Automatically Identifying Online Grooming Chats Using CNN-based Feature Extraction

Svenja Preuß and **Tabea Bayha** and **Luna Pia Bley** and **Vivien Dehne**
Alessa Jordan and **Sophie Reimann** and **Fina Roberto** and **Josephine Romy Zahm**
and **Hanna Siewerts** and **Dirk Labudde** and **Michael Spranger**
University of Applied Sciences Mittweida
Mittweida, Germany
spranger@hs-mittweida.de

Abstract

With the increasing importance of social media in everyone's life, the risk of its misuse by criminals is also increasing. In particular children are at risk of becoming victims of online related crime, especially sexual abuse. For example, sexual predators use online grooming to gain the trust of children and young adults. In this paper, a two-step approach using a CNN to identify sexual predators in social networks is proposed. For the identification of a sexual predator profile an $F_{0.5}$ score of 0.79 and an F_2 score of 0.98 were obtained. The score was lower for the identification of specific line which initialized the grooming process ($F_2 = 0.61$).

1 Introduction

The importance of social networks in today's society is constantly growing. More and more children and young people are turning to digital forms of communication. Studies from Germany show that 71% of children between the ages of 6 and 13 actively use the Internet, and the trend is rising (Feierabend et al., 2020b). The situation is similar for young people between the ages of 12 and 19 (Feierabend et al., 2020a). In one study, 97% of the teenagers surveyed said they used the Internet every day or at least several times a week (Feierabend et al., 2020a). Those developments provide new opportunities for sex predators to gain access to minors, for example, through online grooming.

The Austrian Federal Criminal Police Office (Bundeskriminalamt, 2015) defines online grooming as the targeting of children and young people on the Internet with the aim of establishing sexual relationships. It is a special form of sexual harassment that can lead to physical and sexual abuse. The contact is initiated via the Internet, for example via social media or online video games.

In child online grooming an adult predator uses means of online communication in order to gain access to and trust from a minor in order use the minor for sexual purposes (Wachs et al., 2012).

In many countries, cyber grooming is legally considered a criminal offense. In the U.S., for example, 18 U.S. Code § 2422 criminalizes online grooming. In 2011, the European Parliament passed Directive 2011/92/EU, which obliges member states to enact corresponding legal regulations, including on criminal prosecution. In Germany the criminal law aspect was regulated in § 176/IV StGB.

In an effort to contain such sexual offenses software to identify potential predators is devised (Inches and Crestani, 2012). That kind of Software is supposed to be a preventive measure whose forensic/criminalistic benefit lies in assisting the day-to-day police work and even possibly preventing sexual offenses from happening. The goal is to reduce the expenditure of time needed to identify a potential sexual predator on social media. (Villatoro-Tello et al., 2012; Peersman et al., 2012)

In addition, to support law enforcement, the detection of chats with criminal content and the marking of relevant text lines is necessary. Therefore, this work will primarily focus on these two tasks. The first task is to detect suspicious chats and distinguish them from inconspicuous chats in order to identify the most likely sex offender within the suspicious chat. Subsequently, the offending lines can then be identified.

One contribution of this paper is classification approaches that enable both automatic detection of conversations and chats involving potential sexual predators, and conversation threads that exhibit distinct offender behavior. This is based on a two-stage approach that includes a CNN as a mechanism for selecting useful lexical features and an MLP as a classifier. It is shown that the use of the

CNN can significantly improve the results.

The development and evaluation of the presented approaches were based on the dataset provided as part of the International Sexual Predator Identification Competition at PAN-2012 (Inches and Crestani, 2012). In contrast to this competition, a main focus of this work is the detection of chats with potential sex offenders. Therefore, this dataset had to be annotated with additional annotations based on the tagged chat participants. In the absence of a suitable ground truth for developing a solution to detect the relevant lines within a chat, a gold standard was developed as an additional contribution to be made available for research purposes in collaboration with the owner of the data.

This paper is organized as follows: At first we present some related work in Section 2, followed by an overview of the data and methods used for this paper in Section 3 and 4. In Section 5 we discuss our results and finally conclude with Section 6.

2 Related Work

Sexual predator identification in social networks as a generic text classification problem is often solved by the use of machine learning. There are numerable publications related to grooming on social networks. Often, however, not the chat/conversation itself, but only the messages or the authors are classified (Villatoro-Tello et al., 2012; Pendar, 2007; Morris and Hirst, 2012; Mcghee et al., 2011; Eriksson and Karlgren, 2012).

Assuming that police investigators manually review all the results, the classification of conversations can reduce the amount of chats an investigator hast to read and, thus, reduces the time spent on the investigation. They would only have to reprocess a fraction of all the conversations, namely those that most likely contain a sexual predator. In previous works, if a chat classification was carried out, it represented only an intermediate step or a prefiltering in order to identify the predator (author) (Villatoro-Tello et al., 2012; Cardei and Rebedea, 2017).

In 2012 the Sexual Predator Identification competition, that was part of PAN [1], dealt with the identification of sexual predators in social networks. The best results were achieved by exercising a so-called Two-Step-Classification (Villatoro-Tello et al., 2012; Morris and Hirst, 2012; Peersman et al., 2012; Cardei and Rebedea, 2017). At first the Suspicious Conversation Identification (SCI) is used to sift out conversations featuring potential predators and, afterwards, the Victim from Predator Disclosure (VFP) is applied to classify the conversationalists (Villatoro-Tello et al., 2012). The winning paper by Villatoro-Tello et al. (Villatoro-Tello et al., 2012) tested both support vector machines (SVM) and neural networks (NN), each with a binary and tf-idf weighted Bag of Words (BoW) (with 117015 elements) as input. The SVM with a tf-idf weighting as SCI was able to achieve slightly better results on the validation data, with an $F_{0.5}$ measure of 0.9516, than a neural network (Villatoro-Tello et al., 2012). A later approach, also using an SVM, this time with a sequential minimum optimization, achieved an $F_{0.5}$ measure of 0.938, using a BoW with 1000 words as well as behavioral and interactive-behavioral features (Cardei and Rebedea, 2017).

This work differs from previous work in this area in particular in that it focuses primarily on chat and relevant line classification rather than author classification. To accomplish this, a multilayer perceptron (MLP) is used to classify the conversations, as this form of neural network has performed well in text classification in the past (Villatoro-Tello et al., 2012).

Generally, the examined features can be divided into lexical and behavioural features. Some approaches exclusively used lexical features (Pendar, 2007; Mcghee et al., 2011; Villatoro-Tello et al., 2012), most in form of a bag-of-words model (Villatoro-Tello et al., 2012; Morris and Hirst, 2012; Cardei and Rebedea, 2017) and sometimes extended through the tf-idf weight (Pendar, 2007; Villatoro-Tello et al., 2012; Morris and Hirst, 2012). For the purpose of extracting lexical features we utilized a convolutional neural network (CNN). Until now, in most cases, the terms and conditions of lexical features had to be initialized by the author, for example, in the form of dictionaries. Typically, these dictionaries contain terms that are unique for sexual predators. By using a convolutional layer to extract the lexical features, the network itself should learn which n-grams and phrases are relevant to distinguish between sexual predator and non-predator chats. In this way, not only terms from the vocabulary of sex offenders are learned, but also frequently used phrases of their chat partners and chats of non-offenders.

[1] A series of scientific events and shared tasks on digital text forensics and stylometry. https://pan.webis.de/

In order to improve the classification additional behavioral features were used (Morris and Hirst, 2012; Eriksson and Karlgren, 2012; Cardei and Rebedea, 2017), which ranged from the response time in conversations (Morris and Hirst, 2012) to the number of asked questions within a single message (Cardei and Rebedea, 2017). Results showed that lexical features are very important for identifying relevant conversations, while behavioral features have less of an impact (Cardei and Rebedea, 2017). In addition to the lexical features we surveyed different combinations of behavioral features, some of which are newly developed and others of which have been applied in previous works, including sentiment analysis (Liu et al., 2017).

In order to identify the suspicious lines in conversations, those that show a distinctive predator behavior, dictionaries were used primarily (Mcghee et al., 2011; Peersman et al., 2012). Another approach looked at the so-called predatoriness score, which is calculated from the summed weights of the uni and bi-grams contained in the message, determined by a linear SVM (Morris and Hirst, 2012). The best outcome for suspicious line detection so far was achieved through first classifying the authors and then, if they were flagged as a predator, returning all their lines, which resulted in an $F_{0.5}$ measure of 0.4762 (Popescu and Grozea, 2012). Another approach involved the use of a pre-trained classifier to sort the messages (Mcghee et al., 2011). In order to identify the distinctive lines in conversations we labeled each message to generate a gold standard and trained a CNN, besides testing a new "line-feature". To the best of our knowledge, no publicly available ground truth currently exists for the training data for this specific task. Therefore, providing a gold standard generated by two independent annotators is one of the new contributions of this paper. In order to drive research in this area, it will be made available in cooperation with the data's owner.

3 Data

The data used in this paper was provided by the 2012 Sexual Predator Identification competition (PAN) and together the data sets consist of 222,055 conversations. Within these conversations a sexual predator can communicate with a potential victim or non-predators can converse with each other. The former could resemble a suspicious message, which indicates a predator behavior, in composition or content. However, predators can also write about mundane topics. Therefore, the number of conversations with suspicious messages is limited to less than 4% in this data set to ensure a realistic scenario. (Inches and Crestani, 2012)

	number of conversations		
	overall	w/o pred.	with pred.
before	155,128	151,391	3,737
after	20,788	19,145	1,643
	number of authors		
	overall	w/o pred.	with pred.
before	218,702	218,448	254
after	35,023	34,794	229

Table 1: Test data before and after preprocessing

Preprocessing was used so as to counterbalance the dataset (Table 1).

3.1 Preprocessing of the Data

The reduction and normalization of the data set were required to further analyze the data. Therefore, all conversations who met at least one of the following conditions were removed from the data set:

- more than four participants (authors), because predators do not take part in such conversations
- only one participant (author) (Villatoro-Tello et al., 2012), since one-sided conversations seldom represent suspicious behavior
- each participant sent less than five messages (Villatoro-Tello et al., 2012), assuming that relevant predator behavior is better detectable after "getting acquainted"
- blank conversations, since no text can be analyzed

Additionally, all messages that contained images made from characters were removed as well (Villatoro-Tello et al., 2012) since they only create static and do not provide usable information. These messages include those which are longer than five rows and those whose ratio between symbols and letters is greater than 45%.

Normalizations were made in regard to spelling out abbreviations and the consistent uncapitalization of all letters (Eriksson and Karlgren, 2012). Emoticons were extracted through SoMaJo (Proisl and Uhrig, 2016) and Emot (Shah and Rohilla,

2018) and afterwards each existing emoticon was assigned an ID in the form of $[1-9]{3}-[a-z]{3}, which improved the detection as well as the differentiation of the individual emoticons. In addition, some preprocessing steps required a normalization of XML special characters.

3.2 Preprocessing of the CNN-input

The CNN-input requires the depiction of texts and words in a machine-readable format. Therefore, all words were lemmatized at first. Afterwards, a dictionary was compiled wherein every word got a corresponding ID and unknown words were assigned the ID null. Conversations or messages were portrayed as a list of one-hot vectors with minor density for each occurring word and brought to the same length by means of padding.

3.3 Preprocessing for the line identification

The data provided by the International Sexual Predator Identification competition at PAN-2012 did not include a ground truth for the identification of messages/lines. So, in order to test our supervised learning approach we had to generate our own ground truth by labeling the data manually. Therefore, the training data set was divided into multiple parts and assigned one of the following labels, which are inspired by Peersmann et al. (Peersman et al., 2012) and McGhee et al. (Eriksson and Karlgren, 2012):

0 - irrelevant

1 - sexual theme:

- (erogenous) body parts
- sexual acts
- sexual oriented adjectives, nouns or terms of endearment
- inquiries regarding clothing, especially underwear ("[...]what are you wearing", "what kind of panties do you have on?")

2 - paraphrasing sexual topics with non-sexual terms:

- characteristic words: "teach", "play", "learn"

3 - meeting in person:

- requests to meet in person, video-chat or call
- characteristic words: "meet", "call"

4 - requests for (personal) information:

- pictures, videos, phone number, webcam, address, ...
- characteristic words: "webcam", "cell", "pic", "address"

5 - inquiries about parents, friends, etc. or police:

- securing privacy, so that nobody finds out about the chat or planned actions
- (e.g., "you just cant tell anyone ok", "[...] make sure you delete this stuff", "who is home with you now")

6 - age references:

- child-oriented vocabulary and pet names (e.g., "cutie pie", "princess")
- statements about age or age differences (e.g., "you know im older")
- aware of the culpability (e.g., "your to young ill get in trouble lol") (Peersman et al., 2012)

This labeling process was repeated, so that each section was evaluated by two different persons and thus the unrelated assessments resulted in a Cohen's Kappa of 0.78742. In some cases, when the labels didn't concur, a third person had to reevaluate the messages.

4 Methods

The "Suspicious Conversation Identification", hereafter referred to as SCI, is the main focus of this paper. The SCI separates conversations depending on the participation of sexual predators. Since the data provided by the International Sexual Predator Identification competition at PAN-2012 is labeled on an author basis the following ground truth is applied to the SCI: Every conversation that contains a sexual predator is denoted as a predator-conversation. The "Victim from Predator Disclosure"(VFP) was tested as an addition. It takes the conversations, returned by the SCI, as input and is supposed to distinguish between sexual predators and other authors (e.g. potential victims). Therefore, author-conversation-pairs were created in order to behold each author in every one of his conversations. The VFP was trained on all the conversations that contained at least one predator. Finally, the amount of authors across all conversations that classified as a sexual predator constitutes the end result.

4.1 Classifier

The SCI/VFP classifier is made of two fundamental components, the feature extractor and the actual classifier (Figure 1).

The feature extractor is composed of a CNN which is trained to extract relevant n-grams for the following classification using temporal max pooling. The CNN input consists of texts in the form of one-hot vectors (Input_1). In order to display the similarity between words with regard to their context an embedding layer was integrated ahead of the convolutional layers. In this experimental setup always 40 of the 1-, 3-, 5- and 7-grams were extracted through an one-dimensional convolutional layer. Other lexical/behavioral features were used as an addition to this feature (cf. Subsection 4.2) (Input_2).

The actual classifier is an MLP that consists of two fully connected dense layers. The first dense layer had a size of 20 units, the second had only one unit and served as an output layer. At last the result was scaled to a value between 0 and 1 by a sigmoid function. As a manner of regularization a dropout layer was employed between the layers with a threshold of 0.5.

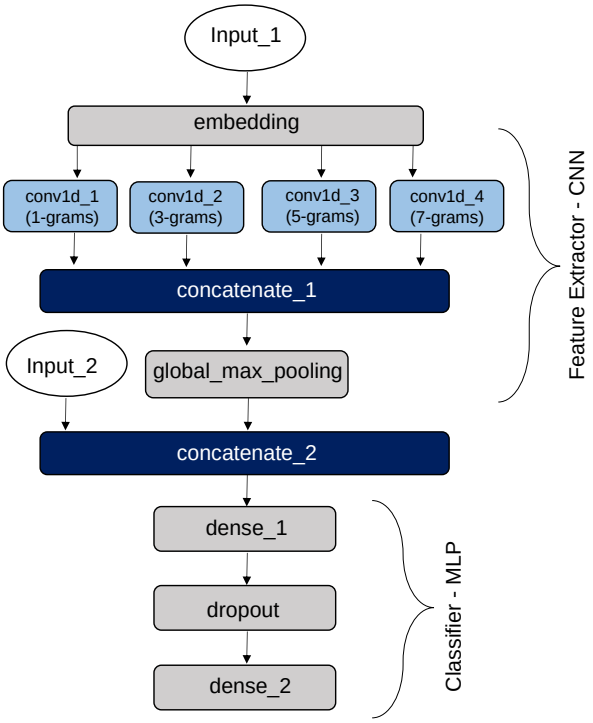

Figure 1: Classifier architecture.

4.2 Feature

The SCI as well as the VFP are based on lexical (LF) and behavioral or conversation based features (BF). The SCI relies on the first feature set (Table 2), which contains conversation-dependent attributes. The second feature set (Table 2) provides the foundation for the VFP. The latter contains similar features to the SCI, which were adjusted to be author-dependent rather than conversation-dependent.

The aforementioned features are based on the corresponding papers (cf. Table 2) and were implemented as follows:

Time of conversation start (TC): The time at which the conversation starts was represented as a figure that was rounded to the nearest whole hour. Every hour is represented two-dimensionally by an x- and y-coordinate in the unit circle so as to obtain a sound result during the change of days.

Duration of a conversation (DoC): For each conversation the duration of a conversation (in minutes) resulted from the difference between the time of the first and last messages.

Number of asked questions (NQ): The number of asked questions was made up of the percentage amount of messages per conversation (feature set 1) or else the amount of messages per author for each conversation (feature set 2) that contained questions. The amount of questions per author for every conversation was determined as well.

Number of messages (NoM): The total number of messages was defined as the amount of sent messages per conversation (feature set 1). In order to identify how dominant an author is in a conversation the percentage amount of messages per author for each conversation was determined (feature set 2).

Number of used emoticons (NoE): For each author the number of used emoticons was counted per conversation utilizing the emoticon-IDs. On the one hand the average number of emoticons per message was calculated for each author, on the other hand the amount of emoticons used by an author compared to the total amount of emoticons in the conversation was determined.

Response time (RT): The response time resulted from the difference between the point in time (in minutes) at which a message was sent and the moment the following message arrived in the conversation. For each conversation the mean response time was determined by calculating the sum over

feature set 1
time of conversation start
duration of conversation
of asked questions (Morris and Hirst, 2012; Cardei and Rebedea, 2017)
of messages (Morris and Hirst, 2012)
sentiment analysis (Liu et al., 2017)
feature set 2
of asked questions (Morris and Hirst, 2012; Cardei and Rebedea, 2017)
of messages (Morris and Hirst, 2012)
of used emoticons (Morris and Hirst, 2012)
response time (Morris and Hirst, 2012; Cardei and Rebedea, 2017)
conversational initiation (Morris and Hirst, 2012; Cardei and Rebedea, 2017)
of words per author (Morris and Hirst, 2012)
sentiment analysis (Liu et al., 2017)

Table 2: Used feature sets (behavioural)

all response times for all authors.

Conversational initiation (CI): The conversational initiation describes which author begins a conversation by sending the first message. Those authors got the value 1 assigned to this feature, other authors got the value 0.

Number of words per author (WA): The word count was defined by the average number of words used in a message by an author. In order to identify the level of participation in a conversation the word count for an author in a conversation was divided by the total word count for that specific conversation.

Sentiment analysis (SA): The sentimemt analysis feature was tested through four different approaches. The first attempt dealt with the Sentistrength tool (Thelwall et al., 2010a), a program that returns values between -1 (not negative) and -5 (very negative) or values between 1 (not positive) and 5 (very positive) in order to score the various sentiments. This entire analysis was based on a dictionary which also took misspelling and negations (e.g. "not nice") into consideration. In addition, a list of boost-words was integrated, whose words, like "very" or "extremely", could amplify the level of positivity/negativity of the sentiment (Thelwall et al., 2010b). The second attempt utilized a similar program, TextBlob, which was based on a dictionary as well. However, the returned score only regards the adjectives that were used and lies between -1 and 1 (Sohangir et al., 2018). The last two attempts did not apply premade tools and trained classifiers instead, by using a data set of 6.3 million tweets (Malafosse, 2019). Both were implemented according to two existing works. On the one hand, the classifier decided whether the sentiment was negative, neutral or positive, but not it's intensity (third approach) (Malafosse, 2019). On the other hand, the classifier was trained in Tensorflow (fourth approach) and returned four values (negative, neutral, positive, mixed) for each text input, which add up to 1 as shown by (Liu et al., 2017).

In this paper, the performance of all features (combined) was tested at first. Then, each feature was surveyed on its own. The features that obtained the best results on the training data were occasionally combined and analyzed again. The final results on the test data arise from those features and feature combinations that achieved the best performances on the training data.

4.3 Line identification

The analysis of lines that show a distinctive predator behavior was conducted under three different rudiments:

1. Usage of the pre-trained CNN from the VFP:
   - the CNN already learned distinctive word patterns in order to identify a sexual predator.
   - single messages from the test data were forwarded as input for the prediction.
2. Usage of a new CNN:
   - a new CNN, whose training was based on the generated ground truth, was created.
   - this classifier used a similar architecture to the SCI and VFP, but the second concatenate layer as well as the input were omitted.
3. Usage of the new CNN in combination with the line feature:

- in addition to the, through the CNN extracted, n-grmas a new feature (line feature) was tested.
- the line feature is based on the assumption that relevant messages are often found in the middle of a conversation. It refers to the message number in relation to the total number of messages in a respective conversation.
- the architecture of the classifier is the same as for the SCI/VFP.

5 Results and Discussion

For the purpose of detecting that epoch, which delivers the best results without overfitting, the overfitting-behavior was analyzed for each epoch for the SCI classifier.

5.1 Sentiment analysis

The sentiment analysis ensued in different manners (cf. Subsection 4.2). Our initial assumption, that conversations with a sexual predator should obtain positive sentiment scores more often than conversations without a predator, was confirmed through the sentiment analysis on a conversational basis. As can be seen in Figure 2, conversations with a sexual predator were to 65.97% positive and conversations without a predator only to 37.66%. Negative sentiment scores were more common for non-predator conversations with 41.62%.

Therefore, our next assumption was that a sexual predator would reach a sentiment score that was distinctly more positive than that of a non-predator (Liu et al., 2017), which couldn't be confirmed through the approach with Tensorflow. According to that the conversational partners of a sexual predator acquired positive scores in 505 conversations, the predators themselves only in 409 conversations. Thus the sentiment scores for predator/non-predator don't allow for a meaningful differentiation.

So far all the tested approaches were nearly indistinguishable. Therefore SentiStrength was used to attain the following results, because of it's easy handling and velocity.

5.2 SCI classification

Already, the lexical features, which were extracted through the CNN, yielded sound results on the validation data, which could be improved by joining the behavioral features. The combination of lexical features and sentiment scores (Table 3) resulted in an $F_{0.5}$ of 0.9935. All the results so far are based on a stratified 5-fold cross-validation.

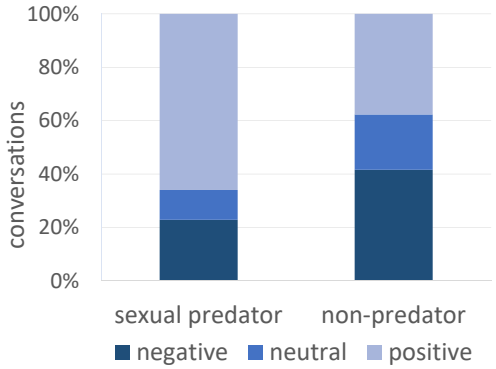

Figure 2: Sentiment values for predaror and non-predator conversations (third approach) (Malafosse, 2019).

Because of these findings a model that trained on lexical features and sentiment scores was reviewed on the test data. With a precision of 0.9982 and a recall of 0.9349 the following F-measures represent the best outcome for a classification of sexual predator chats to date: $F_{0.5}$ of 0.9723 and F_2 of 0.9469.

By reference to this procedure the number of apparently relevant conversations was reduced to 1567, which corresponds to roughly 1% of all the conversations that would have had to be screened manually. Thereby, only 30 conversations were classified as false-positives and 106 conversations were classified as false-negatives. Unfortunately by doing so 18 predators could not be identified. However, it is possible that the false negative classified conversations are attributable to the method, which was used to create the ground truth (cf. Subsection 4.1), where conversations with a sexual predator, that didn't show suspicious behavior, were labeled as relevant.

5.3 VFP classification

The data returned by the SCI created the foundation for the VFP, consisting of altogether 1537 predator conversations and 1567 non-predator conversations.

Similar to the SCI the lexical features constituted a great prerequisite for further analyses based on the training data results. In combination with one

Features	Precision	Recall	$F_{0.5}$
VT2012	-	-	0.9516
CR2017	0.9380	0.9380	0.9380
LF	0.9600	0.9322	0.9541
LF + BF	0.9826	0.9874	0.9835
LF + SA	**0.9935**	0.9891	**0.9926**
LF + TC	0.9881	0.9891	0.9882
LF + DoC	0.9881	**0.9907**	0.9886
LF + CQ	0.9891	0.9814	0.9875
LF + NoM	0.9934	0.9820	0.9910

Table 3: Results for the SCI classification on the training data compared to baseline results from (Villatoro-Tello et al., 2012) (VT2012) and (Cardei and Rebedea, 2017) (CR2017)

other behavioral feature significant improvements could be reached compared to the union of all features (Table 4). All results on the training data are based on a stratified 5-fold cross-validation.

The four most expressive behavioral features were then reviewed on the test data, either in combinations or alone with the lexical features (Table 5). Thereby, the conjunction of lexical features and all four of the aforementioned behavioral features achieved the best result with an $F_{0.5}$ measure of 0.9169 and an F_2 measure of 0.8916. 1466 author-conversation-pairs were returned as relevant, 109 of them were false positives and 179 couldn't be detected (false negatives). In order to identify sexual predators they have to be detected as such in at least one of their conversations. Therefore the end result is determined over all conversations to obtain the exact amount of authors, classified as predators (Table 6). Here the combination of lexical features and the four aforementioned behavioral features achieved the best result as well, with an $F_{0.5}$ measure of 0.7889 and an F_2 measure of 0.9221. The number of classified sexual predators was 213, an additional 70 were false positives and solely 5 predators could not be identified at all.

The obvious difference between the two F-measures is caused by the varying weight and the relatively low precision. Due to the imbalance of authors in the data set the 70 authors, who were incorrectly classified as predators, are a pretty small number compared to the overall 34,794 non-predators. Whereas, compared to the low total number of only 229 sexual predators, the 70 false positives carry a considerable weight, thus causing a low precision.

The usage of the two F-measures is justified through their computation which goes along with different assertions. *$F_{0.5}$-measure*: In order to optimize the expenditure of time that investigators need to find a potential sexual predator, it is better to only have the "right" suspects rather than returning every possible one (Inches and Crestani, 2012). *F_2-measure*: Since the investigators have to double-check the results given by the classifier anyways, it is better to have classified innocent authors as potential suspects (false positives) rather, than to miss out on an actual sexual predator. Therefore, it is important to increase the weight of the recall over the precision.

Features	Precision	Recall	$F_{0.5}$
LF	0.8689	0.8720	0.8693
LF + NoE	0.9279	0.9256	**0.9273**
LF + RT	**0.9302**	0.9147	0.9269
LF + CI	0.9297	0.9070	0.9249
LF + NoM	0.9290	0.9114	0,.252

Table 4: Best results for the VFP classification on the training data.

Features	Precision	Recall	$F_{0.5}$	F_2
LF + NoE	0.9042	0.8665	0.8964	0.8738
LF + CI	0.9201	**0.8841**	0,9127	0.8911
LF + RT	0.9162	0.8613	0.9047	0.8717
LF + NoM	0.9218	0.8750	0.9121	0.8840
all	**0.9256**	0.8835	**0.9169**	**0.8916**

Table 5: Results for the VFP classification on the test data

Features	Precision	Recall	$F_{0.5}$	F_2
VT2012	0.9804	0.7874	0.9346	0.8197
CR2017	1.0000	0.8180	0.9570	0.8489
LF + NoE	0.7241	0.9633	0.7620	0.9036
LF + CI	0.7276	0.9679	0.7656	0.9079
LF + RT	0.7376	0.9541	0.7727	0.9012
LF + NoM	0.7413	0.9725	0.7783	0.9154
all	**0.7527**	**0.9771**	**0.7889**	**0.9221**

Table 6: Final results for author classification over conversations compared to baseline results from (Villatoro-Tello et al., 2012) (VT2012) and (Cardei and Rebedea, 2017) (CR2017).

5.4 Identifying suspicious messages

The results for the line identification (Table 7) were determined by the given ground truth.

The third approach, a CNN that trained on the self-created ground truth, combined with the line feature (LiF), resulted in the best F_3 measure of

Features	Precision	Recall	F_3
PG2012	0.0915	0.8938	0.4762
CNN (VFP)	0.2472	**0.7247**	0.6074
CNN (GT)	0.4590	0.6971	0.6628
CNN + LiF (GT)	**0.4653**	0.7046	**0.6702**

Table 7: Final results for the line classification on the test data, comparing the CNN used for the VFP with the CNN trained on the self-created ground truth (GT) and with the baseline results from (Popescu and Grozea, 2012) (PG2012).

0.6702, with a precision of 0.4653 and a recall of 0.7046. The same CNN without the line feature (second approach) obtained a similar result with an F_3 measure of 0.6628. Those similarities imply that the assumption, that relevant messages occur more often in some paragraphs than in others, is true, however, no significant improvements could be reached.

The pre-trained CNN from the VFP (first approach) reached an F_3 measure of 0.6074. Because of its low precision with only 0.2472 and the greater weighting of the recall the latter has a larger impact on this F-measure.

The results of all three approaches show a greater recall, compared to the precision, which could be explained by the high count of messages that were returned as relevant, regardless of whether they were correctly classified or not. Nevertheless, the approaches that were based on the self-created ground truth (cf. Subsection 4.3) achieved a more balanced relation between precision and recall.

Due to the different approaches used to solve this task the results are difficult to compare. Notwithstanding the above, all three of the aforementioned approaches surpassed the existing results of the Sexual Predator Identification competition at PAN 2012.

6 Conclusion

Both the results of the sexual predator conversation identification and the identification of relevant messages have shown that a CNN can be of great use in extracting lexical features in the form of N-grams. With its help, the results known to us could be exceeded in both areas. The result of the SCI showed that a sentiment analysis in connection with the lexical feature is very well suited to the identification of sexual predator conversations and achieved an $F_{0.5}$ measure of 0.9723. Further tests with feature combinations have not yet been continued. The tests of the VFP showed, however, that the most successful features combined led to an improvement in the end result. Accordingly, a further step would be to combine features of the SCI and see whether this can lead to a further improvement. Especially with the knowledge that other features, such as the number of messages written by each author, showed similarly good results on the training data as the sentiment analysis.

A possible exploratory approach with regard to the VFP could be transfer learning based on the neural network trained for the SCI. The learned features of the SCI are used further and adapted and interpreted for the identification of a sexual predator.

When identifying the relevant messages, a newly tested line feature in conjunction with the lexical features was able to achieve the best results. The CNN that was used for the extraction of lexical features was trained on a self-created ground truth. When annotating the lines, it was particularly noticeable that some messages can be rated as relevant in one context and as irrelevant in another. Only the message "playing" in a sexual context would be a clear word for "paraphrase of sexual topics with non-sexual vocabulary" and thus relevant, but not to be considered relevant in connection with a hobby (sports). At the moment, each message was rated individually without knowing what was previously written. Another sequence-based network, such as an RNN, could possibly differentiate these messages better.

References

Bundeskriminalamt. 2015. Schutz vor (cyber-)grooming. Last accessed: August 14th, 2021.

Claudia Cardei and Traian Rebedea. 2017. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Natural Language Engineering*, 23(4):589—-616.

Gunnar Eriksson and Jussi Karlgren. 2012. Features for Modelling Characteristics of Conversations—Notebook for PAN at CLEF 2012. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy*. CEUR-WS.org.

Sabine Feierabend, Thomas Rathgeb, Hediye Kheredmand, and Stephan Glöckler. 2020a. Jim-studie 2020-jugend, information, medien-basisuntersuchung zum medienumgang 12- bis 19-jähriger. Last accessed: August 14th, 2021.

Sabine Feierabend, Thomas Rathgeb, Hediye Kheredmand, and Stephan Glöckler. 2020b. Kim-studie

2020-kindheit, internet, medien-basisuntersuchung zum medienumgang 6- bis 13-jähriger. Last accessed: August 14th, 2021.

Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors. 2012. *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy*. CEUR-WS.org.

Giacomo Inches and Fabio Crestani. 2012. Overview of the International Sexual Predator Identification Competition at PAN-2012. In (Forner et al., 2012).

Dan Liu, Ching Yee Suen, and Olga Ormandjieva. 2017. A novel way of identifying cyber predators. *Computing Research Repository*, arXiv:1712.03903. Version 1.

Charles Malafosse. 2019. Fasttext sentiment analysis for tweets: A straightforward guide. Last accessed: February 25th, 2019.

India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra Mcbride, and Emma Jakubowski. 2011. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103—-122.

Colin Morris and Graeme Hirst. 2012. Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features—Notebook for PAN at CLEF 2012. In (Forner et al., 2012).

Claudia Peersman, Frederik Vaassen, Vincent Van Asch, and Walter Daelemans. 2012. Conversation Level Constraints on Pedophile Detection in Chat Rooms—Notebook for PAN at CLEF 2012. In (Forner et al., 2012).

Nick Pendar. 2007. Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241.

Marius Popescu and Cristian Grozea. 2012. Kernel Methods and String Kernels for Authorship Analysis—Notebook for PAN at CLEF 2012. In (Forner et al., 2012).

Thomas Proisl and Peter Uhrig. 2016. Somajo: State-of-the-art tokenization for german web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62, Berlin. Association for Computational Linguistics.

Neel Shah and Shubham Rohilla. 2018. Open source emoticons and emoji detection library: Emot. Version 2.1.

Sahar Sohangir, Nicholas Petty, and Dingding Wang. 2018. Financial sentiment lexicon analysis. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 286–289.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010a. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010b. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558.

Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes y Gómez, and Luis Villaseñor-Pineda. 2012. Two-step Approach for Effective Detection of Misbehaving Users in Chats—Notebook for PAN at CLEF 2012. In (Forner et al., 2012).

Sebastian Wachs, Karsten D. Wolf, and Ching-Ching Pan. 2012. Cybergrooming: risk factors, coping strategies and associations with cyberbullying. *Psicothema*, 24(4):628–633.