Journal of Universal Computer Science, vol. 20, no. 2 (2014), 213-239
submitted: 24/10/13, accepted: 31/1/14, appeared: 1/2/14 © J.UCS

Combining Psycho-linguistic, Content-based and Chat-based Features to Detect Predation in Chatrooms

Javier Parapar
(Information Retrieval Lab, Computer Science Department
University of A Coruña, Spain
javierparapar@udc.es)

David E. Losada
(Centro Singular de Investigación en Tecnoloxías da Información (CITIUS)
Universidade de Santiago de Compostela, Spain
david.losada@usc.es)

Álvaro Barreiro
(Information Retrieval Lab, Computer Science Department
University of A Coruña, Spain
barreiro@udc.es)

Abstract: The Digital Age has brought great benefits for the human race but also some drawbacks. Nowadays, people from opposite corners of the World can communicate online via instant messaging services. Unfortunately, this has introduced new kinds of crime. Sexual predators have adapted their predatory strategies to these platforms and, usually, the target victims are kids. The authorities cannot manually track all threats because massive amounts of online conversations take place in a daily basis. Automatic methods for alerting about these crimes need to be designed. This is the main motivation of this paper, where we present a Machine Learning approach to identify suspicious subjects in chat-rooms. We propose novel types of features for representing the chatters and we evaluate different classifiers against the largest benchmark available. This empirical validation shows that our approach is promising for the identification of predatory behaviour. Furthermore, we carefully analyse the characteristics of the learnt classifiers. This preliminary analysis is a first step towards profiling the behaviour of the sexual predators when chatting on the Internet.

Key Words: Sexual predation, Cybercrime, Text Mining, Machine Learning, Support Vector Machines, Psycho-linguistic analysis
Category: H.4, H.3.0

1 Introduction

The Internet has radically changed how people communicate and interact. This has numerous advantages but also introduces new types of crime. The openness and anonymity of most online channels provides an environment that facilitates cybercrimes such as harassment, sexual predation, and other kinds of prey crimes. This is an increasingly important issue, particularly when the target subjects are under-age victims. The number of children who are approached or solicited for sexual purposes through the Internet is staggering [Mcghee et al. 2011] and, unfortunately, online sexual predators always outnumber the law enforcement officers available in police cybercrime units [Pendar 2007].

To illustrate the need to advance cybercrime detection, let us consider the case of S.K.[1], a 14-year-old Estonian teenager who committed suicide in 2009 after being recurrently harassed by a paedophile through the Internet. The paedophile pretended to be a teenager girl in order to gain access to dozens of victims. He could therefore interact with many children in a seemingly natural way. Sadly, S.K. could not bear the constant coercion from the paedophile and this soon led to his suicide. The "Myspace mom" case[2] is another tragic example: L.D. and other cyberbullies pretended to be a teenager boy on Myspace and befriended a teenager girl (M.M). After several weeks exchanging messages they abruptly ended their friendship, telling M.M. that she was cruel. Some days later M.M. committed suicide. These two cases, and many others that occur on a yearly basis across the world, are highly indicative of how severe cyberthreats are[3].

There is a need for technological solutions that can process huge amounts of data and alert about possible offences. Following this line, we make here two main contributions. First, we present an effective and efficient Machine Learning approach to identify suspicious subjects in chatrooms. More specifically, we follow a supervised learning approach that constructs subject classifiers driven by innovative sets of features. The text written by every chatter within a chatroom is important to understand his/her behaviour. We therefore extract the lines written by every individual and compute standard content-based features from this text. However, our representation of the subjects goes well beyond this. Sexual predation is intrinsically a deception activity. In the area of psycho-linguistics, there is evidence that links natural word use to personality, social and situational fluctuations, and other interventions [Pennebaker et al. 2003]. Part of speech particles, such as pronouns, articles, conjunctives or auxiliary verbs, serve as markers of emotional state and can even provide very valuable clues about deception and honesty. We utilise this type of psycho-linguistic evidence to define new features for our classifiers and show that this is a viable avenue to detect sexual predation. Furthermore, we also include additional features based on the global activity of the chatters.

A second contribution of this paper is analytical in nature. We utilise the classifiers learnt for sexual predator identification to study the characteristics that distinguish predators from other subjects. Our classification methods are naturally interpretable and, thus, permit to ascertain the behaviour of malicious Internet users when compared to regular users. This is a valuable contribution, not only to know what features are more discriminative but also to gain some insight into the tactics of the predators. We believe

[1] http://www.publico.es/espana/263683/retrato-de-una-cibervictima

[2] http://www.foxnews.com/story/2007/11/16/mom-myspace-hoax-led-to-daughter-suicide/

[3] It is not our intention to establish a casual relationship between online chatting and suicide. With these two cases, we simply exemplify that in these cases suicide was preceded by a harassing episode through online channels and, therefore, it is important to design new alert software tools.

that this preliminary analysis is important to suggest precautionary measures and to help governmental departments and non-governmental organisations when informing children about possible threats. This analysis also includes a study of the linguistic profile of the predators, relating it to findings in the area of psycho-linguistics.

The rest of this paper is organised as follows. Section 2 reports some studies related to our research and Section 3 presents our classification approach for sexual predation identification. Section 4 contains a careful analysis that gives preliminary insights into the behaviour of the predators. The paper ends with Section 5, where we expose some concluding remarks.

2 Related Work

Data mining and Machine Learning methods have been successfully applied for combating a wide array of crimes [Nissan 2012, Mena 2003]. For instance, clustering and association rules were jointly combined for discovering knowledge from massive real crime datasets [Lee and Estivill-Castro 2011]. In [Kianmehr and Alhajj 2008], Support Vector Machines were employed for predicting crime hot-spot locations. We also take a supervised learning perspective for sexual predation identification in the Internet.

Our approach for detecting sexual predators relies on three different types of features: traditional term weighting features, conversational features and psycho-linguistic features.

Term weighting features. The tf/idf term weighting scheme has been successful in different tasks since its seminal proposal for document retrieval [Spärck-Jones 1972]. In particular, tf/idf has been shown to be effective for Text Classification [Joachims 1998, Dumais et al. 1998].

Conversational features. Features related to the activity of the chat participants and other conversation-based characteristics have been recently exploited by some research teams participating in the PAN 2012's sexual predation identification task [Inches and Crestani 2012]. For instance, Morris and Hirst employed behavioural features such as the user response time, the degree of initiative of a user, and the number of conversations in which a user engages [Morris and Hirst 2012].

Psycho-linguistic features. These features have been traditionally used in tasks where detecting deceptive language is important. For instance, Ott and colleagues exploited linguistic categories for opinion spam detection [Ott et al. 2011], and Cheng and colleagues applied similar categories for author gender identification [Cheng et al. 2011]. Moreover, some PAN 2012 participants applied a psycho-linguistic approach with moderate success [Salmasi and Gillam 2012]. In [Bogdanova et al. 2012b], the authors modelled fixated discourse to detect cyberpaedophiles in chats. To meet this aim, they designed features based on the length of sex-related lexical chains within the conversations. This team of researchers also applied sentiment and emotion-based features to detect sexual predators online. These features worked well against a small dataset [Bogdanova et al. 2012a] but failed against more complex collections [Inches and Crestani 2012].

A wide range of learning strategies have been adopted for sexual predation classification in the literature. Some authors, e.g. [Parapar et al. 2012], designed a single-stage user-level classification method where all the text written by a given user (extracted from all his/her conversations) is considered as a whole. Our current paper is an extended version of [Parapar et al. 2012]. We follow here the same experimental methodology but we include a careful analysis of the best performing classifiers and the most discriminative features. Other teams [Villatoro-Tello et al. 2012, Peersman et al. 2012] applied a two-stage approach with an initial conversation-level classification that tries to filter out conversations with no sexual predation, and a subsequent predator-victim classification. The two-stage method designed in [Villatoro-Tello et al. 2012] was highly effective but the main reason behind such high performance was a pre-processing step that removed 90% of the conversations: a) conversations that had only one participant were removed, b) conversations that had less than six interventions per-user were removed, and c) conversations that had long sequences of unrecognised characters (apparently images) were removed. Such heuristic pruning was favourable for a particular experimental setting but can most likely not be used with other datasets.

Some studies related to profiling sexual predators on the Internet have been published in the literature, e.g. [Malesky 2007, Marcum 2007]. These studies, often based on manually inspecting the data, are limited to a small number of predators. Furthermore, the predators are studied in isolation with no comparative analysis of the differences between predators and regular chatters. Our analysis of the behaviour of sexual predators is also limited, mainly because of the difficulties to compile an assorted collection of personal conversations. However, we work with a large sample of chats and the characteristics of the learnt classifiers are interesting to shape a preliminary profile of sexual predation in chatrooms.

3 Automatic Classification of Chat Participants

Online conversations (e.g., in chats) are composed of a chronological sequence of textual messages written by Internet users. Every individual can be characterised by the lines or messages that he/she writes and, given some training data, sexual predator identification can be approached as a supervised learning problem. We represent the chatters (or subjects) with textual and non-textual features and apply Text Classification [Sebastiani 2002] (TC). TC works from a set of labelled examples (training data) and constructs a classifier able to predict labels for unseen examples. In many application domains, TC drives state-of-the-art solutions. For instance, many email spam classifiers perform above 90% in terms of accuracy [Androutsopoulos et al. 2000]. In the area of TC, a wide range of classifiers and textual representation methods have been proposed [Sebastiani 2002] and applying them to design solutions for finding cyberpredators is a natural and sensible choice.

	PJ	Omegle	IRCs
# conversations	8044	48569	287643
# unique users	731	7018	308662
# unique predators	396	-	-
		Training	
# conversations	2723	14571	49633
# unique users	291	2660	94744
# unique predators	142	-	-
		Test	
# conversations	5321	33998	115809
# unique users	440	4358	213918
# unique predators	254	-	-

Table 1: Main statistics of the PAN 2012 collection. This collection contains chats from Perverted Justice (PJ), Omegle and two IRC logs.

First, we describe the benchmark used for experimentation. Next, we present our classification approach: representation of the subjects, training strategy, and test.

3.1 Chat collection

The construction of a testbed for identifying predatory behaviour in online conversations is a challenge by itself. Some conversation repositories exist but only the PAN 2012 collection [Inches and Crestani 2012] contains regular chat as well as chat containing sexual predators. PAN 2012 is a large collection, with hundred of thousands of conversations, with realistic properties: a low number of true positives (conversations with a potential sexual predator), a large number of potential false positives (people talking about sex or topics that overlap with the topics discussed by predators), and many other non-predatory conversations (non-sexual topics).

Predatory conversations were taken from www.perverted-justice.com (PJ), which has been a common source for different cybercrime datasets. PJ stores logs of online conversations between convicted sexual predators and volunteers posing as under-age teenagers. The false positive set was taken from Omegle[4]. Omegle is a website that allows strangers, connected at the same time to the website, to have anonymous online conversations. This repository contains abusive language and, usually, users engage in cybersex. Other non-predatory conversations were taken from a couple of IRC logs[5] that contain a large volume of general discussions.

To make the chats comparable, the conversations were homogeneously segmented (the message exchange was cut after 25 minutes) and only the conversations with 150 or

[4] http://www.omegle.com
[5] http://www.irclog.org and http://krijnhoetmer.nl/irc-logs

fewer messages were retained. This led to a consistent collection whose main statistics are reported in Table 1. Every conversation and every user was assigned an arbitrary unique identifier. To maintain anonymity, nicknames and email addresses within the messages were replaced by arbitrary tags.

This testbed was the reference collection used in PAN 2012, which was a workshop on "Uncovering plagiarism, authorship and social software misuse" held within the Conference and Labs Evaluation Forum (CLEF). The organisers segmented the collection into a training and a testing split and we follow here the same training and test configuration in our experiments. We adopted PAN 2012's Problem 1 (identify the predators among all users in the different conversations) as our experimental task[6]. The main measures to evaluate performance were Precision, Recall, and F1 (the relevant class is here the set of sexual predators). Observe that only 30% of the collection was dedicated to training and, overall, around 0.1% of the users are sexual predators.

This benchmark contains the largest chat corpus available for experimentation. Given the difficulties to acquire such kind of data, PAN 2012 has quickly become a reference for research on sexual predation in the Internet.

3.2 Representation of the chatters

Every participant often takes part in several chat conversations and interacts with different subjects in different ways. It is therefore quite challenging to understand how to properly represent chatroom users from their interactions. Furthermore, the process of sexual predation is known to happen in phases [Mcghee et al. 2011]: gaining access, deceptive trust development, grooming, isolation, and approach. Every conversation could be classified in accordance to this categorisation and, additionally, every user-to-user interaction could be monitored to estimate what stages of predation have actually occurred. This leads to very intriguing issues related to how to extract relevant patterns of Internet sexual predation from massive amounts of chat conversations.

We are aware that these user-representation challenges are important to advance in sexual predation identification and we plan to explore them in the near future. However, in this study we approach the problem in a simpler way. For every individual, we concatenated together all the lines written by him/her in any conversation in which he/she participated. The resulting text was our document-based representation for the chat participant (i.e., one document per subject). This textual representation is recognizably simplistic but we expected that it still contained the basic clues to identify predation.

3.2.1 Features

From the document-based representations, we extracted the content-based features (tf/idf and LIWC) described below. We also included in our experiments a set of chat-based

[6] In PAN 2012, a second challenge was proposed (Problem 2), where the participants had to identify the part (the lines) of the conversations which are the most distinctive of the predator behaviour.

features, obtained from the global behaviour of the subject in the chatrooms. This acts as a complementary representation of the chatters.

- *tf/idf* features. This is a baseline representation consisting of a standard unigram representation of text. Given the characteristics of chat conversations, we decided to not apply stemming. We simply pruned the vocabulary by removing those terms appearing in 10 or fewer documents[7]. This removal of infrequent terms reduced training time without reducing effectiveness. Each term was weighted with a standard tf/idf weighting scheme [Spärck-Jones 1972]:

$$tf\_idf_{t,d} = (1 + \log(tf_{t,d})) \times \log(\frac{N}{df_t}) \quad (1)$$

  where $tf_{t,d}$ is the term frequency of the term t in the document d, N is the number of documents in the collection and df_t is the number of documents in the collection that contain t.

  We also considered bigrams and trigrams[8] and tested all the combinations of the tf/idf features: unigrams only, bigrams only, trigrams only, unigrams +bigrams, unigrams+trigrams, bigrams+trigrams, and all n-grams. For the sake of clarity, we will only report and discuss those combinations with reasonably good performance.

- *LIWC* features. Predation can arguably be discovered using Psycho-linguistic features. In the area of Psychology [Pennebaker et al. 2003], it has been shown that the words people use in their daily lives can reveal important aspects of their social and psychological worlds. We explored psychological aspects of natural language use with Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al. 2012], which is text analysis software that calculates the degree to which people use different categories of words. The ways that individuals talk and write provide windows into their emotional and cognitive worlds and can be used to analyse aspects such as deception or honesty. LIWC processes textual inputs and produces output variables such as standard linguistic dimensions, word categories tapping into psychological constructs (e.g., affect, cognition), personal concern categories (e.g., work, home, leisure), and some other dimensions (paralinguistic dimensions, punctuation categories, and general descriptor categories). Overall, there are 80 different LIWC dimensions and we processed every document in our collection (as originally written, with no modifications or preprocessing) to obtain 80 LIWC features associated to every individual (Table 2).

[7] Terms whose character size was greater than 20 were also removed.

[8] We excluded bigrams and trigrams occurring in 3 or less documents. The n-grams having a character size equal to or greater than 40 were also removed.

Category	**Abbrev**	**Examples**
Linguistic Processes		
Word count	wc	
words/sentence	wps	
Dictionary words	dic	
Words>6 letters	sixltr	
Function words	funct	
Pronouns	pronoun	I, them, itself
Personal pronouns	ppron	I, them, her
1st pers singular	i	I, me, mine
1st pers plural	we	We, us, our
2nd person	you	You, your, thou
3rd pers singular	shehe	She, her, him
3rd pers plural	they	They, their
Impersonal pronouns	ipron	It, it's, those
Articles	article	A, an, the
Common verb	verb	Walk, went, see
Auxiliary verbs	auxverb	Am, will, have
Past tense	past	Went, ran, had
Present tense	present	Is, does, hear
Future tense	future	Will, gonna
Adverbs	adverb	Very, really, quickly
Prepositions	prep	To, with, above
Conjunctions	conj	And, but,whereas
Negations	negate	No, not, never
Quantifiers	quant	Few, many, much
Numbers	number	Second, thousand
Swear words	swear	Damn, piss, fuck
Psychological Processes		
Social processes	social	Mate, talk,they, child
Family	family	Daughter,husband, aunt
Friends	friend	Buddy, friend, neighbor
Humans	human	Adult, baby, boy
Affective processes	affect	Happy, cried, abandon
Positive emotion	posemo	Love, nice, sweet
Negative emotion	negemo	Hurt, ugly,nasty
Anxiety	anx	Worried,fearful, nervous
Anger	anger	Hate, kill,annoyed

Continued on next page

Category	Abbrev	Examples
Sadness	sad	Crying, grief, sad
Cognitive processes	cogmech	cause, know, ought
Insight	insight	think, know,consider
Causation	cause	because, effect, hence
Discrepancy	discrep	should, would, could
Tentative	tentat	maybe, perhaps, guess
Certainty	certain	always, never
Inhibition	inhib	block,constrain, stop
Inclusive	incl	And, with, include
Exclusive	excl	But, without, exclude
Perceptual processes	percept	Observing, heard, feeling
See	see	View, saw, seen
Hear	hear	Listen, hearing
Feel	feel	Feels, touch
Biological processes	bio	Eat, blood, pain
Body	body	Cheek, hands, spit
Health	health	Clinic, flu, pill
Sexual	sexual	Horny, love, incest
Ingestion	ingest	Dish, eat, pizza
Relativity	relativ	Area, bend, exit, stop
Motion	motion	Arrive, car, go
Space	space	Down, in, thin
Time	time	End, until, season
Personal Concerns		
Work	work	Job, majors, xerox
Achievement	achieve	Earn, hero, win
Leisure	leisure	Cook, chat, movie
Home	home	Apartment, kitchen, family
Money	money	Audit, cash, owe
Religion	relig	Altar, church, mosque
Death	death	Bury, coffin, kill
Spoken categories		
Assent	assent	Agree, OK, yes
Nonfluencies	nonflu	Er, hm, umm
Fillers	filler	Blah, Imean, youknow

Concluded

Table 2: LIWC dimensions

The first two features, wc and wps, are the total count of the number of words and the average number of words per sentence, respectively. The rest of the features are percentages of occurrence of words from different linguistic categories (e.g., % of words in the text that are pronouns). Table 2 includes the LIWC category, an abbreviated name for the category and some examples for each LIWC dimension[9]. The list of words in the table is just illustrative (it is not the complete list of words associated to the category).

– *chat-based* features. We defined 11 additional features (Table 3) that capture some global aspects related to the activity of the individuals in chatrooms. This included features such as the number of subjects contacted by an individual, the percentage of conversations initiated by an individual, the percentage of lines written by an individual, or the average time of day when an individual chats. We expected that this innovative set of features would be indicative of how active, anxious and intense each individual is; and also indicative of the type of conversations in which individuals engage (e.g., 1-to-1 conversations, night/evening conversations). We felt that these features could reveal some trends related to predation.

3.3 Training

The PAN 2012 training collection has a large number of chatters (97689) and our approach handles a large number of features (e.g., more than 10k unigrams). We therefore decided to use LibLinear [Fan et al. 2008], which is a highly effective library for large-scale linear classification. Non-linear classifiers, e.g. SVM with Gaussian kernels, take substantially longer to train and, usually, do not provide any advantage for high dimensional text classification [Joachims 2002].

We extensively tested all the Support Vector Machines (SVMs) and Logistic Regression classifiers (with different regularisation and loss functions). SVMs were consistently better than or equal to Logistic Regression and, therefore, we only report and discuss results for SVM models[10].

The PAN 2012 classification problem is highly unbalanced: 142 out of the 97689 subjects are labelled as predators in the training collection. This introduces the risk of building meaningless classifiers that label every subject as a non-predator. Three main alternatives have been proposed in the literature to address this problem [Nallapati 2004]: oversampling the minority class (by repeating minority examples), undersampling the majority class (by removing some examples from the majority class), and adjusting the misclassification costs. Given the counts of predators and non-predators in PAN 2012, oversampling would lead to a very large training set with too many repetitions, whereas undersampling would lead to a massive removal of non-predators. We

[9] We used the complete LIWC 2007 English Dictionary with no modification.

[10] More specifically, we utilised the L2-regularised L2-loss SVM primal solver. This is option `-s 2` when running the liblinear training script (`train`).

Feature Name	Feature Description
avgLineLengthChars	Average size (chars) of the user's message lines in the collection
avgTimeOfDayOfMessages	Average time of day when every message line was sent by the user. Time of day is measured in minutes from/to midnight (the smallest amount applies)
noOfMessageLines	Number of message lines written by the user in the collection
noOfCharacters	Character count of all the message lines written by the user in the collection
noOfDifferentUsers-Approached	Number of different users approached by the user in the collection
percentOfConversations-Started	Percentage of the conversations started by the user in the collection
avgNoOfUsersInvolved-InParticipedConversations	Average number of users participating in the conversations in which the user participates
percentOfCharacters-InConversations	Percentage of the characters written by the user (computed across all the conversations in which he/she participates)
percentOfLines-InConversations	Percentage of lines written by the user (computed across all the conversations in which he/she participates)
avgTimeBetween-MessageLines	Average time, in minutes, between two consecutive message lines of the user
avgConversationTimeLength	Average conversation length, in minutes, for the user (computed across all the conversations in which he/she participates)

Table 3: Chat-level features associated to a given chat participant.

therefore decided to adjust the misclassification cost to penalise the error of classifying a sexual predator as a non-predator.

With the training collection, we applied 4-fold cross-validation and optimised F1 computed with respect to the positive class (being a predator): $F1 = \frac{2 \cdot P \cdot R}{P+R}$, where $P = TP/(TP+FP)$ and $R = TP/(TP+FN)$.

Performance was relatively insensitive to the SVM cost parameter (C) but very sensitive to the weights that adjust the relative cost of misclassifying positive and negative examples. We therefore focused on fine tuning this weighting. By default, LibLinear assigns a weight equal to 1 to every class label (i.e., $w_1 = 1$, $w_{-1} = 1$). These weights

Feature Set	P	R	F1
tf/idf(1g)	2.85	51.35	5.39
LIWC	4.79	70.95	8.97
chat-based	49.25	66.89	56.73

Table 4: Precision (P), Recall (R) and F1 (in percentage) obtained with the three feature sets when considered independently. Performance is computed with respect to the predatory class.

are multiplied by C and the resulting values are used by the SVM's optimisation process to penalise wrongly classified examples. Since we need to penalise the misclassification of positive examples, we opted for fixing w_{-1} to its default value and iteratively tuning w_1. The SVM cost parameter (C) was fixed to its default value ($C = 1$).

Given the feature sets described in subsection 3.2.1, we did not apply any feature selection strategy but simply configured a complete set of experiments combining the three sets of features. Essentially, we tested all the 1-set, 2-set and 3-set combinations of the feature sets.

For each feature set, the results reported correspond with the highest F1 run (average 4-fold cross-validation F1) obtained after tuning w_1. For the sake of clarity, we do not include the optimal w_1 in every table. The analysis of w_1 will be deferred until subsection 3.3.1.

Table 4 depicts the performance results obtained with the three sets of features (considered independently). Content-based features performed poorly: tf/idf and LIWC both yielded F1 performance lower than 10%. The performance of the chat-based features was substantially higher but it was still rather modest ($F1 = 56.73\%$). The tf/idf results were obtained with unigrams alone, tf/idf(1g)[11]. We also tested the incorporation of bigrams and/or trigrams into the tf/idf features but they did not give much added value. The main conclusion that we extracted from these initial experiments is that taking features from a single set was not enough to have reasonably good effectiveness.

Next, we tested the combination of different sets of features, including different types of n-grams for the tf/idf features. This involved extensive experimentation and validation against the training collection. We only report in Table 5 the most representative runs. All combinations of tf/idf and chat-based features performed very well. Tf/idf unigrams combined with all the other features (tf/idf(1g)+chat-based+LIWC) led also to a very consistent classification strategy. The rest of the combinations (tf/idf(1g)+LIWC, chat-based+LIWC, and tf/idf(1g,3g)+chat-based+LIWC) were clearly inferior.

Another technique that we took into account is scaling. Scaling before applying SVM is known to be important [Hsu et al. 2003]. The main advantage of scaling is to avoid having features in greater numeric range dominating those in smaller numeric

[11] In the tables we use the notation 1g, 2g or 3g to refer to the inclusion of tf/idf unigrams, tf/idf bigrams, and tf/idf trigrams, respectively.

Feature Set	P	R	F1
tf/idf(1g)+chat-based	89.15	80.99	84.87
tf/idf(1g,2g)+chat-based	91.74	78.17	84.41
tf/idf(1g,3g)+chat-based	89.68	79.58	84.33
tf/idf(1g,2g,3g)+chat-based	92.44	77.46	84.29
tf/idf(1g)+LIWC	78.99	76.76	77.86
chat-based+LIWC	45.58	66.22	53.99
tf/idf(1g)+chat-based+LIWC	87.69	80.28	83.82
tf/idf(1g,3g)+chat-based+LIWC	78.36	73.94	76.09

Table 5: Precision (P), Recall (R) and F1 (in percentage) obtained with with feature sets combining tf/idf, LIWC and chat-based. Performance is computed with respect to the predatory class.

ranges. Scaling also avoids numerical difficulties during the calculation. We therefore experimented with scaled features (in the interval [0,1])[12]. The results with scaling were rather unsatisfactory: we never obtained any substantial gain from scaling. The numeric ranges of our features are not highly diverse. This might explain why scaling was not beneficial.

3.3.1 The w_1 weight

The penalty given to positive examples that are misclassified was weighted by w_1 [13]. As recommended in [Hsu et al. 2003], we tried out a grid search approach with exponentially growing sequences of w_1: $2^{-5}, 2^{-4}, ..., 2^{10}$. Once the best w_1 in this sequence was found we conducted a finer grid search on that better region[14]. The w_1 weight was set to the value yielding the highest F1 across all these experiments. Table 6 reports the optimal w_1 weights for the most successful runs.

To further analyse the sensitivity of performance to w_1, we took the three runs that performed the best in terms of F1 -tf/idf(1g)+chat-based, tf/idf(1g)+chat-based+LIWC, and tf/idf(1g, 3g)+chat-based- and plotted how F1 performance changed with varying w_1 (Figure 1). With $w_1 < 1$ performance dropped substantially for the three methods. This is not surprising because w_{-1} was fixed to 1 and, therefore, setting w_1 lower than 1 means that we placed more importance to the correct classification of the negative examples (non-predators). This is not a good choice for our task, which aims at finding sexual predators. With w_1 between 2^2 and 2^5, tf/idf(1g)+chat-based and

[12] Either using `svm_scale` from LibLinear or applying other normalisation methods (e.g., cosine normalisation for the tf/idf features).

[13] As argued above, w_{-1} was fixed to 1 (default value) and we only experimented with varying w_1 values.

[14] For instance, after finding out that $w_1 = 8$ was optimal in the exponentially growing sequence we proceeded to test $w_1 = 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15$.

Feature Set	w_1
tf/idf(1g)+chat-based	11
tf/idf(1g,3g)+chat-based	10
tf/idf(1g)+LIWC	1
tf/idf(1g)+chat-based+LIWC	3
tf/idf(1g,3g)+chat-based+LIWC	3

Table 6: Optimal w_1 weight for the most successful runs.

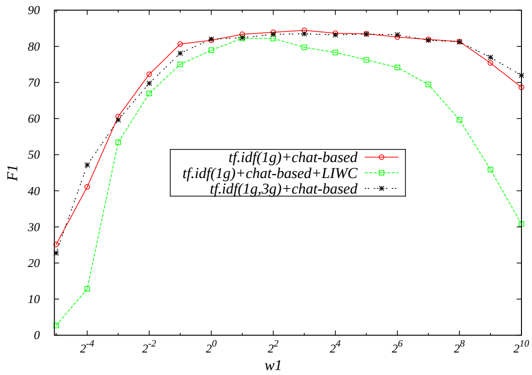

Figure 1: F1 performance with varying w_1 weights for our three best runs

tf/idf(1g,3g)+chat-based had nearly optimal performance. With $w_1 > 2^5$ performance started to fall, showing that we are giving too much emphasis on correctly classifying the positive examples. The figure also shows that tf/idf(1g)+chat-based+LIWC was more unstable than the other two methods (its performance quickly fell with $w_1 > 2^2$).

3.4 Test

The test collection contains 218702 subjects, and 254 of them are positive examples (sexual predators). The percentages of predators in the training and test collections are comparable (around 0.1%) but the absolute numbers of predators are not (254 vs 142). This introduces additional difficulties because the trained classifiers are compelled to extrapolate from few positive examples.

The performance of our best feature sets against the test collection is reported in Table 7. Again, the tf/idf(1g)+chat-based run was the best performing run. In terms of F1, tf/idf(1g,3g)+chat-based, tf/idf(1g,3g) +chat-based+LIWC, and tf/idf(1g)+chat-based+LIWC were not far from the performance obtained by the best run. A similar relative ordering of the runs had already been found with the training collection.

These results suggest that our approach is viable and detects a reasonably high number of predators. Given a massive number of chatters (218702) the best performing clas-

Feature Set	P	R	F1
tf/idf(1g)+chat-based	93.92	66.93	78.16
tf/idf(1g,3g)+chat-based	94.74	63.78	76.24
tf/idf(1g)+LIWC	78.05	62.99	69.72
tf/idf(1g)+chat-based+LIWC	90.11	64.57	75.23
tf/idf(1g,3g)+chat-based+LIWC	93.06	63.39	75.41

Table 7: Precision (P), Recall (R) and F1 (in percentage) of the selected runs against the test collection. Performance is computed with respect to the predatory class.

sifier tags 181 subjects as predators and 170 of them are true positives. This filtering would be very valuable for guiding cybercrime agencies towards potential offenders.

It seems obvious that recall is our main weakness. Comparing our training results (Table 5) against the test results (Table 7) we can clearly see that we achieved higher precision in the test collection but recall fell substantially at test time. This might have something to do with the existence of many predators in the test collection, and some of them might have distinctive characteristics that do not match the trends found for the 142 predators in the training data. This will be the subject of future research.

4 Analysing the most discriminative features

SVMs with linear kernels are a very convenient choice for analysing the classifiers and understand what features discriminate the most. From the perspective of identifying sexual predators, this analysis helps to shed light on the characteristics and cyberpredator behaviour. However, the analysis presented in this section needs to be understood as a preliminary study that is limited by the characteristics of the experimental collection. As argued above, the compilation of a large testbed of conversations is a challenge that poses important privacy implications. The PAN 2012 benchmark was carefully designed to include assorted types of chatters and has become a benchmark of reference for large-scale identification of sexual predators. However, the set of control groups is limited (e.g., there are not real children participating in these chats) and, therefore, we need to be cautious about extrapolating any conclusion beyond the range of this empirical study. Our analytical inspection of discriminative features reveals interesting trends but these findings will need to be further validated.

From a Data Analysis perspective, this study is *observational* because chat data were collected in a way that did not interfere with how the data arose. Observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a casual connection [Diez et al. 2012]. We study here the association between predictors (input features) and the binary response variable, which encodes whether the chatter is a predator. This analysis helps to understand what features or predictors distinguish predators from regular chatters. However, by no means do we claim a casual connection between the discriminative features and predatory behaviour[15].

4.1 Feature weights

The weights (w_j) of the separating hyperplane of a SVM model can be used to assess the relevance of each feature [Guyon et al. 2008, Chang and Lin 2008]. The larger $|w_j|$ is, the more important the jth feature is in the decision function of the SVM. Only linear SVM models have this indication, which naturally facilitates the analysis of the classifiers. This useful property has been used to gain knowledge of data and, for instance, to do feature selection [Brank et al. 2002, Chang and Lin 2008].

A proper and direct comparison of the weights can only be done if all features are scaled into the same range. We therefore focus on the classifier constructed from the three sets of features scaled into [0,1]. This classifier's performance was not the highest but we compared the trends reported here with those obtained with other scaled (and non scaled) classifiers and we can confirm that the relative importance of the features and the main conclusions found remained the same.

Table 8 presents the top 50 features ranked by decreasing absolute weight ($|w_j|$). Some of the tf/idf features explicitly referred to brands (e.g., the name of a business or establishment). To avoid any reference to a company, brand or trademark, these features are listed as **removed**. A positive weight ($w_j > 0$) means that high values of the feature are indicative of the membership of the individuals into the non-predatory class. In contrast, a negative weight ($w_j < 0$) means that high values of the feature are indicative of the membership of the individuals into the predatory class.

Every type of feature (chat-based, tf/idf or LIWC) contributed some features to this top 50 set. Only two chat-based features appear in this set but they are highly influential (ranks #1 and #4). The rest of the features (9 from LIWC and 39 from tf/idf) have weights in a wide range of values.

To do a proper analysis of the tactics and behaviour of the predators, we report in Tables 9 and 10 the ranked list of weights for the most influential chat-based and LIWC features, respectively. In the next subsections we individually analyse the main findings associated to every feature set.

4.1.1 Chat-based features

Table 9 presents the chat-based features. The five features missing (percentOfConversationsStarted, avgLineLengthChars, percentOfCharactersOwnedInConversations, noOfMessageLines, and noOfCharacters) had negligible weights and, therefore, are hardly

[15] To investigate the possibility of a casual connection we would need to apply a strategy based on randomisation or experimental control. This is unviable for the sexual predator case because it implies direct control of the subjects (e.g., *assigning* values of the features to randomised groups in a way which breaks possible dependencies with omitted variables and noise [Shalizi 2013]).

w_j	Feature	Set
9.08304	avgNoOfUsersInvolvedInPart...	CHAT
-2.03234	**removed**	tf/idf
1.68924	asl	tf/idf
1.3731	percentOfLinesOwnedInConv...	CHAT
-1.35502	there	tf/idf
-1.33426	ok	tf/idf
-1.30861	comming	tf/idf
-1.2002	outta	tf/idf
1.19346	m	tf/idf
1.17607	f	tf/idf
-1.13619	i	tf/idf
-1.11683	conj	LIWC
-1.05679	adress	tf/idf
-1.05179	doing	tf/idf
-1.03892	adverb	LIWC
-1.03726	awhile	tf/idf
-1.03559	around	tf/idf
-1.02502	going	tf/idf
0.987659	heyy	tf/idf
-0.98089	town	tf/idf
-0.953722	lately	tf/idf
0.85208	cause	LIWC
0.847594	my	tf/idf
-0.801311	havent	tf/idf
0.784078	like	tf/idf
-0.768532	soon	tf/idf
-0.764126	to	tf/idf
-0.752287	been	tf/idf
-0.732419	ready	tf/idf
-0.731663	directions	tf/idf
0.726415	Dic	LIWC
-0.725751	hours	tf/idf
-0.711857	be	tf/idf
-0.706	were	tf/idf
0.705927	hii	tf/idf
0.702052	dont	tf/idf
-0.70151	for	tf/idf
-0.697089	past	LIWC
-0.68416	relativ	LIWC
-0.674775	im	tf/idf
-0.673178	right	tf/idf
-0.672832	i	LIWC
-0.666966	call	tf/idf
-0.655601	hour	tf/idf
-0.653158	negate	LIWC
-0.65173	pm	tf/idf
-0.64111	miss	tf/idf
-0.63563	excl	LIWC
-0.634284	whats	tf/idf
-0.634007	at	tf/idf

Continued on right column

Table 8: List of the 50 features with the highest $|w_j|$ in the *tf/idf(1g)+chat-based+LIWC (scaled)* classifier. The features are ranked by decreasing $|w_j|$.

w_j	Feature	Set
9.08304	avgNoOfUsersInvolvedInParticipedConversations	CHAT
1.3731	percentOfLinesOwnedInConversations	CHAT
-0.595727	avgConversationTimeLength	CHAT
0.540913	avgTimeOfDayOfMessageLines	CHAT
0.308283	noOfDifferentUsersAdressed	CHAT
0.241387	avgTimeBetweenMessageLines	CHAT

Table 9: List of the most discriminative chat-based features with their weights (w_j) in the *tf/idf(1g)+chat-based+LIWC (scaled)* classifier. The features are ranked by decreasing $|w_j|$.

discriminative for sexual predator identification. In contrast, the remaining six features seem important for our classification task.

The first feature, avgNoOfUsersInvolvedInParticipedConversations, which is also the top feature in the overall list of weights (Table 8) has a positive weight that is quite high. This feature encodes the average number of chatters that participate in the conversations with a given individual. Since the weight is positive, the higher this average is, the more likely the individual is classified as a non-predator. Or, alternatively, if the chatter is always involved into one-to-one conversations then he/she has more chance of being labelled as a predator. This seems to suggest that predators tend to avoid conversations with many participants: predators might prefer to isolate the potential victims, avoiding others who might uncover the predator's actual goals. The absolute value of the weight is substantially greater than all other weights, meaning that this feature is highly important for predator classification.

The second chat-based feature, percentOfLinesOwnedInConversations, represents the percentage of lines written by a chatter (computed across all conversations in which he/she engages). This feature estimates how active a chatter is on average. The weight of the feature is positive and, therefore, a low percentage of lines written goes in favour of labelling the subject as a predator. There might be two main factors that explain this outcome. First, in predator-to-victim conversations the predator may formulate brief questions, trying to gain some knowledge about the victim and these questions might be followed by a sequence of lines where the victim elaborates on the answer. Second, the predators, in their hunt for victims, might engage in different conversations but quickly move to other chats. This may happen when they realise that there is no chance of predation (either because other users entered into the chatroom or because the user addressed does not emphatise with the predator). These factors would explain why a predator contributes on average fewer lines to the conversations when compared to a non-predator.

The feature avgConversationTimeLength captures the average duration of the conversations. The duration of a conversation is computed as the time difference between the first line and the last line of the conversation. This feature gets a negative weight, meaning that predators, on average, tend to engage in conversations that last longer. It is quite natural that the predator-to-victim conversations are long because predation is known to be a process that involves several stages [Mcghee et al. 2011]: gaining access, deceptive trust development, grooming, isolation, and approach.

The analysis of the next feature, avgTimeOfDayOfMessageLines, is also quite revealing. The feature computes the average time of day when the chatter writes his/her lines. This is computed as the minutes from/to the closest midnight (e.g., a line written at 23:10 computes as 50, a line written at 02:30 computes as 150, and a line written at 10:30 computes as 630). The feature's weight is positive. This means that non-predators tend to engage in conversations that happen not so close to midnight (high values of the feature). In contrast, predators often write their messages in the evening or at night (low values of the feature).

The analysis of the noOfDifferentUsersAdressed feature is quite intriguing. We would expect that predators, in their hunt for victims, approach a massive number of chatters. But the feature's positive weight reveals otherwise. It seems that non-predators are more active than predators at contacting with different individuals. This might have something to do with the characteristics of the collection, which also includes conversations between strangers (from the Omegle repository) who get in touch for sexual purposes (e.g., two adults to engage in cybersex). These individuals might be very active at contacting new people when searching for a good match. Furthermore, observe that predators often engage in longer conversations (feature avgConversationTimeLength). Therefore, it is plausible to think that, once they establish stable contact with a limited number of individuals, they concentrate on these conversations (rather than approaching many other chatters). Note also that the predatory process is inherently multi-stage, while an adult-to-adult contact might be simpler and direct to the point. This might explain why non-predators contact a higher number of people. These issues require further investigation and these hypotheses need to be corroborated with other sources of data.

Finally, avgTimeBetweenMessageLines represents the average time (minutes) between two consecutive lines of the chatter. This tries to capture how eager or anxious the chat participant is. The feature's positive weight reveals that non-predators tend to take more time between two consecutive messages. In contrast, predators are quicker at replying.

4.1.2 LIWC features

The LIWC feature weights (Table 10) reveal that predators present a higher usage of conjunctions, adverbs, past tense, relativity, the first person of singular, negations, exclusive particles, auxiliary verbs, motion verbs, negative emotions, social processes, time, affect words, present tense, inclusive particles, and sadness words[16]. It is quite interesting to analyse how these features relate to the objectives and stages of predation. Deceptive trust development, which is one of the phases of sexual predation, naturally leads to a higher use of words related to social processes (e.g., mate, talk, family, friends), affect words (e.g., happy, cry, abandon), and the first person of singular (I, me), because the cybercriminal is trying to bind ties with the victim. In the approach phase, when the predator aims at meeting the victim, is also natural to find higher counts of motion verbs (e.g., arrive, car, go), time words (e.g., until, end), and present tense (e.g., is, does).

In the field of Linguistics, Rayson and colleagues [Rayson et al. 2001] examined several types of spoken and writing genres and found large distributional differences between informative writing and imaginative writing. The former typically consists of more nouns, adjectives, prepositions, determiners, and conjunctions, while the latter

[16] Again, features with absolute weight below 0.1 are disregarded because their importance to the classification decision is negligible.

w_j	Feature	Set
-1.11683	conj	LIWC
-1.03892	adverb	LIWC
0.85208	cause	LIWC
0.726415	Dic	LIWC
-0.697089	past	LIWC
-0.68416	relativ	LIWC
-0.672832	i	LIWC
-0.653158	negate	LIWC
-0.63563	excl	LIWC
0.570359	verb	LIWC
0.544757	assent	LIWC
0.522944	bio	LIWC
0.514321	Sixltr	LIWC
0.510082	humans	LIWC
-0.470985	auxverb	LIWC
0.421325	cogmech	LIWC
0.402671	tentat	LIWC
-0.389815	motion	LIWC
-0.380115	negemo	LIWC
-0.370932	social	LIWC
-0.351631	time	LIWC
0.339894	article	LIWC
0.324658	death	LIWC
0.322068	percept	LIWC
-0.314585	affect	LIWC
0.312213	sexual	LIWC
-0.306323	present	LIWC
-0.300528	incl	LIWC
0.287556	funct	LIWC

Continued on right column

w_j	Feature	Set
0.278842	we	LIWC
0.264249	anger	LIWC
0.256469	body	LIWC
0.25328	Quote	LIWC
0.242801	insight	LIWC
0.206753	Colon	LIWC
0.206474	future	LIWC
0.193264	achieve	LIWC
0.190692	you	LIWC
0.177476	Parenth	LIWC
0.177426	nonfl	LIWC
0.173498	space	LIWC
0.170883	quant	LIWC
0.168805	ipron	LIWC
0.16311	leisure	LIWC
0.16205	Period	LIWC
0.160283	swear	LIWC
0.153304	feel	LIWC
0.152275	discrep	LIWC
0.142551	Comma	LIWC
0.137961	anx	LIWC
0.12101	work	LIWC
0.120065	they	LIWC
0.119163	AllPct	LIWC
0.117608	hear	LIWC
0.113561	preps	LIWC
0.107819	Apostro	LIWC
-0.101542	sad	LIWC
0.100451	Exclam	LIWC
0.0972495	shehe	LIWC

Table 10: List of the most discriminative LIWC features with their weights (w_j) in the *tf/idf(1g)+chat-based+LIWC (scaled)* classifier. The features are ranked by decreasing $|w_j|$.

consists of more verbs, adverbs, pronouns, and pre-determiners. The linguistic instruments used by predators largely overlap with an imaginative writing profile (predators use more verbs, adverbs, and personal pronouns). Perhaps, the most notable exception is the high usage of conjunctions by predators, which is not standard in imaginative writing. Still, Rayson and colleagues observed that some conjunctions, such as 'but', are more common in imaginative writing than in informative writing. Additionally, task-oriented speech (e.g., committee meetings, sermons, lectures, court proceedings) also shows an increased use of conjunctions [Rayson et al. 2001]. The predator's chat communications also have a clear intention and, linguistically, this might be reflected in a higher use of conjunctions.

In the area of Psycho-linguistics different studies analysed how deception and honesty affect the use of the language. Sexual predation largely involves deceiving victims and, therefore, we expected the predator's statements to exemplify the psychological effects of lying. The following classes of word categories have been implicated in deception: pronoun use, emotion words, markers of cognitive complexity, and motion verbs. Let us compare our results with those found in the area of psycho-linguistics:

- *Pronoun use*. There is evidence in the literature [Pennebaker et al. 2003, Mihalcea and Strapparava 2009, Newman et al. 2003, Hancock et al. 2007] that shows that liars often avoid self-references or statements of ownership. Deceptive communication is characterised by fewer first-person singular pronouns (I, me, my) because liars need to distance themselves from their stories and avoid taking responsibility for their behaviour (e.g., in a trial testimony). However, the sexual predators in our study often utilised the first person of singular (the feature "i" from LIWC has a negative weight). We hypothesize that, although predators aim to deceive the victim, they still need to develop trust (trust development stage) and this prevents them from distancing themselves from the themes that are discussed. In [Skillicorn and Lamb 2013], Skillicorn and Lamb analysed deception in interrogation settings and also found that those being deceptive show higher rates of first-person pronouns. An increased use of first-person singular pronouns by liars was also found in other domains such as opinion spam detection [Ott et al. 2011]. Deceptive opinions (e.g., about a hotel or restaurant) have a large number of first-person singular pronouns because deceivers attempt to enhance the credibility of their reviews by emphasising their own presence in the reviewed place. In a similar way, predator-to-victim conversations are intrinsically personal, leading to higher counts of first-person singular pronouns. Overall, first-person singular pronouns seem to be quite discriminative across different domains; either because of an artificially high use of first person (e.g., spam opinions or sexual predation) or because of artificially low use (e.g., trial testimony).

  Persistent utilisation of first person singular has also been related to neuroticism [Pennebaker and King 1999], anxious disposition [Weintraub 1989], and Machiavellianism [Ickes et al. 1999]. In the future, it will be interesting to study whether or not the predator's high use of first-person pronouns stems from one of these psychological states or simply comes from the personal nature of the conversations. This would help to shape the psychological profile of cyberpredators.

The cyberpredators in our study are also characterised by fewer third-person pronouns (the LIWC features "they" and "shehe" have a positive weight). This outcome is consistent with some studies in the literature where deceptive communication was associated with fewer third-person pronouns [Newman et al. 2003].

- *Emotion words*. Deception has been related to heightened anxiety and, in some cases, guilt. This leads to elevation in the use of negative emotion words during deception compared with telling the truth [Pennebaker et al. 2003, Newman et al. 2003, Vrij 2011]. In our data, sexual predators fit with an anxiety or guilt linguistic profile (higher usage of negative emotion words: negative weight for the "negemo" LIWC feature). Additionally, the predators' profile reveals an increased use of sad words (negative weight for the "sad" LIWC feature).

- *Markers of cognitive complexity*. Markers of cognitive complexity (e.g., exclusive words) have been associated with truth-telling [Pennebaker et al. 2003, Newman et al. 2003]. Exclusive words (e.g., but, except, without, and exclude) require the speaker to distinguish what is in a category from what is not in a category. In the art of deception, it is too complex to invent what was done versus what was not done and, therefore, truth-tellers often use far more exclusive words than liars do. However, cyberpredators in our collection show a high usage of exclusive words (the "excl" LIWC feature has a negative weight). This outcome requires further study. Chat communications with the aim of predation substantially differ from other scenarios where deception happens (e.g., an accused criminal testifying in a trial). These contextual differences need to be carefully studied to have a complete understanding of the pyscholinguistic profile of sexual predators.

- *Motion verbs*. Liars tend to use more motion verbs than truth-tellers [Newman et al. 2003]. False stories are fabricated and, often, some of the liar's cognitive resources need to be taken up by the effort of creating a believable story. Motion verbs (walk, go, carry) provide simple and concrete descriptions, and are more readily accessible than words that focus on evaluations and judgements (think, believe). This perfectly matches with the psycho-linguistic profile of cyberpredators in our collection, which shows a higher use of motion words and a lower use of insight words when compared to non-predators (the "motion" and "insight" LIWC features have a negative and positive weight, respectively).

4.1.3 tf/idf features

The list of tf/idf features that appear among the top 50 features (Table 8) is also quite revealing. There is a significant set of words that have a negative weight (i.e., highly used by predators) and explicitly refer to spatial or location information (there, address, around, town, directions, at, and **removed**, which refers to a explicit location). Similarly, many time words are often utilised by predators (awhile, lately, soon, hours, hour, pm). The high presence of these two classes of words confirms that time and location are important in the process of predation. In other types of texts, such as deceptive opinion reviews, explicit spatial words scarcely appear. This is because liars have difficulty encoding spatial information into spam reviews (e.g., simply because they did not visit the restaurant or hotel that is being criticised) [Ott et al. 2011].

The top 50 features also include some action words commonly used by predators, such as coming, doing, or going; and some other words, such as call or miss, which are illustrative of the way in which predators make the approach to the victim. Other tf/idf features, such as asf, m and f, are very discriminative but have a positive weight (i.e., more frequently used by non-predators). The term asl stands for age-sex-location and is employed in the chats to get basic information from another chatter. Similarly, m and f, are shorthands for male and female, respectively. The high counts of these words in non-predatory conversations must be simply due to the presence of chats, such as those extracted from Omegle, that contain plenty of conversations between adults with multiple purposes (including sexual purposes).

Finally, observe that the negative weights for the words i and im are consistent with the findings discussed above about first person pronouns.

4.2 Final remarks

The most conclusive findings of our analysis are:

- The sexual predators in our dataset have the tendency to engage into one-to-one conversations, rather than into conversations with multiple partners.
- On average, the conversations in which predators participate last longer. Predators write only a small percentage of the conversation lines but tend to react quickly (i.e., small time between consecutive messages).
- The predators' chats happen closer to midnight than the non-predators' chats.
- Predators contact a limited number of people via chat when compared to non-predators.
- The predator's linguistic profile shows an increased use of first person pronouns, negative emotion words, affect words, sadness words, time, motion and location words.
- Some parts of the predator's linguistic profile (e.g., negative emotion words, motion words) are known to be indicative of deceptive language.

– The process of predation, which includes stages such as deceptive trust development or approach, seems to have a clear influence on the linguistic style of the predators, showing a significant use of affect words, time, and location words.

5 Conclusions

In this paper we have presented effective automatic methods for detecting sexual predation in chatrooms. We have successfully shown that a learning-based method is a feasible way to approach this problem and we have proposed innovative sets of features to drive the classification of chat participants as predators or non-predators. Our experiments demonstrated that the set of features utilised and the relative weighting of the misclassification costs in the SVMs are two main factors that should be taken into account to optimise performance. Furthermore, we carefully analysed the relative importance of the classifier's features, as a preliminary effort to understand the psycholinguistic, contextual and behavioural characteristics of sexual predators in the Internet.

Our approach is promising for intelligence gathering and prioritisation of investigative resources. For instance, as a tool to assist police cybercrime units in their hunt for sexual predators in the Internet. Completely automating the process of gathering evidence to capture predators is far from reachable. However, new alert tools powered by intelligent classification technology would be very valuable to mark dangering situations that need to be monitored.

In the future, we plan to apply more evolved representations of the Internet subjects, taking into account the sequential process of predation. We also want to further study the discriminative characteristics of the predators, trying to validate our findings against other data sources, and applying alternative data analysis techniques.

Acknowledgments

This work was supported by the *"Ministerio de Economía y Competitividad"* of the Goverment of Spain under the research project TIN2012-33867.

References

[Androutsopoulos et al. 2000] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, and Constantine D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 160–167, New York, NY, USA, 2000. ACM. ISBN 1-58113-226-3.

[Bogdanova et al. 2012a] Dasha Bogdanova, Saint Petersburg, Paolo Rosso, and Thamar Solorio. On the Impact of Sentiment and Emotion Based Features in Detecting Online Sexual Predators. In *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, number July, pages 110–118, 2012a.

[Bogdanova et al. 2012b] Dasha Bogdanova, Saint Petersburg, Paolo Rosso, and Thamar Solorio. Modelling Fixated Discourse in Chats with Cyberpedophiles. In *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*, pages 86–90, 2012b.

[Brank et al. 2002] Janez Brank, Marko Grobelnik, Natasa Milic-Frayling, and Dunja Mladenic. Feature selection using linear support vector machines. *Microsoft Research Technical Report*, 5(2):187–192, 2002. ISSN 15684539.

[Chang and Lin 2008] Yin-wen Chang and Chih-jen Lin. Feature Ranking Using Linear SVM. *Journal of Machine Learning Research - Proceedings Track*, 3:53–64, 2008.

[Cheng et al. 2011] Na Cheng, R. Chandramouli, and K.P. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, July 2011. ISSN 17422876.

[Diez et al. 2012] David M. Diez, Christopher D. Barr, and Mine Çetinkaya-Rundel. *OpenIntro Statistics: Second Edition*. CreateSpace Independent Publishing Platform, 2012.

[Dumais et al. 1998] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 148–155, New York, NY, USA, 1998. ACM. ISBN 1-58113-061-9.

[Fan et al. 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008. ISSN 1532-4435.

[Guyon et al. 2008] Isabelle Guyon, Constantin F. Aliferis, Gregory F. Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander R. Statnikov. Design and analysis of the causation and prediction challenge. *Journal of Machine Learning Research - Proceedings Track*, 3:1–33, 2008.

[Hancock et al. 2007] Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, 45(1):1–23, December 2007.

[Hsu et al. 2003] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[Ickes et al. 1999] William Ickes, Susan Reidhead, and Miles Patterson. Machiavellianism and self-monitoring: as different as "me" and "you". *Social Cognition*, 4:58–74, 1999.

[Inches and Crestani 2012] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at PAN-2012. In *Proceedings of the PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse (within CLEF 2012)*, 2012.

[Joachims 2002] T. Joachims. *Learning to classify text using Suport Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

[Joachims 1998] Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK, 1998. Springer-Verlag. ISBN 3-540-64417-2.

[Kianmehr and Alhajj 2008] K. Kianmehr and R. Alhajj. Effectiveness of support vector machine for crime hot-spots prediction. *Applied Artificial Intelligence*, 22(5):433—-458, 2008.

[Lee and Estivill-Castro 2011] I. Lee and V. Estivill-Castro. Exploration of massive crime data sets through data mining techniques. *Applied Artificial Intelligence*, 25(5):362—-379, 2011.

[Malesky 2007] L Alvin Malesky. Predatory online behavior: modus operandi of convicted sex offenders in identifying potential victims and contacting minors over the internet. *Journal of Child Sexual Abuse*, 16(2):23–32, January 2007. ISSN 1053-8712.

[Marcum 2007] Catherine D Marcum. Interpreting the intentions of internet predators:an examination of online predatory behavior. *Journal of Child Sexual Abuse*, 16(4):99–114, January 2007. ISSN 1053-8712.

[Mcghee et al. 2011] India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra Mcbride, and Emma Jakubowski. Learning to identify internet sexual predation. *Int. J. Electron. Commerce*, 15(3):103–122, April 2011. ISSN 1086-4415.

[Mena 2003] J Mena. *Investigative data mining for security and criminal detection*. Boston, MA: Butterworth, 2003.

[Mihalcea and Strapparava 2009] Rada Mihalcea and Carlo Strapparava. The lie detector: explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 309–312, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Morris and Hirst 2012] Colin Morris and Graeme Hirst. Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[Nallapati 2004] Ramesh Nallapati. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 64–71, New York, NY, USA, 2004. ACM.

[Newman et al. 2003] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 2003.

[Nissan 2012] E. Nissan. An overview of data mining for combating crime. *Applied Artificial Intelligence*, 26(8):760—-786, 2012.

[Ott et al. 2011] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Parapar et al. 2012] Javier Parapar, David E. Losada, and Alvaro Barreiro. A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[Peersman et al. 2012] Claudia Peersman, Frederik Vaassen, Vincent Van Asch, and Walter Daelemans. Conversation Level Constraints on Pedophile Detection in Chat Rooms. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[Pendar 2007] Nick Pendar. Toward spotting the pedophile: Telling victim from predator in text chats. In *Proc. First IEEE International Conference on Semantic Computing*, pages 235–241, 2007.

[Pennebaker and King 1999] James W. Pennebaker and Laura A. King. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77: 1296–1312, 1999.

[Pennebaker et al. 2003] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577, 2003.

[Pennebaker et al. 2012] James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. The development and psychometric properties of LIWC2007, June 2012.

[Rayson et al. 2001] Paul Rayson, Andrew Wilson, and Geoffrey Leech. Grammatical word class variation within the British National Corpus sampler. *Language and Computers*, 36 (1):295–306, 2001.

[Salmasi and Gillam 2012] Anna Vartapetiance Salmasi and Lee Gillam. Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[Sebastiani 2002] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[Shalizi 2013] Cosma R. Shalizi. *Advanced Data Analysis from an Elementary point of view*. Cambridge University press, 2013.

[Skillicorn and Lamb 2013] D. Skillicorn and C. Lamb. Extending textual models of deception to interrogation settings. *Linguistic Evidence in Security, Law and Intelligence*, 1(1):13—-40, 2013.

[Spärck-Jones 1972] Karen Spärck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[Villatoro-Tello et al. 2012] Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y Gómez, and Luis Villaseñor Pineda. A Two-step Approach for Effective Detection of Misbehaving Users in Chats - Notebook for PAN at CLEF 2012. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[Vrij 2011] Aldert Vrij. *Detecting lies and deceit: the psychology of lying and the implications for professional practice*. Wiley, 2011.

[Weintraub 1989] Walter Weintraub. *Verbal Behavior in everyday life*. Springer, 1989.