linguistic features, we observe that, although such features alone do not improve upon the baseline, they do provide a more discrimative feature space than those discussed so far. Based on combined feature selection [10] we obtained the top 5 most discriminative LIWC dictionaries of each stage. These top features are presented in Table 7. We see that the dictionaries characterising each stage reveal patterns highlighting the mindset of a predator on each stage.

Our results also show that discourse features are good discriminators in stage classification. In particular, for the *Trust Development* stage, all discourse features alone improve precision upon the baseline. Moreover the Label discourse feature consistently

	Child Grooming Stages											
	Trust Development			***Grooming***			***Approach***			***Average***		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
	N-gram Features											
$1-gram$	0.746	1.0	0.855	0.782	1.0	0.877	0.774	1.0	0.872	0.767	1.0	0.868
$2-gram$	0.629	1.0	0.772	0.663	1.0	0.798	0.654	1.0	0.791	0.649	1.0	0.787
$3-gram$	0.561	1.0	0.719	0.578	1.0	0.733	0.574	1.0	0.73	0.571	1.0	0.727
	Syntactic Features											
POS	0.653	0.344	0.451	0.584	0.559	0.571	0.628	0.671	0.649	0.621	0.525	0.557
	Sentiment Polarity Features											
$Polarity$	0.521	0.548	0.534	0.517	0.546	0.531	0.502	0.416	0.455	0.513	0.503	0.507
	Content Features											
$Readability$	0.565	0.315	0.405	0.513	0.293	0.373	0.584	0.595	0.590	0.554	0.401	0.456
$Complexity$	0.504	0.676	0.578	0.598	0.503	0.546	0.636	0.591	0.613	0.579	0.59	0.579
$Length$	0.512	0.417	0.460	0.614	0.187	0.287	0.693	0.288	0.407	0.606	0.297	0.385
	Psycho-Linguistic Features											
$LIWC$	0.662	0.719	0.689	0.724	0.619	0.668	0.666	0.668	0.667	0.684	0.669	0.675
	Discourse Features											
$Frame$	0.752	0.228	0.350	0.753	0.368	0.494	0.769	0.342	0.474	0.758	0.313	0.439
$Label$	0.778	0.306	0.439	0.850	0.414	0.557	0.813	0.400	0.536	0.814	0.373	0.511
FRL	0.755	0.235	0.358	0.751	0.377	0.502	0.742	0.365	0.490	0.749	0.326	0.45
	All Features											
All	**0.792**	0.823	0.807	**0.876**	0.888	**0.882**	**0.872**	0.887	**0.879**	**0.847**	0.866	0.856

Table 6. Presents results for the three stages in oline child grooming. The values highlighted in bold corresponds to the best results obtained in P, R, and F measure, while the light-shaded cells indicate the best feature which alone improve P upon the BoW baseline. Significance levels: p-value < 0.01.

outperforms the baseline in precision for all three stages, providing an average boost in precision of 4.63% (t-test with $\alpha < 0.01$). Results for feature selection on the discourse features presented in Table 7 also provide an insight of the discourse patterns used in each stage, which will be further discussed in Section 6.2.

We finally trained classifiers combining all these features. Table 6 reports the best classification performance which where obtained by excluding bigrams and trigrams. We observed that although sentiment and content features alone are not good discriminators of the grooming stages they help in boosting performance when used with the rest of the features. Our results show that the combined-features classifiers do consistently outperform the baseline in precision on all three stages with an average boost of 8% (t-test with $\alpha < 0.01$) for the cost of a drop in recall of 13.3%. While the recall measure does not reach the one of the baseline on all stages, it does provide a good aver-

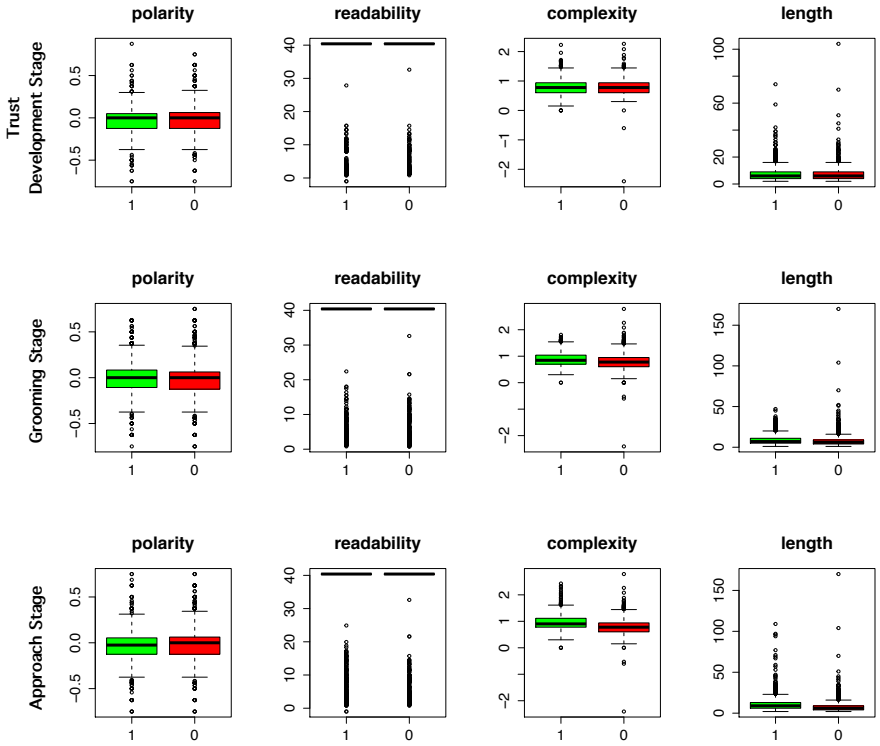

Fig. 2. Sentiment Polarity and Content Features distributions in the online grooming stages. From top to bottom, Trust Development, Grooming and Approach stages.

aged recall of 86.6%. The combined-feature classifiers also improve upon the baseline in F-measure on the *Grooming* and *Approach* stages with an average boost of 0.6%.

6.2 Feature Analysis

In this paper we focus on the characterisation of three typical online grooming stages. Each of them presenting different variations in the use of language and therefore different complexity when being modelled in a classification system. Our results show that the early stage, *Trust Development*, of the online grooming stages is more challenging when modelled using content, syntactical and sentiment features. However for all three stages the use of psycho-linguistic and discourse patterns appeared to be beneficial. Particularly the analysis of these two feature spaces facilitate the profiling of the predator discourse in each stage (see Table 7).

For the *Trust Development* stage, the top LIWC dictionaries in the psycho-linguistic profiling of the predator suggest the use of affect words (e.g, sweetheart, fun), assent (e.g.,absolutly, alright), cogMech (e.g, believe, secret) during the establisment of trust.

	Child Grooming Stages		
Feature	**Trust Development**	**Grooming**	**Approach**
LIWC Dictionaries	affect, assent, cogMech, future, home, insight, negate, ppron, see, tentant, you	assent, body, sexual,friends, death, filler, home, incl, sad, you	conj, discrep, funct, future, leisure, motion, prep, relativ, social, verbs
Frame Features	physical_artworks, similarity, coincidence, containers, desirabiity	observable_body_parts, activity_ongoing, cause_fluidic_motion, cause_to_be_wet, clothing_parts	capability, arriving, come_together, stimulus_focus, visiting
FRL	act, coincidence, evaluee, emotional_state, trust	manipulation, activity, agent, desiring, experiencer	capability, event, goal, building, stimulus, visiting
Label	artifact, picture, send, act, trust	you, cock, pussy, body_part,sex	address, afternoon, beautiful, booted, call

Table 7. Top discriminative features for each online grooming stage.

For this stage the discourse pattern features FRL and Label, highlight the request and examination of media content (e.g., pictures). These features also suggest the relevance of emotional engagement in facilitating the building of trust relationships. For the *Grooming* stage the top psycho-linguistic features profiling predators reveal for example the use of body (e.g., naked, dick), sexual (e.g., condom, orgasm), and friends (e.g, sweetie, honey) related words. Such psycholinguistic patterns are similar to those highlighted by the discourse Label feature. Moreover the FRL and frame features characterise the context of the use of such words within this stage. Finally for the *Approach* stage, top psycho-linguistic features include conj (e.g.,also, then), discrep (e.g., hopefully, must) funct (e.g., immediatly, shall) words, while the discourse features suggest the use of stimulus frames as well as temporal (e.g., event) and locative-related frames (e.g., arriving, visiting) characterising the goal of a predator to achieve physical approach with a minor.

The sentiment polarity features studied in this paper do not appear to be discriminative of the stages. However top frames in each stage, including the *emotional_state*, *desiring*, and *stimulus_focus* frames, suggest that the use of more fine-grained emotions could be useful in characterising these stages.

7 Discussion

Previous work on the qualititative characterisation of online grooming stages in chat-room conversations [9] observed that in some cases the online grooming stages are not sequential. For example a predator could convince a child to meet in person during the *Trust Development* stage. Therefore it is possible for a conversation to move back and forth between stages indicating grooming obstacles or difficulties faced by the predator. In this work we focus on the categorisation of chat-room sentences into the typical online grooming stages. The classification of individual chat-lines enables the tracking of such stages at different points on the timeline of a conversation.

While in this work we did not focus on the appearance of such stages on a timeline, it could be possible to add temporal features to characterise such back-forth changes between stages within a conversation. Also the study of short vs. long conversations might need different tactics or yield new insights. Here we studied chat-lines of the merged conversations of our dataset, however we could study chat-lines at the level of independent predator conversations in order to generate multiple predator profiling. Moreover our study is based on those conversations which lead to convicted-paedophiles, however further studies could address differences between convicted and non-convicted peadophile conversations.

In chat-rooms it is common to find regular chat-conversations with sexual content between teens or between adults. These type of conversations pose serious challenges to systems which only focus on the predator-victim characterisation since in such systems the majority of features involves sexual content. The use of stages for the characterisation of predator conversations could potentially help systems in reducing the number of false positives when exposed to non-peadophile conversations with sexual content since predators' luring stages are not common in standard online sexual conversations [2].

One of the major policing concerns is to gather accurate evidence. Therefore providing systems with a low false positive rate is fundamental. While the proposed baseline offers a 100% recall, our experiments show that the proposed discourse patterns alone and the combined-merged classifiers can boost performance at the expense of a slight drop in recall reducing in this way false positive rates.

8 Conclusions

In this work we have presented a supervised approach for the automatic classification of online grooming stages. To the best of our knowledge this is the first study focusing on the automatic classification of such stages from the psycho-linguistic and discourse patterns perspective. Such features provide an insight of the mindset, and discourse patterns of predators in online grooming stages. Our experiments show that the discourse Label feature alone consistenly outperforms our baseline in precision for all three stages. Moreover when using the combined-features classifiers our results show an improvement upon both precision and F-measure for both the *Grooming* and the *Approach* stages. Our results also show that the combined-features classifiers do consistently outperform the baseline in precision on all three stages with an average boost of 8% (t-test with $\alpha < 0.01$) for the cost of a drop in recall of 13.3%.

These results demonstrate the feasibility of the use of psycho-linguistic and discourse features for the automatic detection of online grooming stages. This opens new possibilities for adressing predator grooming behaviour online, where policing organisations can act in a preventive way by addressing grooming at early stages or in a reactive way by avoiding/intervining in the approach stage. We believe that the work in this paper has the potential to also open new possibilities into understanding the victim entrapment cycle.

Acknowledgments. This work was supported in part by the OU Policing Research Consortium in collaboration with Dorset Police, UK.

References

1. C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *In Proc. of COLING/ACL*, 1998.
2. D. Bogdanova, P. Rosso, and T. Solorio. Exploring high-level features for detecting cyberpedophilia. *Comput. Speech Lang.*, 28(1):108–120, Jan. 2014.
3. C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.
4. D. Das, N. Schneider, D. Chen, and N. A. Smith. Semafor 1.0: A probabilistic frame-semantic parser. Technical report, Carnegie Mellon University Technical Report CMU-LTI-10-001, 2010.
5. V. Egan, J. Hoskinson, and D. Shewan. Perverted justice: a content analysis of the language used by offenders detected attempting to solicit children for sex. *Antisocial Behavior: Causes, Correlations and Treatments*, 20(3):273297, 2011.
6. H. J. Escalante, L. Inaoe, L. Enrique, E. No, E. Villatoro-tello, U. Cuajimalpa, A. Jurez, L. Enrique, E. No, and L. Villaseor. Sexual predator detection in chats with chained classifiers. In *In Proc. of ACL*, 2013.
7. S. W. et al. European online grooming project - final report. Technical report, European Comission Safer Internet Plus Programme, 2012.
8. R. Gunning. *The technique of clear writing*. McGraw-Hill, International Book, 1952.
9. A. Gupta, P. Kumaraguru, and A. Sureka. Characterizing pedophile conversations on the internet using online grooming. *CoRR*, 2012.
10. M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.
11. G. Inches and F. Crestani. Overview of the international sexual predator identification competition at pan-2012. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF*, 2012.
12. C. J. K., T. P., and S.-E. N. London: Blackwell, 2004.
13. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. 14th IJCAI - Volume 2*, 1995.
14. A. Kontostathis. Chatcoder: Toward the tracking and categorization of internet predators. In *Proc. Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining . SPARKS, NV. MAY 2009.*, 2009.
15. A. Kontostathis, L. Edwards, J. Bayzick, I. Mcghee, A. Leatherman, and K. Moore. Comparison of rule-based to human analysis of chat logs. In *1st International Workshop on Mining Social Media Programme, Conferencia de la Asociación Española Para La Inteligencia Artificial, 2009*, 2010.
16. C. Lilley, R. Ball, and H. Vernon. The experiences of 11-16 year olds on social networking sites. Technical report, NSPCC, 2014.
17. D. Michalopoulos and I. Mavridis. Utilizing document classification for grooming attack recognition. In *In Proceedings of the IEEE Symposium on Computers and Communications*, 2011.
18. H. Nijman, H. Merckelbach, and M. Cima. Performance intelligence, sexual offending and psychopathy. *Journal of Sexual Aggression*, 15:319–330, 2009.
19. A. I. of Criminology (AIC). Online child grooming laws. Technical report, High tech crime brief no. 17. Canberra: AIC, 2008.
20. A. I. of Criminology (AIC). Children's use of mobile phones. Technical report, GSMA, NTT DOCOMO, 2013.
21. L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. K. K. Rogers. Entrapping the innocent: Toward a theory of child sexual predators luring communication. *Communication Theory*, 17(3):231–251, 2007.

22. R. OConnell. A typology of child cybersexploitation and online grooming practices. Technical report, Cyberspace Research Unit, University of Central Lancashier, 2003.
23. T. Palmer and S. L. Just one click - sexual abuse of children and young people through the internet and mobile phone technology. Technical report, Essex: Barnardo's UK, 2004.
24. N. Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 235–241, Sept 2007.
25. J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of liwc2007. Technical report, Technical report, Austin, TX, LIWC.Net., 2007.
26. J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count (liwc): Liwc2001 manual. Technical report, Erlbaum Publishers, 2001.
27. M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.
28. S. Project. Study of cognition, adolescents and mobile phones (scamp). Technical report, Imperial College London, 2014.
29. M. Rowe and H. Alani. Mining and comparing engagement dynamics across multiple social media platforms. In *In Proc. of ACM 2014 Web Science Conference, Bloomington, Indiana, USA*, page 229238, 2014.
30. A. S., K. M., P. J., and S. J. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119123, 2009.
31. S. L. Salzberg and U. Fayyad. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, pages 317–328, 1997.
32. G. th G and R. L. Protecting children from online sexual predators. Technical report, NSW parliamentary library brieng paper no. 10/07 Sydney: NSW Parliamentary Library, 2007.
33. K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70, 2000.
34. C. Wagner, M. Rowe, M. Strohmaier, and H. Alani. Ignorance isn't bliss: an empirical analysis of attention patterns in online communities. In *In Proc. of 4th IEEE International Conference on Social Computing, Amsterdam, The Netherlands*, 2012.
35. H. Whittle, C. Hamilton-Giachritsis, and A. Beech. Victim's voices: The impact of online grooming and sexual abuse. *Universal Journal of Psychology*, 1(2):59–71, 2013.
36. J. Wolak, K. Mitchell, and D. Finkelhor. Online victimization of youth: Five years later. Technical report, Bulleting 07-06-025, National Center for Missing and Exploted Children, Alexadia, Alexandria, VA., 2006.