NTNU – Trondheim
Norwegian University of
Science and Technology

Detection of cyber grooming during an online conversation

Halvor Kulsrud

Submission date:	June 2019
Responsible professor:	Patrick Bours, IIK
Supervisor:	Patrick Bours, IIK

Norwegian University of Science and Technology
Department of Information Security and Communication Technology

Title: Detection of cyber grooming during an online conversation
Student: Halvor Kulsrud

Problem description:

In an online society, a person can assume any identity they want and be anonymous while posting, commenting and chatting online. Unfortunately does this anonymity also lead to people engaging in unfriendly or even illegal activities. As a result of this, a severe problem is that of cyber grooming, where sexual predators try to build up a trust relationship with children in a chat room to share erotic images or even worse to convince the victim to meet in real life.

To address the societal challenge of cyber grooming, this project aims to detect cyber grooming in an early stage of an online conversation. Such a classification is meant to be used to warn children, platform owners and law enforcement of the possibility that an online chatter is doing something illegal. The objective of the warnings is to reduce the number of incidents caused by online grooming. This project's main interest is to investigate whether it is possible to detect child grooming during an online conversation.

Responsible professor: Patrick Bours, IIK
Supervisor: Patrick Bours, IIK

Abstract

Cyber grooming is a prominent societal problem, and few solutions to mitigate the problem exists. One in five youths have been exposed to unwanted sexual content, and one in nine have experienced unwanted online sexual solicitations. This project aims to detect cyber grooming in an early phase of an online conversation. Three predator identification methods were developed and tested before one was selected and tested on conversation segments and full-length conversations to find out whether it is possible to detect predators at an early stage of a conversation. The Conversation-Based Detection (CBD) approach with two classification stages obtained the best results on the conversation segments. The performance was measured with an $F_{0.5}$-score where the best result was 0.893. The classification method detected 209 out of 254 predators and misclassified 20 non-predatory authors in a dataset with 218702 authors. The CBD approach was further tested on a limited number of messages within the conversations to see how early in the conversations that it could recognize a predator. The CBD approach managed to detect 101 of the 254 predators within 20 messages, 191 within 50 messages and 207 within 80 messages. Intermediate results and manual analysis showed that the combination of terms used in the process of cyber grooming is different from the combination of terms used in general conversation. Not all of the analyzed predators built relations to their victims before they attempted to groom the victims. Most of the analyzed predators applied the same course of conduct to approach a child. However, the pace of the predators varied. Predator detection during online conversations can help to mitigate the societal problem of online grooming. Predator detection is a well-researched area, but it has not been tested in an environment of ongoing conversations before. This thesis puts light on the importance of early detection in order to detect predators before any physical or psychological harm is caused to the victims.

Sammendrag

Internett fasilitert grooming er et fremtredende samfunnsproblem, og få løsninger for å mitigere problemet eksisterer. Én av fem midreårige har vært utsatt for uønsket seksuelt innhold, og en av ni har opplevd uønskede seksuelle forespørsler. Denne oppgaven har som målsetning å oppdage grooming i en tidlig fase av en internett fasilitert samtale. Tre metoder for å oppdage overgripere er utviklet og testet. Deretter ble en av dem valgt til å teste om det er mulig å oppdage overgripere i en tidlig fase av en samtale. En samtalebasert implementasjon med to klassifiseringstrinn oppnådde de beste resultatene. Resultatene ble målt med en $F_{0,5}$-score, og det beste resultatet var 0,893. Klassifiseringsmetoden oppdaget 209 av 254 overgripere og feilklassifiserte 20 brukere som overgripere i et datasett med 218702 brukere. Den samtalebaserte implementasjonen ble ytterligere testet på et begrenset antall meldinger i flere samtaler for å finne ut hvor tidlig i samtalene den kunne gjenkjenne en overgriper. Den samtalebaserte implementasjonen oppdaget 101 av de 254 overgriperne innen 20 meldinger, 191 innen 50 meldinger og 207 innen 80 meldinger. Mellomliggende resultater og manuell analyse viste at kombinasjonen av begreper som brukes i en internett fasilitert grooming prosess er forskjellig fra kombinasjonen av begreper som brukes i generelle samtaler. Ikke alle de analyserte overgriperne opprettet en relasjon til ofrene sine før de forsøkte å groome dem. De fleste overgriperne som ble analysert brukte samme tilnærming for å groome barn. Deteksjon av overgripere i internett fasiliterte samtaler kan bidra til å redusere samfunnsproblemet grooming utgjør. Det er forsket mye på deteksjon av overgripere, men deteksjon har ikke tidligere blitt testet for pågående samtaler. Denne oppgaven legger vekt på betydningen av tidlig deteksjon for å oppdage overgripere før en overgriper gjør psykisk eller fysisk skade på offeret. Arbeidet i denne oppgaven har vist at det er mulig å oppdage grooming i en tidlig fase av en internett fasilitert samtale.

Preface

This thesis is written as the final part of a Master's degree within communications technology at the Norwegian University of Science and Technology (NTNU) in the faculty of Information Technology and Electrical Engineering. The work was conducted from January to mid June 2019. Patrick Bours has been the supervisor and responsible professor for this project.

The topic of this thesis is part of a larger security project currently running at NTNU. The goal of the project is to detect people with fake profiles and child predators based on their typing and stylometric behavior. This thesis is only focusing on detecting child predators to mitigate the societal problem of cyber grooming.

Halvor Bugge Kulsrud
Trondheim, Tuesday 11th June, 2019

Acknowledgments

I would like to thank my supervisor Professor Patrick Bours of the NTNU at the faculty of Information Technology and Electrical Engineering. Patrick was always available whenever I ran into a trouble spot or had a question about my research or writing.

I would also like to thank my co-students for their contribution in technical discussions, feedback and comments during this semester.

Halvor Bugge Kulsrud
Trondheim, Tuesday 11th June, 2019

Contents

List of Figures

List of Tables

List of Terminologies

Artifact An artifact is something created by people for some practical purpose. Examples of artifacts designed and studied in information systems and software engineering research are algorithms, methods, notations, techniques, and even conceptual frameworks [Wie14].

CNN Convolutional Neural Networks (CNNs) are deep neural networks which use multilayer perceptrons to minimize preprocessing.

DataFrame A Python Pandas DataFrame is a two-dimensional data structure.

F-score F-score or F-measure is a binary classification measure of a test's accuracy. It is explained in detail in Subsection 3.4.1.

Neural Network NN is a computer system modeled on the human brain and nervous system. It works as a framework for machine learning algorithms in order to work together and process complex data inputs..

SVM SVMs are supervised learning models that analyze data used for classification and regression analysis. When given a set of training examples, it divides the data into two groups, separated by a gap.

TF-IDF Term Frequency–Inverse Document Frequency is a numerical statistic that weight the importance a word is to a document in a collection or a corpus based on its number of appearances.

List of Acronyms

ABD Author-Based Detection.

ANN Artificial Neural Network.

BoW Bag-of-Words.

CA Continuous Authentication.

CBD Conversation-Based Detection.

CEOP Child Exploitation and Online Protection.

CLEF Conference and Labs of the Evaluation Forum.

CNN Convolutional Neural Network.

CSAM Child Sexual Abuse Material.

EU European Union.

ICAC Internet Crimes Against Children.

IRC Internet Relay Chat.

IWF Internet Watch Foundation.

LIWC Linguistic Inquiry and Word Counting.

MBD Message-Based Detection.

MCE Missing Children Europe.

MLE Maximum Likelihood Estimation.

MLP Multi-layer Perceptron.

NB Naïve Bayes.

NCIS National Criminal Investigation Service.

NCMEC National Center for Missing and Exploited Children.

NLP Natural Language Processing.

NLTK Natural Language Toolkit.

NN Neural Network.

NSD Norwegian Centre for Research Data.

NTNU Norwegian University of Science and Technology.

PJ Perverted Justice.

RNN Recurrent Neural Network.

SCI Suspicious Conversations Identification.

SVM Support Vector Machine.

TF-IDF Term Frequency–Inverse Document Frequency.

UK United Kingdom.

USA United States of America.

VFP Victim From Predator disclosure.

WEBIS Web Technology & Information Systems Network.

XML eXtensible Markup Language.

Chapter 1 Introduction

This chapter presents the motivation, research question and hypothesis for this project. It also determines the scope and gives an outline for the remaining chapters.

1.1 Motivation

In an online society, a person can assume any identity they want and be anonymous while posting, commenting and chatting online. Unfortunately, anonymity also leads to people engaging in unfriendly or even illegal activities. As a result of this is the severe problem of cyber grooming, where sexual predators try to build up a trust relationship with children in a chat room to share erotic images or even worse to convince the victim to meet in real life.

Online grooming is a significant problem in today's society, where people spend more and more time online. In 2015, more than 80% of youth in the United States of America (USA) had access to the Internet and children aged 5-16 spent on average 6.5 hours per day on devices connected to the Internet [MVA+18]. The Internet provides many opportunities and is an excellent source of information. However, the Internet is also a mostly unregulated place and thus can put youth in risk of dangers such as unwanted online solicitation. Online solicitation is a scenario where a peer or adult requests to engage in unwanted sexual activities or sexual talk online. Youth have lower socio-cognitive sophistication on a general basis when compared to adults [MVA+18]. It makes youth less likely to foresee potential threats when interacting online. Studies have revealed that 25% of youth reported that they were considerably distressed or afraid as a result of online solicitation [MVA+18]. Reports and investigations of online sexual exposure and solicitation of youth have increased over time. Findings suggest that approximately one in five youths have been exposed to unwanted sexual content, and one in nine have experienced unwanted online sexual solicitations [MVA+18]. The findings do not account for unreported incidents, which often happens because children might feel guilty, ashamed or not even know that they were abused.

The National Center for Missing and Exploited Children (NCMEC) received more than 8.2 million reports to their CyberTipline about Child Sexual Abuse Material (CSAM) in 2016 [ECP18]. The number of reports was almost double the amount from the year before and eight times more than in 2014. The CyberTipline works as an online mechanism for members of the public and electronic service providers to report incidents of suspected child sexual exploitation [ECP18]. In 2009, there were 8144 arrests for technology-facilitated sexual crimes against children in the USA [ECP18]. In the United Kingdom (UK) there were 1247 offenses reported for taking, making or distributing child abuse images in 2012/2013 [ECP18]. Online solicitation is the reason behind a large portion of the produced images. Organizations and governments that are working to protect children from online predators advice both parents and children to educate themselves on how to use the Internet safely. However, most of the population is either unaware or ignore advice offered from government and children associations on how to protect children online.

According to NCMEC *"Online Enticement involves an individual communicating with someone believed to be a child via the Internet with the intent to commit a sexual offense or abduction. This is a broad category of online exploitation and includes sextortion, in which a child is being groomed to take sexually explicit images and/or ultimately meet face-to-face with someone for sexual purposes, or to engage in a sexual conversation online or, in some instances, to sell/trade the child's sexual images"* [Nat]. When analyzing reports of online enticement from the CyberTipline, it was deduced that the age of the victims ranged from 1 to 17 and that the average age was 15. Almost all of the children said that they did not know the extorter prior to the communication.

This project aims to detect cyber grooming in an early stage of an online conversation to address the societal challenge of cyber grooming. Such a classification is meant to be used to warn children, platform owners and law enforcement of the possibility that an online chatter is doing something illegal. The objective of the warnings is to reduce the number of incidents caused by online grooming. The following research question was developed in adherence to this project:

- **RQ1:** To what extent is it possible to detect child grooming during an online conversation?

Four hypotheses were made to expedite the research question. They will be explored in the result chapter and thoroughly discussed in the discussion and conclusion chapters.

- **H1:** Terms used in the process of cyber grooming are categorically different from the terms used in general conversations.
- **H2:** Predators must build relations to the victims before they attempt to groom them.
- **H3:** Predators apply the same course of conduct to approach a child.
- **H4:** Grooming cannot be detected during the initial phase of an online conversation.

1.2 Scope

This thesis is a part of a larger security project running at NTNU, where this work consists of detecting predators from online communication platforms. The work includes analyzing chat logs by looking at single messages and complete transcripts. The work mainly focuses on conversations with two participants. Capturing Instant Messaging or Internet Rely Chat is not a part of the scope for this project. Complete and available transcripts were used instead.

1.3 Outline

The remainder of this thesis is structured as follows:

Chapter 2 presents the background for this project. The background chapter contains information and explanation of cyber grooming. It includes a detailed summary of related work and state of the art for online predator identification. Lastly, it describes technical information about machine learning and legislation on cyber grooming.

Chapter 3 presents the methodology that have been used during this project.

Chapter 4 describes the dataset that has been used in this project. The chapter includes where the collection of conversations were gathered from and how the collection is structured.

Chapter 5 presents the results that have been obtained during the project. The chapter presents results that make the foundation to answer the research question.

Chapter 6 discuss the presented results from Chapter 5 in light of the research question and hypothesis presented in the introduction.

Chapter 7 presents concluding remarks and suggestions for future work.

Chapter 2 Background

This chapter presents three definitions of the term cyber grooming. The chapter proceeds by going through related work in the field of predator identification and predator detection. After that, it describes technical information about machine learning techniques and legislation covering cyber grooming.

2.1 Cyber Grooming

Grooming is the process where a predator builds trust with a child with the intention of sexual abuse. Grooming usually includes lowering the child’s inhibitions to sexual content. The word cyber is normally used to describe something that involves computers and networks. Thus, cyber grooming is when a predator is grooming a child over the Internet.

[NMEL18] defines child grooming or sexual grooming as *“a communication process by which a perpetrator applies affinity seeking strategies, while simultaneously engaging in sexual desensitization and information acquisition about targeted victims in order to develop relationships that result in need fulfillment”* such as physical sexual solicitation. As such, the term pedophile or sexual predator is used to describing such people, and these terms are often used interchangeably [NMEL18].

[MBK+11] defines grooming as *“the subtle communication strategies that sexual abusers use to prepare their potential victims to accept the sexual conduct”*. Thus, communication that functions as grooming does not directly lead to sexual contact, but instead, desensitizes the victim to sexual remarks or foul language. Successful grooming leaves the victim unaware that any process is underway [MBK+11].

[EEL10] states that *“grooming involves subtle communication strategies that desensitize victims to sexual terminology and reframe sexual acts in child-like terms of play or practice.*

Grooming does not have to occur online, and thus for the above definitions to fit as cyber grooming the communication must take place online. Common for all of the above definitions is the word desensitization, which in this context means to expose victims to sexually explicit language or images. All of the definitions also point out that grooming is a communication process, and such is the phase before the sexual abuse. In this thesis, grooming is defined according to the first definition as described by [NMEL18].

2.2 Related Work

There is plenty of related work within the topic of online predator identification. The Perverted Justice (PJ) website [Per] has been the main source for predatory transcripts used within the research domain. As technology has evolved, manual methods of catching predators are no longer efficient. Thus, there is a need for better and automated methods. This subsection presents the work that has been conducted to improve the methods for identifying predators online and the current state of the art.

Pendar's pilot study [Pen07] on using automatic text categorization techniques in identifying online sexual predators has set the foundation for how to differentiate between predator and victim in text chats. He motivates his study by pointing at the need for a software application that can flag suspicious online chats automatically. He motivates the need with a statement that online sexual predators always outnumber law enforcement officers and volunteers. Besides, an objective of the study is to increase awareness in the research community of this important issue and the attainability of a solution [Pen07]. Pendar divides the sexual content relevant to his study into two groups. The first group consists of interactions between a sexual predator and what that individual believes to be a victim, and the second group consists of consensual interaction between two adults. Pendar points out that data acquisition is a significant problem for some of the subcategories of the first group. However, he points out that the next best thing in this group is available from the PJ website. Pendar's study did not include data from group two, and thus only focused on distinguishing victim and predator. He collected 701 text logs from PJ and split them such that each part only contained one person's messages. He trained a series of SVM and distance-weighted k-NN classifiers and used unigrams, bigrams and trigrams from the training data as features. Furthermore, a combination of document frequency and odds ratio were used for feature extraction. By averaging the odds ratio for all the n-grams from the training set, nine feature sets were built by extracting 5000, 7500 and 10000 unigrams, bigrams and trigrams, which had the highest average odds ratio. When testing the effectiveness in an SVM and a distance-weighted k-NN classifier, the best result for the SVM was achieved using a feature set built on 10000 trigrams. The best result achieved for the k-NN used trigram features with 10000 trigrams and k=30. The SVM and the k-NN, respectively obtained a F-score of 0.908 and 0.943. A few experiments were also made to differentiate between the predators based on their "sliminess", which the predators are scored against on the PJ website. The results were hardly any better than chance. Pendar therefor concluded that predator and child side of text chats use a different subset of the English language, while among predators the language is similar. Pendar concludes that it is possible to distinguish the victim from the predator in a predatory conversation.

Edwards et al. [EEL10] present the state of technology for studying Internet crimes against children and relevant articles related to the study of cybercrime. Their approach is to protect children from cyber predators by integrating communication and computer science theories and methodologies to develop automated tools. They point out the importance of differentiating luring in the real world and online contexts. They include slang, abbreviations, netspeak and emoticons as part of their analysis. To perform a content analysis of Internet predation, they developed a codebook and a dictionary. The codebook and the dictionary were used to make a software program they called ChatCoder. They managed to correctly identify the predator in a predatory conversation 60% of the time with ChatCoder. In a second experiment, they managed to distinguish a small sample of Perverted Justice transcripts from a small sample of non-predatory transcripts 93% of the time. More interestingly, by looking at different language patterns used by predators and clustering them with the k-means algorithm, they managed to find what they believe to be four different types of predators.

Wollis' thesis [Wol11] presents the idea of using automated text analysis to identify different stages in the grooming process. She uses a Linguistic Inquiry and Word Counting (LIWC) program [Pen]. LIWC reads a given text and counts the percentage of words that reflect a given category. The study consist of a three-stage grooming model which is merged from five different phases of the grooming process. She reduced the five phases into three by combining "friendship" with "relationship forming", and "risk assessment" with "exclusivity". The last phase is "sexual". She analyzed transcripts from the perverted justice site [Per]. Wollis assessed the messages of the predators and removed any other messages. She divided each transcript into three parts of equal length based on a simple word count. Each part represents one of the three phases in the grooming process. A problem with LIWC that affect her results is that it only recognizes real words that are represented in its dictionary. Thus, incorrectly spelled words and internet language is not recognized. The result barely supports the author's hypothesis.

Egan, Hoskinson and Shewan [EHS11] focused on finding recurrent themes that indicate cyber grooming. They used content and data analysis in an attempt to solve the problem of cyber grooming. By using content analysis, the authors wish to get insight into the offenders' thought process. A data analysis software called NVivo [QSR] uses conversation transcripts as input. The result from the software analysis was eight recurrent themes to classify the presence of grooming in a conversation. The language offenders used in the conversations indicated a willingness for risk-taking behavior. Offenders arranged offline meetings with little caution. The behavior indicates that minimizing the risk of detection was of little importance.

McGhee et al. [MBK$^+$11] took the previous approaches a step further by using machine learning algorithms to label each line in a conversation. Their approach used communication theories and computer algorithms to identify predatory messages. Different machine learning algorithms classified lines based on phrase matching and rule-based approaches, and the best result was obtained using the nearest neighbor algorithm. It was able to label the lines correct 83.11 percent of the time. The experiment contained 33 unique conversations. The nearest neighbor algorithm outperformed the k-nearest neighbor's algorithm. Two of the label types, grooming and approach, were used to identify incidents of grooming.

Guapta, Kumaraguru and Sureka [GKS12] divided the grooming process into different stages and used those stages to create psycho-linguistic profiles. The purpose was to gain useful insights and patterns. To achieve their purpose they used the same program as [Wol11] did, LIWC [Pen]. The ultimate goal of their study is to build a real-time automated tool that can flag an ongoing conversation on the Internet as a pedophile conversation. Their current work only consists of the initial processes of profiling a perpetrator and do not include any performance measures.

Pandey, Klapaftis, and Manandhar [PKM12] used SVMs to detect the behavioral profile of a predator. Their research introduces a combination of machine learning and computational linguistics to detect predator behavior from online textual chats. They created a data model by training on both predatory and non-predatory chat logs. The resulting method means to be able to detect and raise an alarm whenever it detects a chat to contain predator activity. The final result used SVM with n-grams. In this context, n-grams are the contiguous sequence of n words in a conversation. When using trigrams for SVMs, they correctly classified the profiles with an average accuracy of 76.23 percent over the tested dataset.

Inches et al. [IC12] give an overview of the international sexual predator identification competition at PAN 2012. The competition was concerned with solving two challenges. The first challenge was to identify as many predators as possible from a collection of chat logs containing both predatory and non-predatory conversations. The second challenge was to identify which of the predators' lines that were deemed to reflect grooming behavior.

For the first challenge, a common approach was to start with a pre-filtering stage and a two-stage classifier. The first stage of the classifier usually consisted of distinguishing between predatory (true positive) and non-predatory (false negative) conversations. The first stage was necessary because the datasets were designed to reflect a real-life scenario, where the majority of the conversations were false negatives, and only about one percent of the conversations were true positives. The second stage of the classifier differentiated between the victim and the predator in a suspicious conversation.

The participants in the competition used two main groups of features, lexical features and behavioral features [IC12]. Lexical features are taken directly from the raw text of conversations. Behavioral features, on the other hand, are those features that capture a user's action within a conversation. Examples of behavioral features are the number of times a user starts a conversation, the number of questions asked and message response time. In the classification step SVMs were most used, but other submissions also included neural networks, maximum-entropy, decision trees, k-NN, random forest and Naïve Bayes (NB). Some of the authors combined dictionaries of predatory language with their classifiers.

The second challenge was more laborious and did not include any training data. To cope with this, most of the participants collected all of the lines from whom they had identified as predators in the previous challenge. Then they ran those lines up against a dictionary of perverted language or used a scoring system such as TF-IDF.

Inches et al. conclude that lexical and behavioral features work well for predator identification. Pre-filtering is essential, and there is not one unique method to identify predators, but different approaches exist [IC12].

Peersman et al. [Pee12] present what they deem to be an entirely new way of detecting online predators in chat rooms by combining results based on predictions of individual posts, user and the entire conversations. They participated in the PAN 2012 competition, where the main task they worked on solving was sexual predator identification. They experimented using SVMs with different settings. They made an interesting observation during error analysis that in some cases, both users in a conversation were labeled predator. They suspect that the reason was due to victims mirroring vocabulary of the predator. After using the predator probabilities of the user classifier to find the real predator in each conversation, they managed to achieve a precision of 0.94 and recall of 0.85 which translates into a F-score of 0.90 on the training set. When retraining their models from the F-score of 0.90 on the training set, they managed to achieve an F-score of 0.72 on the test set. The results from identifying single grooming messages were not as good as the online predator identification and Peersman et al. only achieved an F-score of 0.302. However, when evaluated by the F-score with β of 1 (F_1-score), this was the best score achieved for identifying single grooming messages in the PAN 2012 competition.

Villatoro-Tello et al. [VTJGE$^+$12] work differs from previous work according to themselves in that they can *“identify when a chat conversation is a case of child exploitation and subsequently to tell which user is the sexual predator”* as one solution. They are calling their two main stages Suspicious Conversations Identification (SCI) and Victim From Predator disclosure (VFP). The SCI stage act as a filter by distinguishing general chatting from possible cases of online child exploitation. Villatoro-Tello et al. competed in the PAN 2012 competition and were the highest ranked participants for the task of detecting online predators. They did not preprocess the texts from the competition dataset because they did not want to lose potentially valuable information. However, as a mean to focus only on the most important cases and to reduce the computational cost of automatically processing all the information, they added a pre-filtering stage.

The pre-filtering stage removed all conversations that either had only one participant, less than an average of six messages per user or contained long sequences of unrecognized characters. The pre-filtering reduced the number of conversations with approximately 90%, while at the same time keeping almost 92% of the predators. The authors argued that the messages from the removed predators were not sufficient to effectively recognize them as predators. Examples of messages from removed predators are displayed in Table 6.1.

The authors approached the sexual predator identification task as a text classification task. Text classification is the process of assigning tags or categories to text according to its content, and it is one of the fundamental tasks in Natural Language Processing (NLP). To train the SCI classifier, the authors employed text classification techniques to build a model that distinguishes between general chatting and cases of child exploitation [VTJGE$^+$12]. To properly train the SCI, they labeled all the chat conversations that included at least one predator as a suspicious conversation. This lead to a total of 798 suspicious conversations. For the VFP classifier, they divided text conversations containing predators into interventions, where one intervention is all of the messages that are written by one user in one specific conversation. Thus, each user within a predatory conversation had one or more interventions. The VFP classifier discovered 194 examples of victims from the set of interventions.

Villatoro-Tello et al. used NNs and SVMs for classification. The NNs consisted of two layers with a single hidden layer of ten units and for the SVMs, they tested both linear and polynomial kernels. Two-fold cross-validation was used to estimate their performance during the development phase only using training data. Two-fold cross-validation is a way to split the data into two equal parts, to use one part as training data and the other as test data and then swap them around. Their best result from the SCI stage during testing was obtained using SVM with TF-IDF weighting. The best result for the VFP stage was obtained using NN with binary weighting.

The authors best result achieved a precision of 0.9804, recall of 0.7874, which lead to an F-score of 0.9346 when using a β of 0.5 as set by the organizers of the event.

As for future improvements, they suggest to include linguistic features to improve the recall levels of their proposed system.

Meyer's master thesis [Mey15] addresses the challenge of detecting adults pretending to be children. The goal with the thesis is to move a step towards an automated analysis of chat room conversations to detect possible attempts on grooming. Due to the limitation of public transcripts of predators posing as children, transcripts of law enforcement officers posing as children have been analyzed instead. A significant part of the work consists of age estimation. Meyer used Adaboost, SVM and NB classifiers to estimate age from texts. He transformed text documents into feature vectors. Pairwise statistical analysis, the expert knowledge technique and model validation were performed as features selection to reduce the number of features. Meyer experimented on different mixes of book reviews, blogs and chat conversations where there was an equal amount of children and adult authors. After many different experiments, Meyer achieved perfect results to differentiate between a child and someone pretending to be a child. However, he performed experiments with very little data. He used less than 1000 conversations, and only 20 of them contained law enforcement officers. Meyer suspected bias towards the topic and suggested some reasons for the perfect results. Even though he could not point to any specific reason, Meyer believes the reason behind the good results were due to law enforcement officers overplaying their part as children. Meyer backs up his thoughts with analysis on the content of the conversations and also that the comparison between adults and real children only performed slightly better than by chance. An essential feature vector was foreign words, which was largely more used by law enforcement officers than actual children. Foreign words were words from a language other than English, misspellings, slang, abbreviations and emoticons. Meyer believes that law enforcement officers purposely had more misspellings, abbreviations and slang than actual children. Meyer's ending argument is that it is possible to differentiate someone pretending to be a child from both the way adults and children communicate. Thus, he concludes that it is possible to differentiate someone pretending to be a child from a real child.

Ashcroft et al. paper [AKM15] is similar to Meyer's, an approach to identify adults pretending to be children. Their work consists of two steps. The first step is to classify authors on different platforms as adults or children. They classified authors from book reviews, blog posts and online chatroom conversations with the Adaboost algorithm [FS96]. Their next step checks for each child, whether they are genuine children or someone else posing as a child. By using the Adaboost algorithm on both regular chat conversations and predatory conversations, the authors achieved almost a perfect distinction. However, the authors were suspicious of their results. They fear it is more likely that they were identifying law enforcement officers doing their job. Law enforcement officers and regular adults have different behavior in online conversations. Law enforcement officers act more suspicious and direct. Thus, they are most likely not representative for an average adult. Even so, based on further research in the article, the authors conclude that it is possible to determine an adult pretending to be a child. Ashcroft et al. have a very similar approach as Meyer and their conclusion is the same.

A more recent study on the topic of online predator identification [ESO16] compare different text classification methods and introduce their own based on CNNs. Their findings suggest that CNNs have the best result of identifying an online predator. An interesting observation from their study is that CNNs outperformed general pre-trained word vectors and SVMs. Their study also shows that using only one convolution layer led to better results than having a deeper structure with several convolution layers. They introduced one-hot vectors, a method in text classification where the order of words matters. One-hot vectors outperformed methods such as using simple representations of unigrams. By using One-hot CNN they managed to get a F-score of 0.8087.

Mabuza et al. motivate their research by describing the societal problem of cyber grooming and its outcomes [NMEL18]. They present an overview of machine learning technologies and algorithms that have been employed in attempts to mitigate cyber grooming. They conclude that most of the existing solutions use lexical features and luring communication theory as their foundation. In their conclusion, they point out the fact that most of the employed methods are based on supervised learning, and that there have been few attempts on methods such as unsupervised or reinforcement learning. Their research paper is a prestudy and does not include any models. However, the authors want to further investigate and implement semi-supervised deep learning models as future work to improve accuracy on CNN models.

The earlier work presented in this section used the full length of conversation segments in their experiments to detect cyber grooming. It is too late to detect a predator when a conversation has finished. In such cases, there is already a victim. This project implements a method to stop grooming by detecting the predatory conversation before the end of the conversation. It aims to detect grooming as early as possible. This project uses Machine learning and NLP techniques to detect predators. It differs from existing work in that an incremental number of messages are analyzed to detect the predator as early as possible. The early detection is a continuous evaluation for each posted message. It is intended to be used to close the conversation and warn the other user and law enforcement of the grooming attempt.

2.3 Technical Background

This section provides technical information about NLP and machine learning techniques used in the project.

2.3.1 Natural Language Processing Techniques

NLP is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a valuable manner [Gar]. This section presents the two following NLP techniques, Bag-of-Words (BoW) and TF-IDF.

Bag-Of-Words

BoW is a simple yet quite an effective method in NLP. The method consists of counting the occurrences of each word in a text, which is used to create a dictionary. The dictionary is then used to measure the presence of known words in a text. BoW is used to extract features from a text which again can be used for modeling, where machine learning algorithms are popular examples. BoW does not care about the order or structure of words in a text, as its name indicate. The main idea about the method is that texts with similar content are similar texts and that it is possible to learn something about the meaning of the text based on its content [MS99].

Term Frequency-Inverse Document Frequency

TF-IDF is an approach that extends the BoW method by also focusing on the total frequencies of words in a corpus. TF-IDF helps to penalize too frequent words and remove words that occur less than a specified amount of times from the feature space [MS99]. The term frequency part of TF-IDF describes the number of times a term occurs in a text. The inverse document frequency, on the other hand, decreases the weight of terms that occur very frequently in the collection of texts and increase the weight of terms that occur more rarely. It is important to focus on the words that matter and not to focus on words such as syncategorematic words, which are words that cannot stand by themselves, for example, "the", "a" and "of". The TF-IDF method can be further extended to include the n-gram model, which combine consecutive words and add them to the dictionary.

TF-IDF is computed as the product of term frequency and inverse document frequency.

$$tf_{i,j} * idf_i \tag{2.1}$$

A normalized term frequency, which is normalized in order to prevent bias towards longer documents, is given as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{2.2}$$

Where $n_{i,j}$ is the number of occurrences for the term t_i in document d_j and the denominator is the sum of number of occurrences of all terms in document d_j, which is the size of the document. Inverse document frequency can be written as,

$$idf_i = \log \frac{N}{|\{j : t_i \epsilon d_j\}|} \tag{2.3}$$

where N is the total number of documents, and the denominator is the document frequency of the term t_i [MS99].

2.3.2 Machine Learning Classifiers

A classifier is an algorithm which maps input data to specific categories in order to solve a classification problem. This subsection presents the logistic regression, ridge regression, NB, SVM and NN classifiers.

Logistic Regression

Logistic regression is a simple and common method to solve binary classification problems. The logistic model computes the logarithm of the odds as a linear combination of one or more independent variables which are often called predictors [Nav]. The probability of each of the output values lays between zero and one. These values are converted from the logarithm of the odds to probability by a logistic function, which is the reason for the name of the classifier. Logistic regression is estimated using the Maximum Likelihood Estimation (MLE) approach. By maximizing the likelihood function, the parameters that are most likely to produce the observed data can be determined.

Logistic regression is based on a combination of the linear regression equation and the Sigmoid function. The Sigmoid function is shaped like an 'S' formed curve which maps any real-valued number into a value between zero and one. Output from the Sigmoid function above 0.5 is more likely to be classified as one, and output below 0.5 is more likely to be classified as zero. The values correspond to the probability of whether the input belongs to zero or one, where a value of 0.75 corresponds to a probability of 75 % that the input belongs to one. Figure 2.1 illustrates a logistic regression model with its two possible output values.

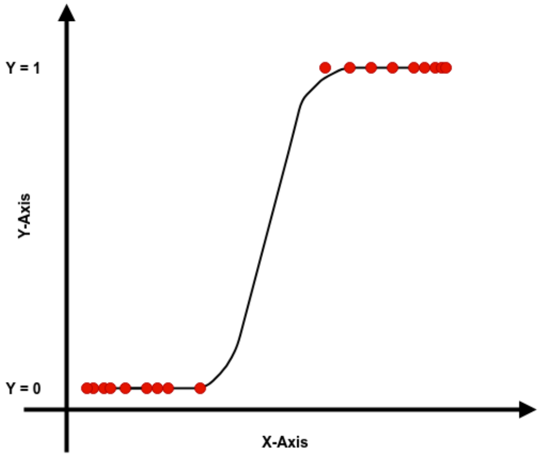


Figure 2.1: Example illustration of a logistic regression model.

The equation for logistic regression is derived from applying the Sigmoid function on linear regression, as shown in Equation 2.6.

Linear regression equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{2.4}$$

Sigmoid function:

$$p = \frac{1}{1 + e^{-y}} \tag{2.5}$$

Logistic regression:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}} \tag{2.6}$$

Advantages of using a logistic regression classifier are that it does not require high computation power, it is easy to implement and widely used by data analysts and scientists. Disadvantages, on the other hand, are that it is not able to handle many features and is vulnerable to overfitting [Nav]. In statistics, overfitting is the production of an analysis which corresponds too closely or exactly to a particular set of data, and may, therefore, fail to fit additional data or predict future observations reliably [Oxf].

Ridge

The ridge classifier is a model that accounts for situations where the number of predictor variables exceeds the number of observations or where a dataset has correlations between predictor variables [Scia]. In comparison with least square regression, ridge regression overcomes the problem where a least square regression model is not defined when the number of predictors exceeds the number of observations. This scenario happens to least square regression because it does not differentiate between important and less important predictors in a model, and thus includes them all. In such situations, the least square regression model will overfit and fail to find unique solutions. Least square regression also has problems when dealing with correlations between predictor values in data. Ridge regression avoids these problems by using biased estimators that have just enough bias to make the estimates reasonably reliable [Scia].

Ridge regression uses L2 regularization, meaning that it adds an L2 penalty which is equal to the square of the magnitude of the coefficients [Scia]. The coefficients are shrunk by a factor which is equal for all of the coefficients such that none of them are eliminated. A tuning parameter (α) is used to control the power of the penalty term. Given an α equal to zero, ridge regression is just the same as least square regression. On the other hand, an α approaching infinity will result in that all coefficients are shrunken to zero. Thus, an ideal penalty will lay somewhere in between the two.

Naïve Bayes

The different variants of NB are all supervised learning algorithms that are based on Bayes' theorem. Bayes' theorem uses prior knowledge of conditions that might be related to an event to describe the probability of the event. The NB algorithms also use what is called a "naive" assumption of conditional independence between every pair of features. NB classifiers have performed well in many real-life scenarios such as document classification and spam filtering even though it uses over-simplified assumptions. The classifiers only require a small amount of training data to estimate its necessary parameters, which makes them very fast compared to more advanced models [Scic].

Bernoulli NB uses multiple features which are independent binary values. The decision rule for Bernoulli NB is

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{2.7}$$

where y is the class variable, x are feature vectors, and i are features [Scic].

Support Vector Machine

SVM is a classifier that is defined by a separating hyperplane. A hyperplane is a (V-1)-dimensional subspace of a V-dimensional vector space [Bis06]. To simplify, this means that the hyperplane of a two-dimensional plane is a one-dimensional line. When an SVM is trained, it outputs a hyperplane which is used to categorize new data.

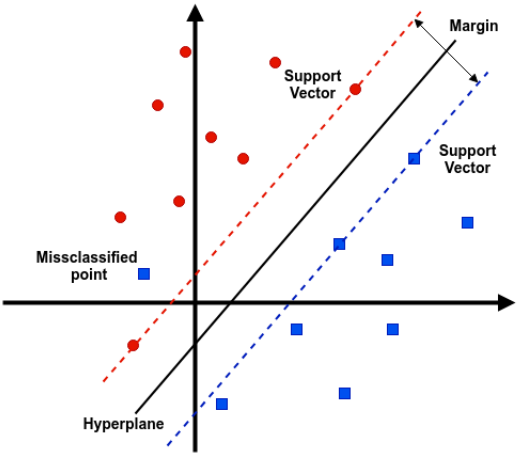


Figure 2.2: A representation of a linear SVM.

With the introduction of the kernel trick, SVM became useful also for non linearly separable data. The idea behind it is that non linearly separable data in a specified dimensional space may be linearly separable in a higher dimensional space [Pat]. SVMs can have different kernels, where some of the options are polynomial, Gaussian, Sigmoid and linear. When it comes to a linear kernel, the learning of the hyperplane is performed by transforming the problem by using linear algebra. To predict a new input, the dot product of the input (x) and each support vector (x_i) is calculated as

$$f(x) = B_0 + \sum a_i * (x, x_i) \tag{2.8}$$

where the inner products of a new input vector (x) are calculated with all the support vectors of the training data. B_0 and a_i are coefficients estimated from the training data [Pat].

The SVM classifier contains a regularization parameter (C) which is used to specify how much misclassifying is tolerated for each training input. Large C values result in a smaller margin for the hyperplane if the hyperplane does a better job of classifying the training input correctly. Small C values ignore a few misclassifications and try to achieve a larger margin for the hyperplane. SVMs use another parameter called gamma, which defines how far the influence of a single training input reaches. Low values consider points that are far away from the separation line, and high values do not.

For an SVM classifier to perform optimally, it needs a good margin. A margin is a separation of a line to the closest training inputs on both sides of the line. A good margin is achieved when there is an equal distance to the closest training inputs on both sides of the separation line. SVMs are effective in high dimensional spaces, memory efficient and often the best choice in binary classification tasks.

Neural Network

Neural Networks (NNs) are a machine learning framework built on the same logic as the biological neural networks that compose animal brains [GB17]. NN attempts to mimic the learning pattern of biological neural networks where interconnected neurons receive inputs and use them to produce outputs [GB17]. To differentiate between neural networks and biological neural networks, it is common to use Artificial Neural Networks (ANNs) when talking about neural networks used in computing systems. ANN will be referred to as NN throughout this thesis.

In this thesis, a supervised learning algorithm called Multi-layer Perceptron (MLP) has been used to represent an NN. When an MLP is given a set of features and a target, it can learn a non-linear function approximator for classification [Scib]. The algorithm is different from logistic regression in that it uses non-linear layers between the input and the output layer. The non-linear layers are called hidden layers.

The leftmost layer in Figure 2.3 is called the input layer and consist of a set of neurons representing the input features. The input features are transformed by each of the neurons in the hidden layer with a weighted linear summation, then followed by a non-linear activation function [Scib]. The output layer then receives the values from the last hidden layer before transforming them into output values.

MLP trains on two arrays (X and y) using a form of gradient descent where the gradients are calculated using backpropagation [Scib]. The X array contains the training samples represented as feature vectors, and the y array contains the labels for the training samples. In classification MLP uses the Cross-Entropy loss function, which measures the performance for a classification model whose output is a probability value between zero and one, to output a vector of probability estimates per sample [Scib]. The main advantage of MLP is its capability to learn non-linear models. Disadvantages include the need for tuning different hyperparameters such as

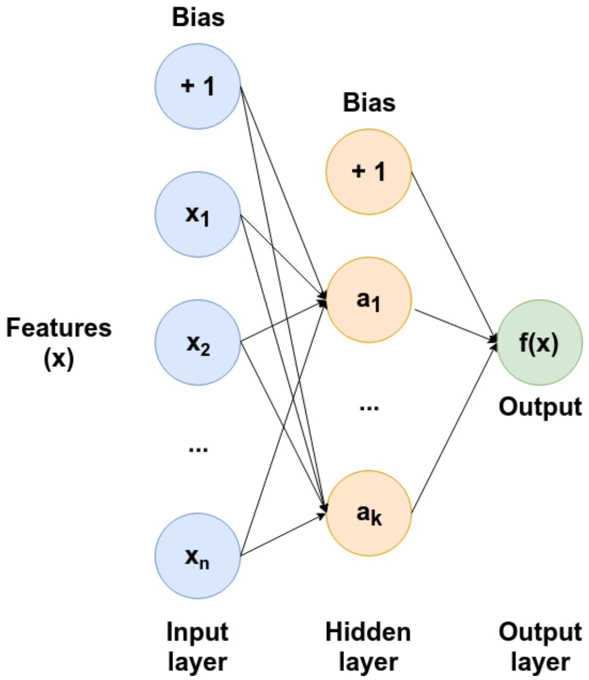


Figure 2.3: Logic representation of a MLP with one hidden layer [Scib].

the number of hidden neurons, layers and iterations. MLP is also sensitive to feature scaling.

2.3.3 Cross-validation

K-fold cross-validation is used to estimate the performance of machine learning models on unseen data. It is performed by dividing training data into different folds and use each of the folds to test on exactly one time and to train on K-1 times. The validation method uses the folds in order to estimate how the model can be expected to perform on a general basis when used in predictions of data that was not included in the training of the model [Scid]. The idea is illustrated in Figure 2.4.

Cross-validation can be thought of as several rounds of the more straightforward method train/test split given that the folds of a train/test split were of equal size and remained the same over each round. Thus, the train/test split is a method where the training data is split into a train and a test part to estimate the performance of the model. Benefits of cross-validation compared to train/test split is that it is a more reliable estimate for out-of-sample performance, it can be used to select tuning parameters, choosing between models and selecting features [Scid]. The main

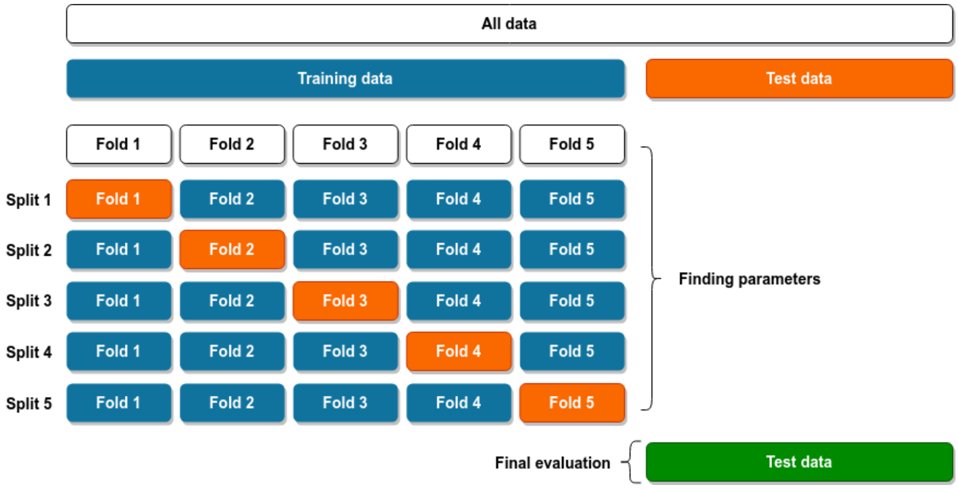


Figure 2.4: An example figure of K-fold cross validation, with K=5 [Scid].

drawback is that it can be computationally expensive.

2.4 Legislation

This section presents the laws concerning online grooming in Norway, the European Union (EU) and the USA. The laws from Norway are presented because this thesis is written in Norway and also to be used as a comparison to the laws in the EU and the USA. American laws are presented because the predatory data used in this project is gathered within the USA. European laws are presented to compare with American and Norwegian laws.

2.4.1 Norway

According to Norway's The Penal Code, Part II. Criminal acts, Chapter 26, Sexual offenses from June 2009, Section 306, Arranging a meeting to commit sexual abuse: *"A penalty of a fine or imprisonment for a term not exceeding one year shall be applied to any person who has arranged a meeting with a child under 16 years of age, and who with intent to commit an act with the child as specified in sections 299-304, section 305 b) or section 311 first paragraph a) has arrived at the meeting place or a place where the meeting place may be observed"* [Min]. This legislation covers both grooming and online grooming. It should be noted that section 306 applies before any sexual activities have occurred, and even before a meeting has occurred. It is enough for the predator to show up close to an arranged meeting place for this law to apply.

Section 302. Sexual activity with a child between 14 and 16 years of age states that: *"Any person who engages in sexual activity with a child between 14 and 16 years of age shall be subject to imprisonment for a term not exceeding six years, unless the conduct also falls within the scope of other provisions. The same penalty shall be applied to any person who makes a child between 14 and 16 years of age perform acts corresponding to sexual activity on himself/herself"* [Min]. This law covers predators that interact with children in activities such as cybersex, sexual activities performed in front of a webcam and when encouraging a child to take sexual photos of itself.

Section 303. Aggravated sexual activity, etc. with a child between 14 and 16 years of age: *"Aggravated violation of section 302 is punishable by imprisonment for a term not exceeding 15 years. The same applies if the offender has previously been convicted of acts specified in sections 291, 299 or 302. In determining whether a violation of section 302 is aggravated, particular weight shall be given to whether*

a) the act was committed by multiple persons acting together,
b) the act was committed in a particularly painful or offensive manner, or
c) the aggrieved person died or suffered considerable harm to body or health as a result of the act. A sexually transmitted disease is always considered considerable harm to body or health pursuant to this section" [Min].

Section 304. Sexual act with a child under 16 years of age states that: *"Any person who performs a sexual act with a child under 16 years of age shall be subject to imprisonment for a term not exceeding three years, unless the conduct falls within the scope of section 299"* [Min]. Section 299-301 deals with sexual assault on a child under 14 years of age and has a penalty of imprisonment for a term not exceeding 21 years. Section 304 is less strict than section 302 and 303, and covers all forms of sexual acts.

Section 305. Sexually offensive conduct, etc. directed at a child under 16 years of age *"A penalty of a fine or imprisonment for a term not exceeding one year shall be applied to any person who*

a) by words or conduct exhibits sexually offensive or other indecent conduct in the presence of or directed at a child under 16 years of age.
b) forces or induces a child under 16 years of age to exhibit sexually offensive or other indecent conduct, unless the situation falls within the scope of stricter provisions" [Min].

The sections presented above are gathered from The Penal Code in Norway [Min], which have been translated by ministries and other public authorities from Norwegian to English. *"The translations are not official; they are provided for information purposes only. In the event of any inconsistency, the Norwegian version shall prevail"*

[Min]. These sections are the most relevant when discussing online grooming as all of them have the potential to be used in litigation against any predator depending on their actions. Online and offline actions are considered equal.

2.4.2 European Union

Legislation for online grooming in the EU is found in Directive 2011/92/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA. According to article 1, which states the subject matter: *"This Directive establishes minimum rules concerning the definition of criminal offenses and sanctions in the area of sexual abuse and sexual exploitation of children, child pornography and solicitation of children for sexual purposes. It also introduces provisions to strengthen the prevention of those crimes and the protection of the victims thereof"* [EURa].

Laws related to online grooming are defined in Article 6: Solicitation of children for sexual purposes. *1. Member States shall take the necessary measures to ensure that the following intentional conduct is punishable: the proposal, by means of information and communication technology, by an adult to meet a child who has not reached the age of sexual consent, for the purpose of committing any of the offences referred to in Article 3(4) and Article 5(6), where that proposal was followed by material acts leading to such a meeting, shall be punishable by a maximum term of imprisonment of at least 1 year* [EURa]. Thus, chatting itself is not punishable without material acts leading to a meeting or production of child pornography. However, once again, it should be noted that these are the minimum rules for the member states of the EU. It is up to each member state whether to implement more strict laws, such as making the communication itself punishable.

"2. Member States shall take the necessary measures to ensure that an attempt, by means of information and communication technology, to commit the offenses provided for in Article 5(2) and (3) by an adult soliciting a child who has not reached the age of sexual consent to provide child pornography depicting that child is punishable" [EURa]. Once again, these laws do not make it punishable to send predatory text messages, nor do they address the situation where a predator is sending predatory pictures. However, they do make it punishable for predators to "knowingly" receive pictures of their victims.

Where the referenced articles, from Article 3: Offences concerning sexual abuse and Article 5: Offences concerning child pornography, states:
Article 3(4): Engaging in sexual activities with a child who has not reached the age of sexual consent shall be punishable by a maximum term of imprisonment of at least 5 years.

Article 5(2): Acquisition or possession of child pornography shall be punishable by a maximum term of imprisonment of at least 1 year.
Article 5(3): Knowingly obtaining access, by means of information and communication technology, to child pornography shall be punishable by a maximum term of imprisonment of at least 1 year.
Article 5(6): Production of child pornography shall be punishable by a maximum term of imprisonment of at least 3 years [EURa].

The legislation works differently for the European Union than for a single country. EU treaties are achieved by several types of legal acts. They are divided into regulations, directives, decisions, recommendations and opinions. The following definitions are gathered from [Eurb]: *A regulation is a binding legislative act. It must be applied in its entirety across the EU. A directive is a legislative act that sets out a goal that all EU countries must achieve. However, it is up to the individual countries to devise their own laws on how to reach these goals. A decision is binding on those to whom it is addressed (e.g. an EU country or an individual company) and is directly applicable. A recommendation is not binding. A recommendation allows the institutions to make their views known and to suggest a line of action without imposing any legal obligation on those to whom it is addressed. An opinion is an instrument that allows the institutions to make a statement in a non-binding fashion, in other words without imposing any legal obligation on those to whom it is addressed. It can be issued by the main EU institutions (Commission, Council, Parliament), the Committee of the Regions and the European Economic and Social Committee. While laws are being made, the committees give opinions from their specific regional or economic and social viewpoint.*

Article 27: Transposition states *"1. Member States shall bring into force the laws, regulations and administrative provisions necessary to comply with this Directive by 18 December 2013"* [EURa]. In other words, this means that these are the laws that apply in regard to online grooming as a minimum requirement in each of the EU member states. Each member state can make their own laws as long as they comply with the EU Directives.

2.4.3 United States of America

According to 18 U.S. Code §2422. Coercion and enticement from 2015:
"(a) Whoever knowingly persuades, induces, entices, or coerces any individual to travel in interstate or foreign commerce, or in any Territory or Possession of the United States, to engage in prostitution, or in any sexual activity for which any person can be charged with a criminal offense, or attempts to do so, shall be fined under this title or imprisoned not more than 20 years, or both.
(b) Whoever, using the mail or any facility or means of interstate or foreign com-

merce, or within the special maritime and territorial jurisdiction of the United States knowingly persuades, induces, entices, or coerces any individual who has not attained the age of 18 years, to engage in prostitution or any sexual activity for which any person can be charged with a criminal offense, or attempts to do so, shall be fined under this title and imprisoned not less than 10 years or for life" [Leg]. 18 U.S §2422(a) is general in terms of age and not directed at minors. 18 U.S §2422(b) on the other hand, is concerned about minors, which here translates to any person less than 18 years old. Thus, part b is the one of interest for this project.

18 U.S. Code §2425. Use of interstate facilities to transmit information about a minor, states that: *"Whoever, using the mail or any facility or means of interstate or foreign commerce, or within the special maritime and territorial jurisdiction of the United States, knowingly initiates the transmission of the name, address, telephone number, social security number, or electronic mail address of another individual, knowing that such other individual has not attained the age of 16 years, with the intent to entice, encourage, offer, or solicit any person to engage in any sexual activity for which any person can be charged with a criminal offense, or attempts to do so, shall be fined under this title, imprisoned not more than 5 years, or both"* [Leg].

18 U.S. §2422(b) makes it a federal offense to entice or persuade a minor to sexual activity. While 18 U.S. §2425 makes it a federal offense to transmit information about a minor that is under 16 years old with the purpose of solicitation or similar of any person to engage in sexual activity. Thus, §2425 ensures that sharing of information on a minor is punishable. These are the national laws within the domain of cyber grooming in the USA. The national laws are quite general and broad, and not specific for different kind of violations. However, some states have additional laws to cope with cyber grooming. An example of this is the law in Florida that makes "Use of a Computer to Seduce a Child" a felony.

2.4.4 Comparison

The legislation in Norway, EU and USA are respectively from 2009, 2011 and 2015/2017. Thus, the laws in the USA are the most recently reviewed, where the latest one was updated only two years ago. The Norwegian and EU laws, on the other hand, have been around for some years. Section 306 of The Penal Code is the law in Norway that can be directly linked to online grooming. Although the law is quite old and not directly aimed at online communication, it covers both online and offline communication. The law specifies that the perpetrator must have arrived at or close to the meeting place in order to be punished by a fine or imprisonment up to one year. However, it should be noted that more severe punishments may apply depending on the content of the conversation as covered by the other sections.

European law 6(1) is closely related to the Norwegian one, and specify that a meeting must take place for the communication to be punishable. It also states that member states must make the act punishable with a maximum of at least one year. The European legislation is more general than the Norwegian legislation and has the potential for a more severe punishment depending on a member state's decision on the magnitude of the punishment. The laws in the USA are even more general than the European ones and aimed towards illegal sexual activities. Online grooming is a part of these sexual activities. The minimum penalty for online grooming in the USA is more severe than in Norway and the EU.

Chapter 3 Methodology

This chapter presents the methodology used to investigate and answer the research question and hypothesis from Chapter 1. The methodology chapter is inspired and based on the book "Design science methodology for information systems and software engineering" [Wie14] written by Roel Wieringa. The methodology was chosen because it provides guidelines for doing design science in information systems and software engineering systems, which fits well with the purpose of this thesis.

3.1 Design Science

Design science is the design and investigation of artifacts in context. Design science iterates over two activities: designing an artifact that improves something for stakeholders and empirically investigating the performance of an artifact in context [Wie14]. The artifact of this thesis is the method to detect cyber grooming during an online conversation, and the context consists of mitigating the societal problem of cyber grooming. Design science problems are improvement problems, and it is the interaction between the artifact and the problem in the context that contributes to solving the problem [Wie14].

This project strives to solve a design problem. The problem is to design a method to detect grooming during an online conversation. The technical research goal of the design problem is to redesign existing predator identification methods to detect a predator in an early phase of a conversation. The technical research goal aims to meet the social context goal, which is to mitigate the societal problem of cyber grooming. This chapter will focus on the highlighted parts of Figure 3.1, which address the part of design science that covers design problems.

A design science project iterates over the activities of designing and investigating. The design task itself is decomposed into three tasks, namely, problem investigation, treatment design, and treatment validation [Wie14]. These three tasks are called

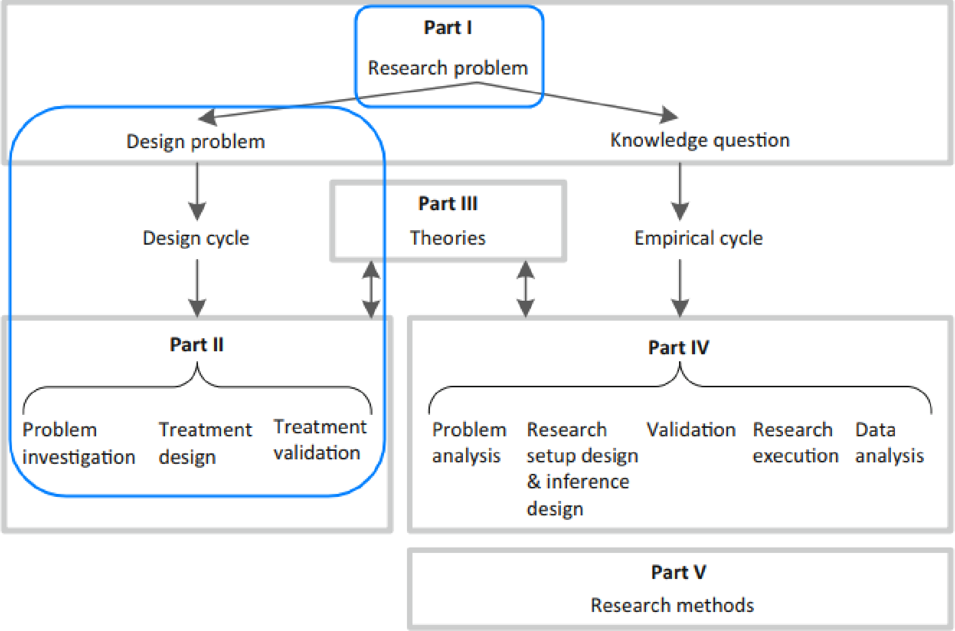


Figure 3.1: A modified overview of the design science methodology [Wie14].

the design cycle because they are iterated several times in a design science research project.

3.2 Problem Investigation

Problem investigation is the investigation of real-world problems as a preparation for the design of a treatment for the problem. In problem investigation, the research goal is to investigate an improvement problem before an artifact is designed and when no requirements for an artifact have been identified yet. The research goal is to improve a problematic situation, and the first task is to identify, describe, explain and evaluate the problem to be treated [Wie14].

There are many ways to investigate implementations and problems, such as reading scientific, professional, and technical literature, and interviewing experts [Wie14]. The method chosen in this project was to conduct a systematic literature review. It was chosen to gain sufficient knowledge about the problem of cyber grooming and state of the art.

3.2.1 Systematic Literature Review

Systematic literature reviews are means to identify, evaluate and interpret research that is deemed relevant for a specific topic according to Kitchenham [Kit04]. Kitchenham lists summarizing existing evidence, identifying gaps in the research topic, and providing background for new research as advantages of using a systematic approach to conduct a literature review. The main reason for conducting a systematic literature review compared to a literature review is that it is more thorough and fair, and thus of greater scientific value. In order to be systematic, it is important that the author identifies and reports research that does not support their hypothesis as well as identifying and reporting research that does so. The main stages of a systematic literature review consist of defining a question, searching for relevant data, extract the relevant data, assess the quality of the data, and analyze and combine the data. The following paragraphs will provide the approach used in this project.

Search Engines and Academic Platforms

Google Scholar[1] is a web search engine freely accessible to everyone. It indexes text and metadata of scholarly literature. Google Scholar was the starting point for gathering information in this project. Through the Google Scholar search engine, several informative platforms were found. Most of these platforms are networking sites made specifically so scientists and researchers can come together and share their work. Some of the main platforms that were used to gather information in this project are:

- Academia
- Academic Journals Database
- IEEE Xplore
- ResearchGate
- ScienceDirect
- Semantic Scholar
- SpringerLink

Flow of Information Gathering

To get a better understanding of cyber grooming and related work in the field of detecting predatory conversations; articles, papers, thesis, websites, news articles, conversation transcripts and related tools were researched. Some of the most used search words were grooming, predator, exploitation, pedophile, perverted, sexual, justice, identification, conversation, chats, cyber, machine, learning, NLP, legislation, laws and online. The words were used either alone or combined with other search

[1] https://scholar.google.com/

words. Whenever a relevant source of information was found, its referenced material was also investigated to see if the referenced material was relevant for this study. If the referenced material was deemed relevant, then it was also included in this study. In order to include any material, it had to contribute to either answer the research question or to be used as background for the discussion of the research question and hypothesis. Many sources of information were studied during this project, and some of them have been important to understand the topic, but not been included in the reference section. The material in the reference section has been directly used in this study and cited accordingly.

3.3 Treatment Design

The term treatment means for an artifact interacting with a problem context to treat a real-world problem. It differs from the term solution in that treatments may solve a problem only partially or not at all [Wie14]. In order to design a treatment for the problem at hand, requirements are necessary.

A requirement is a property of the treatment desired by some stakeholder, who has committed resources to realize the property. Requirements provide useful guidelines for searching for possible treatments [Wie14]. Three requirements have been designed for the predator detection method:

- **R1:** The method must be automated.
- **R2:** The method must return predictions before the end of the conversations.
- **R3:** The method must measure the results with the use of precision and recall metrics.

If the predator detection is automated and manages to return predictions before the end of a conversation, it contributes to the social context goal of mitigating the societal problem of cyber grooming. Furthermore, if the performance of the method is measured in precision and recall, it is possible to compare it with similar solutions and to view its effectiveness.

The treatment design task in this project includes methods within machine learning and NLP. Machine learning and NLP methods were chosen because they are state of the art within text classification, which is the essence of the design problem. The following subsections present those methods.

3.3.1 Model Implementation

This subsection presents the methods applied to the two datasets from Chapter 4. The different approaches used in this project to solve the predator identification task are presented first, followed by a more detailed explanation of the techniques used in those approaches.

Message-Based Detection

To get familiar with the dataset, its content and different machine learning techniques, the first approach to identify predators was based on single messages. By looking at single messages, it was possible to use the datasets without separating the content into authors or conversations. For the MBD approach, each message was labeled as either belonging to a predator or not. The idea behind it was to train a classifier to recognize all messages produced by a predator as predatory and all other messages as non-predatory. The approach did obviously not go very well as predators write normal messages such as "hi", "cool", "good" and so on, while non-predatory persons also write some messages which are similar to predatory messages. The approach returned poor results and was quickly dismissed.

Conversation-Based Detection

The second approach is based on Villatoro-Tello et al. [VTJGE+12], where the authors proposed a new methodology for solving the problem of sexual predator identification with a two-stage classification system. The first stage consisted of detecting conversations where a predator was involved and the second stage to differentiate between the victim and the predator. Figure 3.2 illustrates their proposed system.

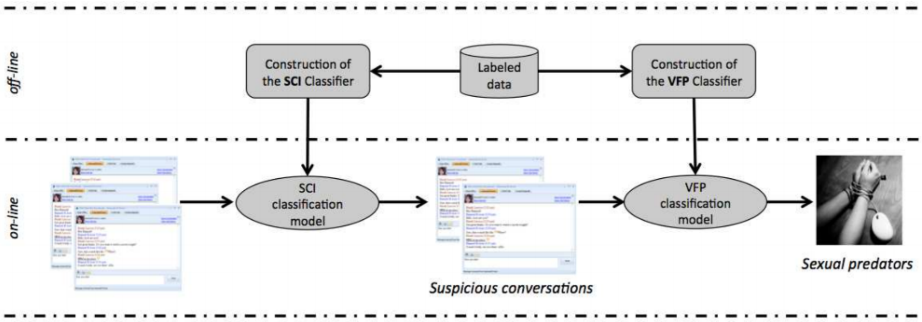


Figure 3.2: General overview of the proposed sexual predators identification system in [VTJGE+12].

In Figure 3.2, SCI stands for Suspicious Conversations Identification and VFP stands for Victim From Predator. Two similar stages have been implemented and tested in this project due to the good results obtained by Villatoro-Tello et al.

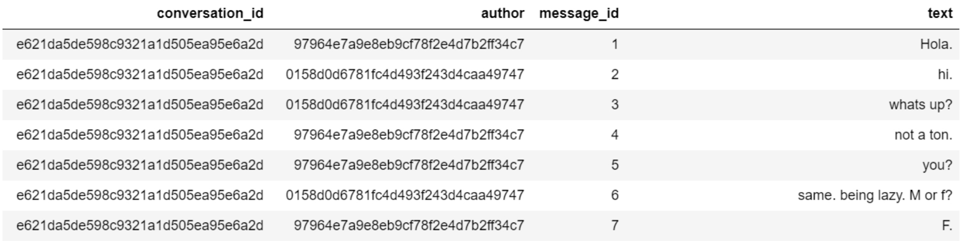
Author-Based Detection

A third approach used in this project to identify predators in online conversations gathered all the messages sent by a single author and classified whether that author was predatory or not. In comparison with the proposed solution in [VTJGE+12], this method can be thought of as a single-stage predatory author identification. Different binary classifiers were used to divide all the authors into two groups, where one of the groups was predatory, and the other was non-predatory.

3.3.2 Data Gathering

The dataset used in this project is described in detail in Chapter 4. It was gathered from PAN[2] and no pre-processing of the text was necessary. The pre-processing was not necessary due to two reasons. The first reason was that the organizers of the PAN 2012 competition had already prepared the data. The second reason was that any further text pre-processing could potentially remove valuable information.

The original dataset is structured as two large eXtensible Markup Language (XML) files. The data was restructured in this project into Pandas DataFrames[3] using the Python programming language for easier data handling. Figure 3.1 shows a snippet of the new structure for the dataset.



conversation_id	author	message_id	text
e621da5de598c9321a1d505ea95e6a2d	97964e7a9e8eb9cf78f2e4d7b2ff34c7	1	Hola.
e621da5de598c9321a1d505ea95e6a2d	0158d0d6781fc4d493f243d4caa49747	2	hi.
e621da5de598c9321a1d505ea95e6a2d	0158d0d6781fc4d493f243d4caa49747	3	whats up?
e621da5de598c9321a1d505ea95e6a2d	97964e7a9e8eb9cf78f2e4d7b2ff34c7	4	not a ton.
e621da5de598c9321a1d505ea95e6a2d	97964e7a9e8eb9cf78f2e4d7b2ff34c7	5	you?
e621da5de598c9321a1d505ea95e6a2d	0158d0d6781fc4d493f243d4caa49747	6	same. being lazy. M or f?
e621da5de598c9321a1d505ea95e6a2d	97964e7a9e8eb9cf78f2e4d7b2ff34c7	7	F.

Table 3.1: A small excerpt of the dataset structured as a DataFrame.

3.3.3 Pre-filtering

Pre-filtering in this project consisted of removing data which is considered irrelevant to identify a predator. Conversations containing exactly two users is an example of one pre-filtering criterion, meaning that all other conversations were removed.

[2] https://pan.webis.de/clef12/pan12-web/author-identification.html

[3] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html

Pre-filtering was an important step to reduce the computational cost of the machine learning algorithms and also to make those algorithms focus on the important information in the datasets.

3.3.4 Pre-processing

Pre-processing was applied on the text within messages and consisted of transforming all letters into lowercase, replacing the characters "/", "(", ")", "{", "}", "[", "]", "|", "@", "," and ";" with spaces, and removing all other characters than letters, numbers, white spaces, "#", "+" and "_". Furthermore, stopwords gathered from the Natural Language Toolkit (NLTK)[4] were removed, and consecutive white spaces transformed into a single white space. Testing was performed with and without the use of pre-processing to investigate whether it could enhance the performance of the predator detection methods.

3.3.5 Data Preparation

In order to more easily review results, a new column called "label" was added to the dataset. The values of the "label" column were either "0" (non-predatory) or "1" (predatory). Some additional data preparation were conducted for the conversation and author based detection approaches. The additional preparation consisted of merging messages sent within a conversation and messages sent by the same authors into longer texts. Those texts were used to make features in the CBD and ABD approaches. The label values for single messages, author merged messages and conversation merged messages were set to "1" if the author id matched the author id of one of the predators. The predatory author ids were provided along with the dataset. Comparing the label with the prediction of a classifier revealed the prediction's correctness.

3.3.6 Features

In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon being observed [Bis06]. In this project, features were made by transforming text from messages into a vector with two different approaches, BoW and TF-IDF. BoW and TF-IDF were chosen because of their simplicity and their great results within text classification tasks. The concepts of BoW and TF-IDF are explained in Subsection 2.3.1.

Bag-Of-Words

BoW was implemented by three steps in this project:

[4] https://www.nltk.org/nltk_data/

1. The N most popular words in each dataset were enumerated and used to make a dictionary. N is in the number of features that are extracted from the BoW method, and they are used to train the different classifiers. Furthermore, N is a parameter which can be adjusted to achieve better results during classification.

2. For each message, author and conversation in the dataset, a zero vector with dimension equal to N was created. Zero vectors were created for messages, authors and conversations to test different approaches in a search for the best results.

3. For each of the approaches, words from the texts were iterated. When a word matches any of the words in the dictionary, the word's corresponding value in the zero vector is increased by one. The process is repeated for every word in a text.

Term Frequency-Inverse Document Frequency

TF-IDF was implemented with the method TfidfVectorizer[5] from the scikit-learn[6] machine learning library. The method was chosen in an addition to the BoW because of its extended features.

3.3.7 Classifiers

A classifier is an algorithm which maps input data to specific categories in order to solve a classification problem. Classifiers were used in this project to predict and classify whether an author was predatory or not. Logistic regressing, ridge, NB, SVM and NN classifiers were chosen, all of which are described in Subsection 2.3.2. Five different classifiers were used to compare their performances. It was not prior knowledge which classifier would perform well. Previous work used different classifiers and different approaches, which made it hard to compare the performances. The classifiers chosen for this project were known to produce good results for binary classification tasks. The classifiers were implemented from the scikit-learn machine learning library.

3.3.8 Cross-validation

Cross-validation is used to estimate the performance of machine learning models on unseen data. It was chosen to estimate the performance of the machine learning models while training on the training data in this project because cross-validation makes it possible to train on all the data. Furthermore, it gives more reliable results because of its multiple iterations. 10-fold cross-validation was used on the training

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[6] https://scikit-learn.org/stable/

data in this project. K=10 is a value that has been found through experimentation to generally result in a model performance estimate with low bias and modest variance [JWHT14], which is the reason why it was used here.

3.4 Treatment Validation

To validate a treatment is to justify that it would contribute to the social context goal if implemented. The goal of validation is to predict how an artifact will interact with its context, without actually observing an implemented artifact in a real-world context [Wie14].

To validate the treatment in this project, it is necessary to show that the requirements of the treatment design are satisfied. To show that the requirements are satisfied, a validation model is used and compared with the model implementation of this project. The work conducted in this project is based on the PAN 2012 competition, which provided a validation model for its contestants. The same validation model is used for this project, such that it is possible to compare the results in this project with the results of the competition. The $\mathrm{F}_{0.5}$-score was the validation of the results of the competition. The following subsection presents the metrics to compute that score.

3.4.1 Performance Measurements

For the evaluation of results obtained from testing, the standard Information Retrieval measure of Precision (P), Recall (R) and harmonic mean between precision and recall (F) was used [Sas07]. Precision and recall is defined as following:

$$Precision = \frac{\text{Number of relevant items retrieved}}{\text{Number of retrieved items}} \tag{3.1}$$

$$Recall = \frac{\text{Number of relevant items retrieved}}{\text{Number of relevant items}} \tag{3.2}$$

The standard F_1 measure, where precision and recall is equally weighted is defined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.3}$$

However, it is not always desirable to use the equally weighted metric. In order to weigh either precision or recall as more important than the other, the following general formula for any positive real β is used:

$$F_\beta = (1 + \beta^2) \cdot (\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}) \tag{3.4}$$

For any positive real β higher than one, recall is emphasized, and if it is lower than one precision is emphasized.

Chapter 4 Dataset

This chapter contains information about where the dataset used in this project was gathered from and characteristics of the dataset. It includes information about where the collection of conversations composing the data was found, how it was assembled, and how the conversations were defined.

4.1 PAN

The dataset used in this project was gathered from [PAN]. *"PAN fosters digital text forensics research by organizing shared task evaluations. Shared tasks are computer science events that invite researchers and practitioners to work on a specific problem of interest, the task"* [PAN]. The Web Technology & Information Systems Network (WEBIS) hosts PAN. WEBIS *"meets challenges of the information society by conducting basic research, developing technology, and implementing and evaluating prototypes for future information systems. Our research contributes to web mining and retrieval, machine learning, computational linguistics, and symbolic AI".*

The dataset was first used in the Sexual Predator Identification Competition of PAN 2012, where PAN was part of the Conference and Labs of the Evaluation Forum (CLEF) in Rome. The goal of the competition was to identify online predators within a collection of conversations, and which lines inside a predatory conversation were the most distinctive for a predator's bad behavior.

4.2 Dataset Characteristics

The dataset is stored as two XML files structured in the format of logical trees. Each tree's root is called "conversations". The roots have many children called "conversation id", which is identified by a unique id for each conversation. Each conversation can have one or more messages called "message line". Each message has a line number. The line number indicates what position that message has within the conversation. The first message of a conversation starts with number one, and this number increases for every posted message. Line numbers can be used together with the conversation id to identify a specific message uniquely. Each message has three children called "author", "time" and "text". "Author" is a unique id for every user that has posted a message. Time tells what hour and minute of the day the message was posted, and text is the actual content of the message. A visual representation of the tree structure is presented in Figure 4.1.

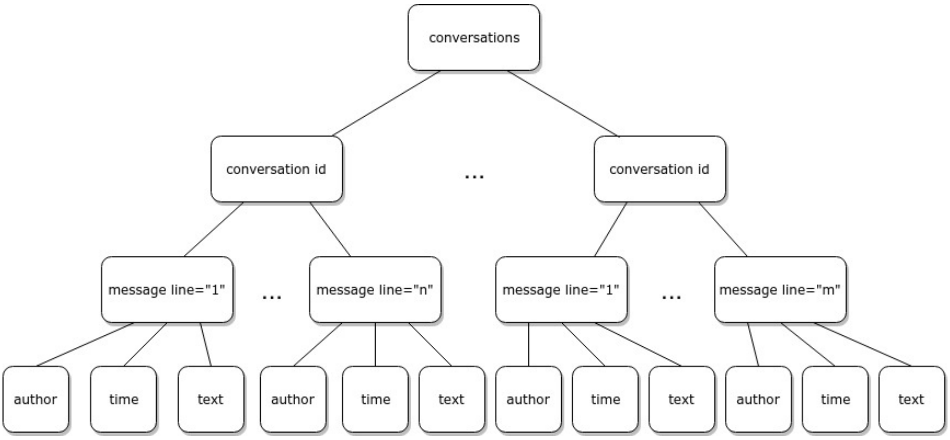


Figure 4.1: Tree structure of the dataset.

The size of the dataset is significant, but it only contains a small number of conversations with a sexual predator. Besides, there are many non-predatory conversations about sex-related topics and a large number of general, non-sex-related conversations. The dataset only contains a small number of predatory conversations to make its environment realistic. The training set consist of exactly 66927 conversations, while the testing set consist of 155128 conversations.

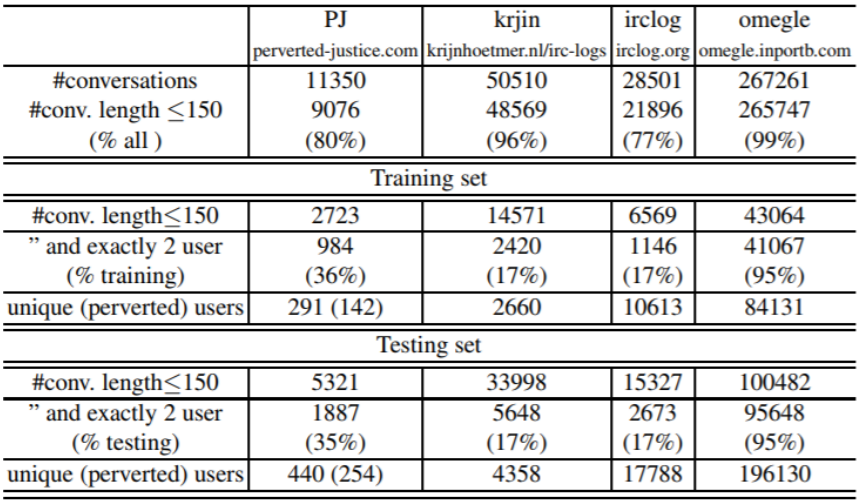
Conversations of interest as described by [Pen07] consist of predatory and non-predatory conversations such as:

- I) Predator/Other interaction
  a) Predator/Victim (victim is underage)
  b) Predator/Pseudo-Victim (volunteer posing as child)
  c) Predator/Pseudo-Victim (law enforcement officer posing as child)

- II) Adult/Adult (consensual relationship)

Class Ia and Ic are not included in this dataset because such conversations are difficult to obtain since they involve police or law enforcement agencies. The lack of such conversations is a limitation of the dataset and thus also a limitation of the entire project. This limitation is addressed in Section 4.4.

The predatory conversations within the dataset were gathered from a website called Perverted Justice (PJ). Those conversations are classified as class Ib, implying that predators chatted with volunteers posing as underage victims. These conversations are the true positives of the dataset and those which should be extracted in the online predator identification task. The PJ website was founded in 2002 by Frank Fencepost and Xavier Von Erick, and it claims that they have convicted 623 predators since their first conviction in June 2004 [Per]. According to Von Erick, they cultivate cooperation with the police and work with the law to get justice. Their work contributed to new laws for online solicitation in the USA and to spread public awareness about adults attempting to sexually assault minors they met online. Their work reached its height when they worked with the National Broadcasting Company who made a television series called "To Catch a Predator" which led to hundreds of convictions and millions of viewers. The program consisted of confronting, interviewing and arresting predators that met up with a decoy under the pretense of sexual conduct after chatting with a pseudo-victim online.



	PJ perverted-justice.com	krjin krijnhoetmer.nl/irc-logs	irclog irclog.org	omegle omegle.inportb.com
#conversations	11350	50510	28501	267261
#conv. length ≤150	9076	48569	21896	265747
(% all)	(80%)	(96%)	(77%)	(99%)
Training set				
#conv. length≤150	2723	14571	6569	43064
” and exactly 2 user	984	2420	1146	41067
(% training)	(36%)	(17%)	(17%)	(95%)
unique (perverted) users	291 (142)	2660	10613	84131
Testing set				
#conv. length≤150	5321	33998	15327	100482
” and exactly 2 user	1887	5648	2673	95648
(% testing)	(35%)	(17%)	(17%)	(95%)
unique (perverted) users	440 (254)	4358	17788	196130

Table 4.1: Properties of the dataset gathered from [IC12].

Class II consisted of conversations from an Omegle repository. Omegle is a website where two strangers have anonymous online conversations. The Omegle conversations included in the dataset is a random sample from more than one million Omegle conversations. Some of the conversations contain abusive language, and some users also engaged in cybersex. Such conversations made the basis for the potential false positives in the dataset. In order to add some variety and also include more true negatives to the dataset topics about general discussions were added. These conversations are Internet Relay Chat (IRC) logs, called krjin and irclog in Table 4.1.

A few notable things were performed to combine all of the different conversations. First of all, a conversation was defined as an exchange of messages where there were no more than 25 minutes of a break between two messages. If there were more than 25 minutes between two messages, those two messages were considered as part of two different conversations. The organizers of PAN 2012 *"empirically observed that this was a reasonable threshold for a topic change in the conversation or the starting of a totally new one"* [IC12]. Furthermore, the organizers noticed that the vast majority of conversations were below 150 messages. According to [IC12], they decided only to include those conversations that were equal to or less than 150 messages. While examining the dataset, the length requirement turned out to not be enforced. Thus, the dataset used in this project also contains conversations with more than 150 messages. Lastly, the organizers generated arbitrary unique ids for each of the conversations and each of the users. The ids also replaced nicknames in the messages, and arbitrary tags replaced email addresses. These measures contributed to anonymize the users to keep their privacy, although the conversations were already public information.

The dataset contains a training set and a testing set, where the training set consists of 30 percent of the entire collection of conversations. The organizers of the competition decided to divide the dataset it that way as the training set was intended for practicing rather than training in the context of machine learning.

4.3 Ethics and Privacy

During the making of this thesis, the Norwegian Centre for Research Data (NSD) was contacted with regards to using the dataset described in this chapter. According to them, there was no need to file an application as long as the data was the same as the data used in PAN 2012. As described earlier, the data from PAN 2012 were anonymized by replacing usernames and nicknames with arbitrary unique ids and emails with arbitrary tags. However, it is important to bear in mind that the data should be used carefully in any future works in order to not break any laws. In addition to this, the organizers of the PAN 2012 competition posted the following message in their task description. *"Given the public nature of the dataset, we ask the participants not to use external or online resources for resolving this task (e.g. search engines) but to extract evidence from the provided datasets only"* [PAN].

The use of a predator identification system on a messaging platform is also a privacy concern. It is not up to the author of this thesis to decide the boundaries or legality of such a system. This project only investigates the feasibility of a functioning predator identification system. However, it is of the author's opinion that a predator identification system handled correctly would be of great societal value. The data handled by such systems would have to be anonymized and only released to the correct authorities.

4.4 Limitations

A limitation of the dataset and thereby also the work in this study is the lack of real underage victims. Data used in both this work and related work stems from conversations between trained professionals posing as children and online perpetrators. Grooming data was collected from the PJ website. Lack of data from conversations between a predator and an actual underage victim could affect the precision of the results. Lack of such data relates to the natural implications of privacy issues. Thus, conversations between a predator and an underage victim are not included in this study.

Chapter 5
Results

This chapter contains the results from the three different approaches, Message-Based Detection (MBD), Conversation-Based Detection (CBD) and Author-Based Detection (ABD), presented in Chapter 3. Each of the approaches used the techniques presented in Section 3.3. Lastly, Section 5.4 presents the result from using the CBD approach for early detection of predatory conversations. The combination of the CBD approach and early detection is the proposed treatment to the design problem in this thesis.

5.1 Message-Based Detection

The idea behind the MBD approach was to detect all messages produced by a predator as predatory based only on one message at a time. The approach was only tested on the training set due to its poor results.

5.1.1 Pre-filtering

According to [IC12], the organizers of the PAN 2012 competition, pre-filtering of unrelated conversations was an important approach used by the top performers in the competition. Pre-filtering was important to get a more balanced dataset with a more even distribution of predators and non-predatory authors. It was therefore decided to use similar pre-filtering techniques in this work. The first round of pre-filtering removed all conversations that only contained one user and all conversations that had less than six messages per user on average. Those criteria reduced the number of conversations with more than 80 %. Five predators were also removed as displayed in the "Filtered data 1" column of Table 5.1. The loss of pre-filtered predators is addressed in Chapter 6. The second round of pre-filtering removed all conversations with more than two users. "Filtered data 2" shows that the number of conversations was further reduced while none of the predators were removed. The last round of pre-filtering removed all empty messages and messages with more than eight characters that did not contain any letters from the English alphabet nor any numbers. "Filtered data 3" displays the result after all of the pre-filtering techniques have been applied.

Number of	Original data	Filtered data 1	Filtered data 2	Filtered data 3
Conversations	66927	12245	8692	8692
Messages	903607	638278	384315	381189
Users	97689	22180	15228	15223
Predators	142	137	137	137

Table 5.1: The number of conversations, messages, users and predators in the training set before and after pre-filtering.

5.1.2 Pre-processing

The MBD approach did not use cross-validation while training the classifiers. Instead, it used the train/test split method with 80 % of the data as training data and 20 % as testing data. After dividing the data, all messages were pre-processed. Table 5.2 displays the ten most common processed and non-processed training and testing words from the pre-filtered training set.

	Raw text		Processed text	
	Training	Testing	Training	Testing
1	i: 34428	to: 9633	u: 29417	omegle: 8311
2	you: 27677	with: 9290	: 26660	u: 6768
3	u: 25025	a: 8695	like: 10189	: 6362
4	to: 22383	i: 8283	lol: 10058	say: 4432
5	a: 20341	Omegle: 8202	im: 9287	apos: 4237
6	the: 16288	the: 8053	ok: 8217	sent: 4140
7	and: 15082	you: 6801	hi: 6434	messages: 4117
8	I: 12345	are: 6380	iapos: 5623	strangers: 4110
9	it: 11967	not: 5980	ur: 5308	claiming: 4103
10	me: 10776	u: 5845	know: 5021	represent: 4103

Table 5.2: The top ten most frequent words before and after pre-processing the training and testing parts of the pre-filtered training set.

Table 5.2 displays how many times the top ten most frequent words were used in the two datasets. In the pre-processed part of the table, the second highest occurrence of the training part and the third highest occurrence of the testing part were empty phrases caused by messages that only contained characters and words that were removed by the pre-processing technique. The empty phrases were caused by the pre-processing and are not the same messages that were removed in the pre-filtering. The number of occurrences for some of the words in the processed part of the table is higher than for the corresponding words in the non-processed part of the table. The higher numbers stem from replacing capital letters with lowercase letters.

Table 5.2 presents mostly common words, which are frequent in regular sentences. However, there are a few words that do stand out, such as "omegle", "iapos" and "apos". The word "omegle" beeing frequent was not surprising considering that the majority of the dataset consists of conversations from an Omegle repository. The reason why "apos" and "iapos" were so frequent was due to the formatting used for some of the conversations in the dataset, where every apostrophe was encoded into "&amp;apos;". The encoding was not reverted in this project because it was the same case for the PAN 2012 competition and changing it could affect the comparison between this study's result and the results of the competition. The characters "&amp;" are the XML encoded version of "&" and is interpreted only as "&". This observation means that "&apos;" becomes the output. After the pre-processing of "&apos;", "&" is removed and "apos" gets connected with the character(s) that is in front of "&". E.g., "i'm" is in some conversations written as "i&amp;apos;", which becomes "i&apos;m", which is transformed into "iapos m" by the pre-processing technique.

5.1.3 Features

After the pre-processing, the messages were transformed into numeric vectors using both the BoW and the TF-IDF methods. The transformation resulted in 5000 features as determined by the 5000 most used words for the BoW method and 18953 features for the TF-IDF method when using unigrams and bigrams, a maximum document frequency of 90 % and a minimum document frequency of five documents. A sparse matrix representation, which is a representation containing mostly zero values, stored the numeric vectors.

Table 5.3 shows the most distinct features for the logistic regression classifier and the ridge regression classifier, respectively. The columns labeled "Positive" contain the highest weighted features towards labeling a message as predatory. The "Negative" columns contain the highest weighted features towards labeling a message as non-predatory. Several negative weighted features ended with "apos", which was due to the same formatting problem as mentioned earlier. The formatting problem only occurred in one part of the imported dataset where all the conversations were non-predatory, which made it a distinctive feature to recognize non-predatory conversations.

	Raw text				**Processed text**			
	LogReg		Ridge		LogReg		Ridge	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
1	sweetie	i&apos;m	ok	hi	sweetie	iapos	ok	hi
2	hun	it&apos;s	:-*	i&apos;m	lil slut	donapos	hun	asl
3	aimee	don&apos;t	lol	you?	aimee	itapos	oh ok	iapos
4	:>	you&apos;re	oh ok	haha	hun	youapos	call	f
5	mwah	i&apos;ll	hun	asl?	mwah	quot	lol	haha
6	i'm	nice to	k	f	b f	thatapos	sweetie	name
7	o ok	that&apos;s	sweetie	asl	ur dad	msn	want	itapos
8	cutie	wat	yes	u?	truck	nice meet	thinking	donapos
9	=p	msn	i'm	m	itll	wat	tonight	nice meet
10	i c	f	o ok	?	cutie	asl	ill	wat

Table 5.3: The highest weighted TF-IDF features for the MBD approach.

5.1.4 Classification

The classification results from the MBD approach using a logistic regression classifier and a ridge regression classifier are presented in Table 5.4. It contains the same metrics as used in PAN 2012, precision, recall, F_1-score and $F_{0.5}$-score. Accuracy is not included for any of the results in this thesis because of the imbalance between predators and non-predatory authors in the dataset. The imbalance of the dataset would result in very high accuracy, e.g., classifying all authors as non-predatory would result in an accuracy of more than 99 % when performed on the full dataset.

	Raw text				**Processed text**			
	LogReg		Ridge		LogReg		Ridge	
	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF
Precision	0.56	0.68	0.17	0.22	0.56	0.62	0.18	0.21
Recall	0.12	0.11	0.79	0.74	0.08	0.07	0.68	0.68
F_1-score	0.20	0.20	0.28	0.33	0.15	0.14	0.29	0.31
$F_{0.5}$-score	0.32	0.34	0.20	0.25	0.21	0.24	0.21	0.24

Table 5.4: Classification results for the MBD approach.

The classification results were not good, as measured by the $F_{0.5}$-score. The bad results were not unexpected. The assumption behind the MBD approach was that all messages produced by a predator could be detected as predatory messages, even without the rest of the conversation. The assumption is obviously not correct. Predators write perfectly normal messages. Furthermore, context is, in most cases, necessary to determine whether a message is predatory or not. The highest achieved result with the logistic regression classifier managed to label 841 out of 7371 messages produced by a predator correctly while misclassifying 398 out of 66997 messages produced by non-predatory authors. The ridge regression classifier with TF-IDF features predicted 5417 out of 7371 messages produced by a predator correctly. However, it also misclassified 19657 out of 66997 messages as predatory. The SVM classifier was also tested for this approach, but it did not work at all, and it ended up classifying every single message as non-predatory.

Due to the poor results, the MBD approach was not used on the testing set. Despite the poor results, it was used as a stepping stone for the CBD approach.

5.2 Conversation-Based Detection

The CBD approach used two classifiers, one to detect predatory conversations and another to differentiate between the victim and the predator in the predatory conversations. The CBD approach reused the pre-filtering and pre-processing techniques from the MBD approach.

5.2.1 Pre-filtering

Since the MBD approach was dismissed before trying it out on the testing set, pre-filtering of the testing set was not performed during that approach. In the CBD approach, pre-filtering of the testing set was performed with the same criteria as used for the training set. The result of pre-filtering the testing set is displayed in Table 5.5.

Number of	**Original data**	**Filtered data**
Conversations	155128	20131
Messages	2058781	837114
Authors	218702	35470
Predators	254	228

Table 5.5: The number of conversations, messages, authors and predators in the testing set before and after pre-filtering.

The pre-filtering of the testing set reduced the number of conversations with 87 % and the number of authors with 84 % while reducing the number of predators with 10 %. The pre-filtering technique removed 26 predators, which made the highest possible recall measure for the PAN 2012 competition slightly less than 0.90. The removed predators are discussed in Subsection 6.2.1.

5.2.2 Pre-processing

The CBD approach used 10-fold cross-validation while training all of the different classifiers. In comparison with the MBD approach, the CBD approach was therefore able to train on all of the training data.

	Raw text		**Processed text**	
	Training	Testing	Training	Testing
1	i: 42711	i: 90935	u: 36185	faggot: 89561
2	you: 34478	you: 77074	like: 12699	u: 75491
3	to: 32016	FAGGOT: 71449	lol: 12132	obama: 69648
4	u: 30870	OBAMA: 69595	im: 11666	little: 32946
5	a: 29036	to: 64054	ok: 10031	boys: 31136
6	the: 24341	u: 62148	omegle: 8831	rape: 30928
7	and: 18638	a: 57864	hi: 8122	im: 25344
8	are: 15472	the: 52174	iapos: 7007	like: 24492
9	I: 15120	and: 43682	say: 6855	lol: 23034
10	it: 14892	I: 36669	ur: 6501	ok: 19200

Table 5.6: The top ten most used words before and after pre-processing the training and testing sets.

Table 5.6 contains a lot of syncategorematic[1] words for both parts of the dataset, which are common in all types of texts. Those are the kind of words that the TF-IDF technique penalizes as they are not important for the meaning of the message. TF-IDF was not implemented until after the pre-processing, meaning that Table 5.6 and Table 5.7 still contain the syncategorematic words.

Table 5.7 contains a subset of Table 5.6, which represents the top ten most frequent words that were written by predators. There are many similarities between Table 5.6 and Table 5.7, and it was not until after the top ten most used words that they diverged. Once again, the number of occurrences for some of the words in the processed part of the tables are higher than for the corresponding words in the non-processed part of the tables. The higher numbers stem from replacing capital letters with lowercase letters. The words that combined the most negative weighted features in the MBD approach occurred in the full dataset, but not in the predatory part, indicating that those negative features were correctly recognized as negative.

[1] https://www.dictionary.com/browse/syncategorematic

	Raw text		**Processed text**	
	Training	Testing	Training	Testing
1	i: 13166	so: 3192	i: 23325	ok: 5132
2	u: 12784	do: 3176	u: 20897	do: 4923
3	to: 7228	and: 3171	to: 12360	so: 4876
4	you: 6427	what: 3112	you: 10772	that: 4772
5	lol: 4917	like: 3091	lol: 7714	my: 4577
6	me: 4430	that: 2816	me: 6919	like: 4438
7	a: 4061	my: 2813	a: 6535	what: 4386
8	it: 3944	I: 2561	it: 6396	in: 3736
9	ok: 3803	not: 2325	and: 5537	be: 3736
10	the: 3559	im: 2285	the: 5325	im: 3653

Table 5.7: The top ten most used words before and after pre-processing the predatory conversations of the training and testing sets

The classifiers in the CBD approach trained without the pre-processing technique because it was discovered during testing that the approach performed better without any pre-processing.

5.2.3 Features

Features in the CBD approach were built using the BoW and TF-IDF techniques. For the suspicious conversation classifier, BoW features were built from a dictionary of the 15000 most frequent words. A total of 17670 TF-IDF features were built with unigrams, bigrams, trigrams, a maximum document frequency of 90 % and a minimum document frequency of 15 documents. The victim from predator classifier used the same parameters except for a reduced minimum document frequency of five documents because there were fewer documents used as input for the second classifier. The total number of features for the TF-IDF approach of the second classifier was 16052, while the number of BoW features remained the same.

5.2.4 Training Suspicious Conversations Classifier

The classifiers used to differentiate between conversations with and without the presence of a predator was implemented using the same BoW and TF-IDF methods as in the MBD approach. However, for the CBD approach, the classifiers were differentiating between entire conversations instead of single messages. Thus, the feature vectors were built from conversations instead of messages.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.95	0.23	0.62	0.94	0.94	0.98	0.98	0.65	0.97	0.97
Recall	0.92	1.00	0.76	0.93	0.95	0.81	0.72	0.76	0.95	0.95
F_1-score	0.94	0.37	0.68	0.93	0.95	0.88	0.83	0.70	0.96	0.96
$F_{0.5}$-score	0.94	0.27	0.64	0.94	0.94	0.94	0.91	0.67	0.97	0.97

Table 5.8: The mean results from of a 10-fold cross-validation when training different classifiers to differentiate between predatory and non-predatory conversations.

Table 5.8 presents the mean precision, recall, F_1-score and $F_{0.5}$-score from a 10-fold cross-validation of training the suspicious conversation classifier. The best results of training classifiers to differentiate between predatory and non-predatory conversations were achieved using NN and SVM classifiers with features from the TF-IDF technique. TF-IDF outperformed BoW for all the different classifiers except from the logistic regression classifier.

5.2.5 Testing Suspicious Conversations Classifier

After training the different classifiers on the entire pre-filtered training set, the classifiers were applied on the testing set without any information about the content inside the testing set. Table 5.9 shows that most of the classifiers performed slightly better on the testing set as compared with the results from training the classifiers.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.98	0.19	0.62	0.95	0.97	1.00	1.00	0.63	1.00	0.99
Recall	0.90	1.00	0.80	0.90	0.92	0.81	0.70	0.78	0.94	0.92
F_1-score	0.94	0.31	0.70	0.93	0.95	0.90	0.83	0.69	0.97	0.96
$F_{0.5}$-score	0.97	0.22	0.65	0.94	0.96	0.96	0.92	0.65	0.99	0.98

Table 5.9: The results from testing the classifiers trained on the training set on the testing set to differentiate between predatory and non-predatory conversations.

The suspicious conversation stage of the CBD approach was intended as a filter between predatory and non-predatory conversation. All the conversations that became labeled as predatory were extracted and then used as input for the second classifier that differentiates between a victim and a predator.

Figure 5.1 and Figure 5.2 present confusion matrices of the two highest $F_{0.5}$-scores from Table 5.9. The confusion matrices display the numbers of correctly and incorrectly predicted conversations. The NN classifier labeled 8 out of 18527 non-predatory conversations as predatory and 130 out of 1604 predatory conversations as non-predatory. The corresponding numbers for the SVM classifier were 3 out of 18527 and 94 out of 1604. The total number of conversations in the pre-filtered testing part of the dataset was 20131. Since the predators in the dataset participated in one or more conversations, incorrectly classified predatory conversations did not necessarily mean that a predator was lost. Whenever a predator participated in more than one conversation, it was usually because the organizers of PAN 2012 had divided a longer conversation into several conversation segments. Some of the conversation segments contained predatory behavior, while others did not. The conversation segments without predatory behavior contributed to the relatively high number of misclassified predatory conversations.

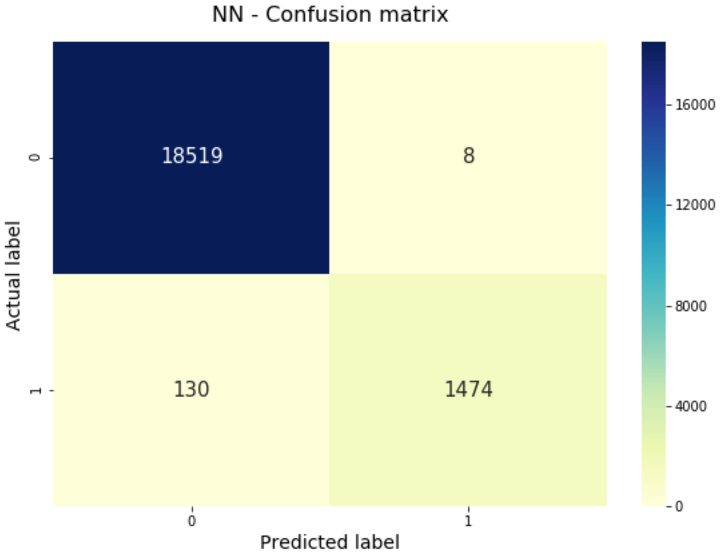


Figure 5.1: Confusion matrix for the result of predicting predatory conversations with a NN classifier.

A low number of false positives was substantial in the first classifier to achieve a high precision score and a high $F_{0.5}$-score. There was more tolerance for a higher number of false negatives for the $F_{0.5}$-score because of the emphasized precision and also because a false positive did not necessarily result in a lost predator because of the divided conversations. Ideally, it would be better with a higher number of false positives and a lower number of false negatives as more predators would remain, while it would still be possible to remove the false positives in the last classification stage.

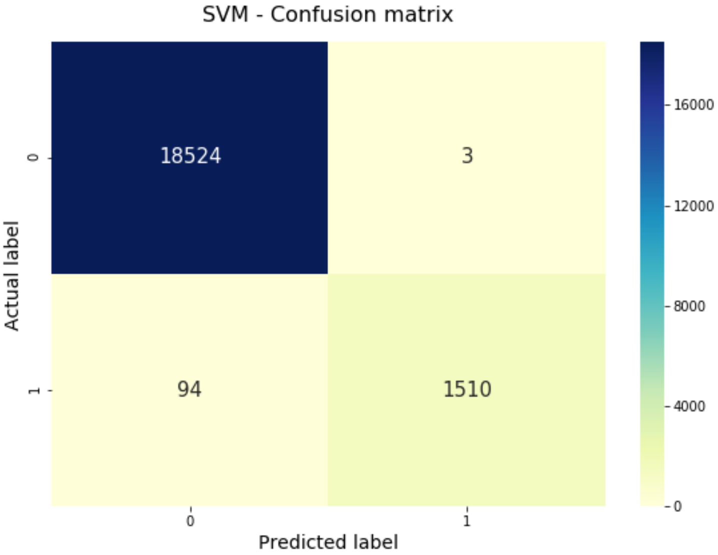


Figure 5.2: Confusion matrix for the result of predicting predatory conversations with an SVM classifier.

5.2.6 Training Victim From Predator Classifier

The second classifier of the CBD approach needed to differentiate between a victim and a predator within a predatory conversation. For non-predatory conversations, the classifier should classify both authors as victims. The datasets did not contain any conversations with more than one predator. If there were any such conversations, they should ideally be detected as non-predatory conversations because neither of the participants would be victims. The victim from predator classifiers trained on all of the predatory conversations from the pre-filtered training set to learn the difference between victims and predators. The victim from predator classifier did not train on the non-predatory conversations.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.92	0.96	0.92	0.91	0.92	0.93	0.94	0.93	0.96	0.95
Recall	0.93	0.79	0.90	0.89	0.93	0.93	0.94	0.80	0.97	0.96
F_1-score	0.92	0.87	0.91	0.90	0.92	0.93	0.94	0.86	0.96	0.96
$F_{0.5}$-score	0.92	0.92	0.92	0.91	0.92	0.93	0.94	0.90	0.96	0.96

Table 5.10: The mean result from a 10-fold cross-validation when training different classifiers to differentiate between victim and predator in predatory conversations.

The training result for the second classifier was good but not as good as for the first classifier. The result is displayed in Table 5.10. It was more difficult to differentiate between victim and predator than to differentiate between predatory and non-predatory conversations. The victim from predator classifier was trained with significantly less data than the suspicious conversation classifier. The dataset did not contain a lot of predatory conversations to train on.

5.2.7 Testing Victim From Predator Classifier

The victim from predator classifier was tested on all of the predatory conversations in the pre-filtered testing set. The classifier was first tested on the conversations that were labeled as predatory by the organizers of PAN 2012. It was tested on those conversations to see how well the classifier could perform without relying on the suspicious conversation classifier. This subsection presents those results, while the next subsection presents how well the victim from predator classifier performed on the conversations labeled as suspicious by the first classifier.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.91	0.97	0.95	0.89	0.90	0.95	0.96	0.96	0.95	0.94
Recall	0.88	0.76	0.84	0.84	0.87	0.91	0.91	0.76	0.92	0.91
F_1-score	0.89	0.85	0.89	0.86	0.89	0.93	0.93	0.85	0.93	0.93
$F_{0.5}$-score	0.90	0.92	0.93	0.88	0.90	0.94	0.95	0.91	0.94	0.93

Table 5.11: The result from testing the victim from predator classifiers on the predatory conversations from the pre-filtered testing set.

Table 5.11 shows that all of the classifiers performed reasonably well and that the ridge classifier outperformed the other classifiers slightly. All of the conversations that were tested on for the victim from predator classifier were predatory with one victim and one predator. A misclassification was not unlikely to result in two wrongly classified authors as the classifier would mix the roles of the victim and predator.

Figure 5.3 and Figure 5.4 present the confusion matrices for the SVM and Ridge classifiers that achieved the highest performance when applied on the testing set. There were more false positives and false negatives for the victim from predator classifier than for the suspicious conversation classifier even though there were more than six times as many conversations for the suspicious conversation classifier.

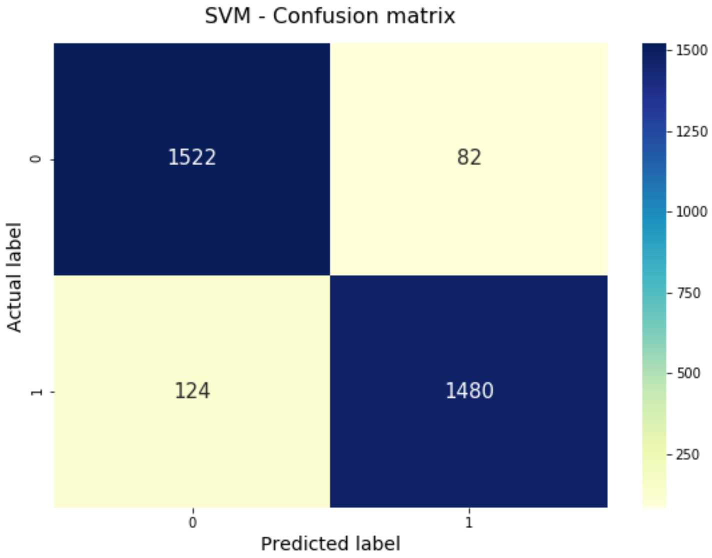


Figure 5.3: Confusion matrix for the result of predicting victim (0) from predator (1) with an SVM classifier.

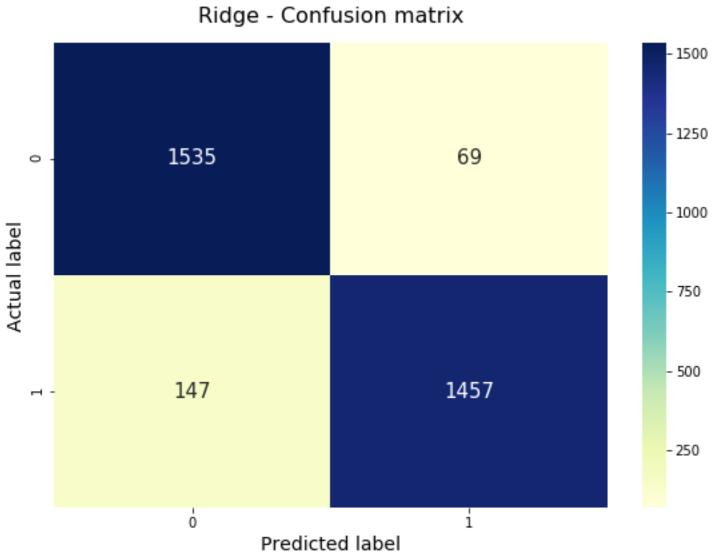


Figure 5.4: Confusion matrix for the result of predicting victim (0) from predator (1) with a Ridge classifier.

5.2.8 Testing Victim From Predator Classifier On Suspicious Conversations

The SVM implemented suspicious conversation classifier predicted 1513 conversations from the pre-filtered testing set to be predatory. The victim from predator classifier used those conversations as input in the two-stage classification system.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.92	0.97	0.95	0.89	0.91	0.95	0.96	0.96	0.95	0.94
Recall	0.88	0.76	0.85	0.85	0.87	0.91	0.91	0.78	0.92	0.92
F_1-score	0.90	0.85	0.90	0.87	0.89	0.93	0.93	0.86	0.94	0.93
$F_{0.5}$-score	0.91	0.92	0.93	0.88	0.90	0.94	0.95	0.91	0.95	0.94

Table 5.12: The results from testing the victim from predator classifiers on the suspicious conversations from the pre-filtered testing set.

As seen from Table 5.12, SVM and Ridge had the best performances in differentiating between victim and predator in suspicious conversations. If the goal were to differentiate between predator and victim in as many conversations as possible, the SVM-based or Ridge-based classifiers would be preferred. However, the goal of the PAN 2012 competition was to detect as many unique predators as possible. The number of correctly differentiated predators from victims was not equal to the number of unique predators because the predators could participate in more than one conversation. Therefore, the results from the two-stage classification system were adjusted for the entire testing set and the number of unique predators classified.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.799	0.898	0.899	0.768	0.781	0.855	0.891	0.913	0.863	0.843
Recall	0.862	0.799	0.843	0.846	0.858	0.862	0.870	0.823	0.866	0.866
F_1-score	0.830	0.846	0.870	0.805	0.818	0.859	0.880	0.865	0.864	0.854
$F_{0.5}$-score	0.811	0.877	0.887	0.782	0.796	0.857	0.887	0.893	0.863	0.847

Table 5.13: The competition results of the two-stage CBD approach to identify unique predators in a corpus.

Table 5.13 presents the scores from the two-stage CBD approach used to identify unique predators in a corpus. The first stage used a SVM classifier, while the classifiers for the second stage and their results combined with the first SVM classifier are presented in the table. The result is presented in the same format as the format used for the PAN 2012 competition.

For the victim from predator classification of unique predators, NB achieved the highest performance, and Ridge regression achieved the second highest. The final result combined the two classification stages of detecting predatory conversations and then differentiating between the victim and predator. The best result in this project was achieved using SVM for the first classification stage and NB for the second classification stage. TF-IDF features obtained better results than BoW features.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
TP	219	203	214	215	218	219	221	209	220	220
FP	55	23	24	65	61	37	27	20	35	41
FN	35	51	40	39	36	35	33	45	34	34
TN	218393	218425	218424	218383	218387	218411	218421	218428	218413	218407

Table 5.14: The confusion matrix values from the classification of unique predators from the CBD approach.

Table 5.14 presents confusion metrics for all of the different classifiers implemented with both BoW and TF-IDF features towards identifying unique predators. T, F, P and N stands for true, false, positive and negative, respectively. TP is the number of correctly classified predators, FP is the number of misclassified predators, FN is the number of predators that were not detected, and TN is all of the correctly classified authors that were not predators. TP, FP, FN and TN were used to calculate precision and recall. The table shows that classifiers implemented with TF-IDF features generally managed to detect more unique predators than classifiers implemented with BoW features.

The NB classifier detected less unique predators than all of the other classifiers. However, it performed better than the other classifiers because it had the least amount of true positives, resulting in higher precision, which the PAN 2012 competition emphasized. The two-stage classification system implemented with an SVM for the first classifier and NB for the second classifier correctly classified 209 out of 254 unique predators and misclassified 20 out of 218702 authors as predators.

Table 5.15 presents the best results of the top ten participants from the PAN 2012 competition. The participants submitted several solutions, but only the best result is part of the table. The competition had 16 participating teams, and the best result achieved an $F_{0.5}$-score of 0.935. The best result in this project was an $F_{0.5}$-score of 0.893, which would place as number three in the competition.

Participant run	RETR.	REL.	P	R	$F_{\beta=1}$	$F_{\beta=0.5}$	Rank
villatorotello-run-2012-06-15-2157g	204	200	0.9804	0.7874	0.8734	0.9346	1
snider12-run-2012-06-16-0032	186	183	0.9839	0.7205	0.8318	0.9168	2
parapar12-run-2012-06-15-0959j	181	170	0.9392	0.6693	0.7816	0.8691	3
morris12-run-2012-06-16-0752-main	159	154	0.9686	0.6063	0.7458	0.8652	4
eriksson12-run-2012-06-15-1949	265	227	0.8566	0.8937	0.8748	0.8638	5
peersman12-run-2012-06-15-1559	170	152	0.8941	0.5984	0.7170	0.8137	6
grozea12-run-2012-06-14-1706b	215	163	0.7581	0.6417	0.6951	0.7316	7
sitarz12-run-2012-0615-1515	218	159	0.7294	0.6260	0.6737	0.7060	8
vartapetiance12-run-2012-06-15-1411	160	99	0.6188	0.3898	0.4783	0.5537	9
kontostathis-run-2012-06-16-0317e	475	170	0.3579	0.6693	0.4664	0.3946	10

Table 5.15: A modified table of the top ten participants from the PAN 2012 competition [IC12]

5.3 Author Based Detection

The ABD approach combined all the messages sent by the same author to one long text. Messages from different conversations were merged to get as much information as possible for one single author. The process was repeated for all of the authors, and each author was connected only to their messages. Thus, only one classification stage was necessary for this approach.

5.3.1 Pre-filtering

The ABD approach reused the pre-filtered data from the two previous approaches. Since the pre-filtering removed conversations with only one author and conversations with less than six messages per user on average, some of the messages produced by different authors were removed. However, those messages were not important.

5.3.2 Pre-processing

The ABD approach used the same pre-processing technique as the two other approaches and achieved better results with pre-processing than without. The highest $F_{0.5}$-score for the ABD approach was 0.891 with pre-processing and 0.883 without pre-processing. For the following results, the pre-processing technique was applied to the ABD approach.

5.3.3 Features

Features were built from the BoW and TF-IDF techniques. BoW used a dictionary size of 20000 words which constructed 35470 numeric vectors representing the number of occurrences for the words in the dictionary. TF-IDF was implemented with unigram, bigram, trigrams, a maximum document frequency of 90 % and a minimum document frequency of 15 documents which resulted in 13584 features.

5.3.4 Classification

The classifiers in the ABD approach trained on all of the pre-filtered messages in the training set. The results in Table 5.16 stems from a 10-fold cross-validation.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.81	0.02	0.22	0.81	0.82	0.86	0.92	0.25	0.91	0.92
Recall	0.73	0.99	0.67	0.77	0.80	0.35	0.51	0.82	0.79	0.83
F_1-score	0.76	0.04	0.33	0.78	0.80	0.48	0.63	0.39	0.84	0.87
$F_{0.5}$-score	0.79	0.03	0.25	0.80	0.82	0.66	0.79	0.29	0.88	0.90

Table 5.16: Results from training the ABD approach on the pre-filtered training set.

Table 5.16 presents the training results from the ABD approach on the pre-filtered training set. All but one of the classifiers with features built from TF-IDF performed better than those that were built from BoW, the exception was for the logistic regression classifier. NN obtained the best performance, while SVM came second. The NB implemented classifiers and the Ridge classifier built on BoW features did not produce usable results.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.93	0.01	0.18	0.84	0.90	0.94	0.95	0.21	0.96	0.98
Recall	0.54	1.00	0.63	0.60	0.82	0.27	0.44	0.71	0.70	0.72
F_1-score	0.69	0.03	0.28	0.70	0.86	0.42	0.60	0.32	0.81	0.83
$F_{0.5}$-score	0.81	0.02	0.21	0.77	0.88	0.62	0.77	0.24	0.90	0.92

Table 5.17: Results from testing the ABD approach on the pre-filtered testing set.

Table 5.17 presents the results from testing the ABD approach on the testing set. The testing results were slightly better than the training results. The best performance was achieved with the TF-IDF based NN classifier.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
Precision	0.925	0.014	0.179	0.835	0.903	0.938	0.953	0.210	0.964	0.982
Recall	0.488	0.894	0.567	0.539	0.732	0.240	0.398	0.642	0.626	0.650
F_1-score	0.639	0.028	0.272	0.656	0.809	0.382	0.561	0.317	0.759	0.782
$F_{0.5}$-score	0.785	0.018	0.207	0.753	0.863	0.593	0.745	0.243	0.870	0.891

Table 5.18: Results from testing the ABD approach on the entire dataset in adherence with the PAN 2012 competition format.

Table 5.18 presents the results of the ABD approach on the same format as the PAN 2012 competition. The highest $F_{0.5}$-score from the ABD approach was 0.002 lower than for the CBD approach. The ABD approach achieved a very high precision score with the NN classifier. The precision for the NN implemented classifier of the ABD approach achieved a score of 0.982, which was 0.069 higher than for the CBD approach. The ABD approach had a considerably lower recall score. It was only 0.650 compared to 0.823 for the CBD approach.

	BoW					TF-IDF				
	LogReg	Ridge	NB	SVM	NN	LogReg	Ridge	NB	SVM	NN
TP	124	227	144	137	186	61	101	163	159	165
FP	10	15730	662	27	20	4	5	613	6	3
FN	130	27	110	117	68	193	153	91	95	89
TN	218438	202718	217786	218421	218428	218444	218443	217835	218442	218445

Table 5.19: The values of the confusion matrices from the classification of unique predators with the ABD approach.

Table 5.19 presents the predictions for the different classifiers towards labeling different authors as unique predators. The number of correctly predicted predators were considerably lower than for the CBD approach. Despite the low number of true positive, the ABD approach achieved a similar $F_{0.5}$-score as the CBD approach due to its low number of false positives.

5.4 Early Detection

The CBD approach obtained the highest $F_{0.5}$-score and the highest number of detected predators among the three presented approaches and was selected for the early detection task. This section presents results from using the CBD approach on an incremental number of messages to identify the predators in an early phase of the conversations. The two-stage classification approach trained on all of the pre-filtered data from the training set as in the previous sections. The suspicious conversation classifier trained on all of the pre-filtered training data, and the victim from predator classifier trained on all of the predatory conversations from the pre-filtered training data. Predator classification was performed on all of the pre-filtered conversations N times. N represents the number of iterations, which is equal to the number of messages that the classifiers tested on. The first classification was performed with only one message in each conversation, then two messages and after that three messages, until it reached N messages. From experimenting with different values of N, N=100 was observed to be enough messages for the classifiers to stabilize.

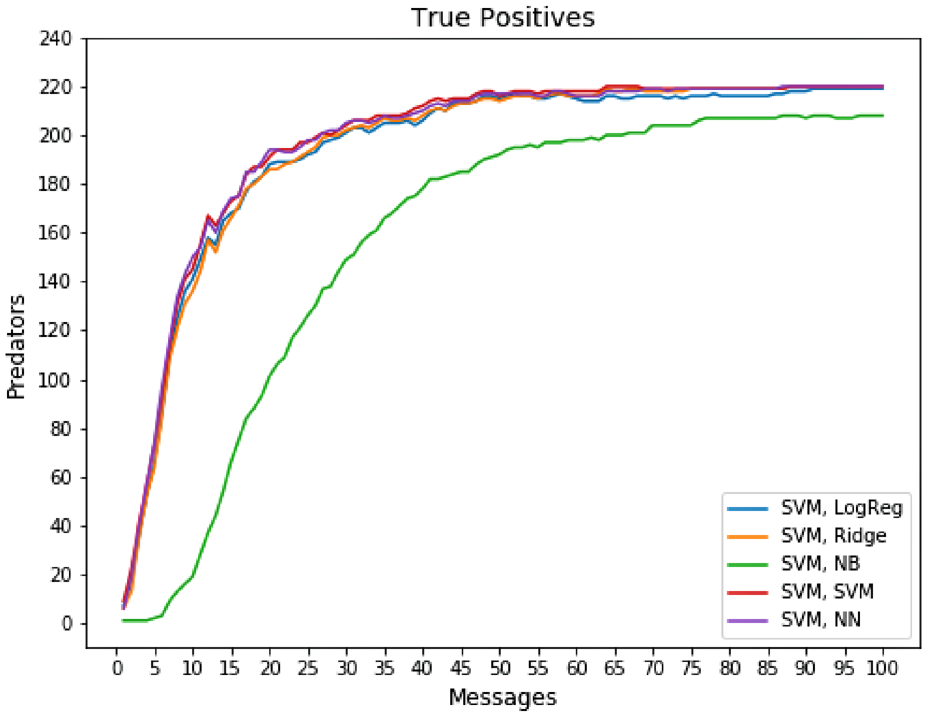


Figure 5.5: The number of unique predators correctly classified based on the number of messages provided during early detection.

Figure 5.5 presents the number of unique predators that were correctly classified by the TF-IDF implemented classifiers. The TF-IDF implemented classifiers generally performed better than the BoW implemented classifiers and are therefore presented. The legend in the lower right corner shows which classifiers that were used for the two classification stages. The suspicious conversation classifier was implemented with SVM in all of the experiments. The Ridge, SVM and NN classifiers detected

220 unique predators as their best results. Logistic regression detected 219 and NB detected 208. The corresponding values for the classification of full-size conversations were 221, 220, 220, 219 and 209. All of the classifiers detected 200 predators within 30 messages except from the NB classifier which did so after 63 messages.

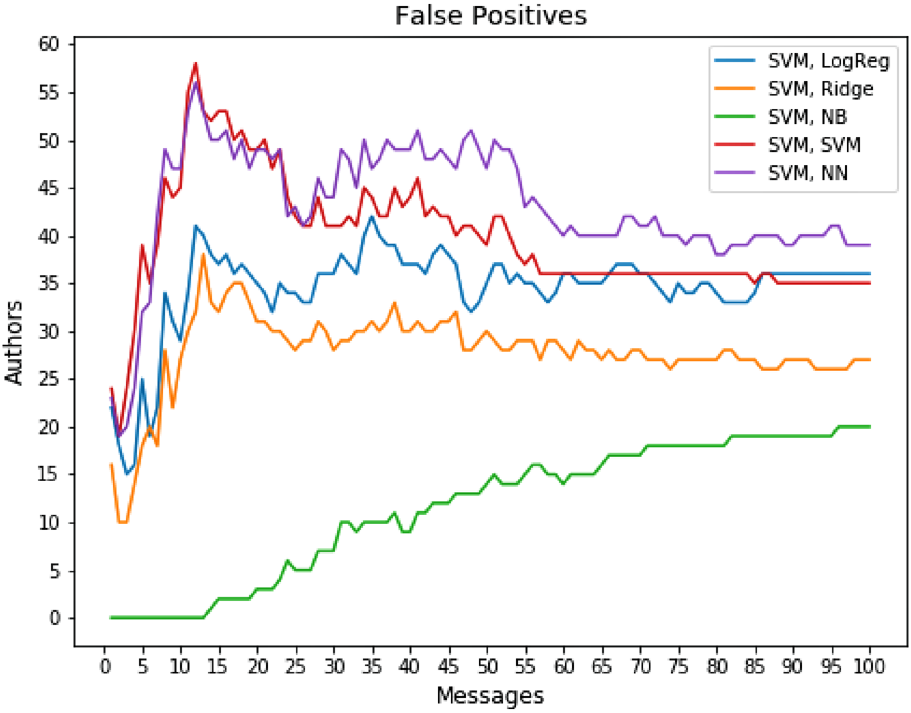


Figure 5.6: The number of unique authors misclassified as predators based on the number of messages provided during early detection.

Figure 5.6 presents the number of non-predatory authors misclassified as predators. The classifiers started by misclassifying a low number of authors when there were few messages for each conversation used as input. The classifiers were more careful at the beginning of a conversation than later on because there was usually not enough information to go on. Furthermore, in most cases, the first classification stage had not evaluated the conversations as suspicious. Therefore, the conversations were not evaluated by the second classifier. When the classifiers received more information, both the number of true positives and false positives increased until a certain point where they had enough information to classify most of the authors correctly. After this point, as the classifiers received more information than earlier, the number of false positives slowly decreased. The exception was the NB classifier, which used more time to detect the same number of predators as the other classifiers. The number of false positives for the NB classifier slowly increased with its number of true positives.

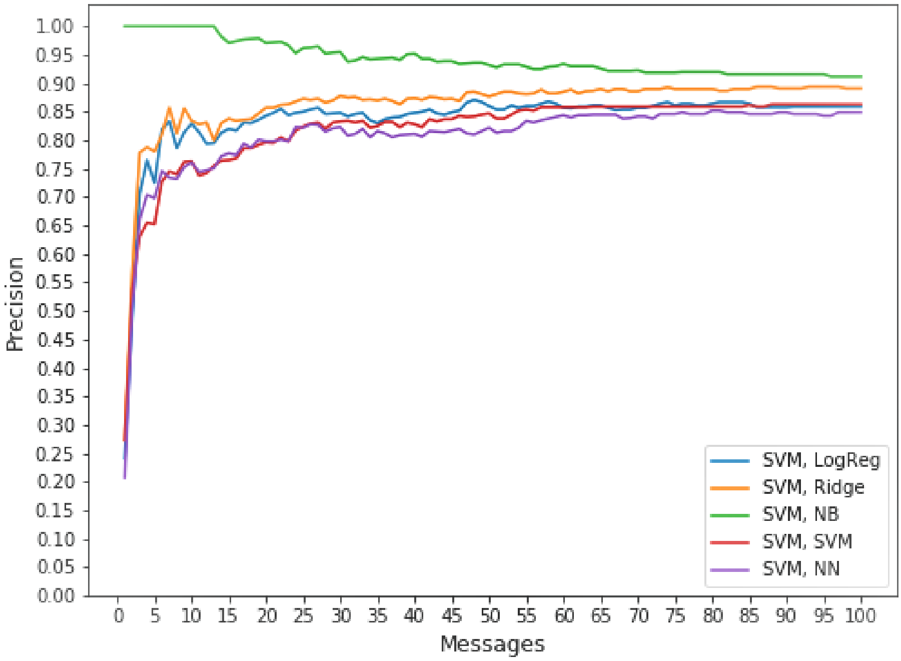


Figure 5.7: The precision score for classification of unique predators based on the number of messages provided during early detection.

Figure 5.7 presents the precision score for the classification of predatory authors. Precision is the number of true positives divided by the sum of true positives and false positives. The precision score reflects how many of the predicted predators that were predators. Due to the NB classifier's more careful approach, it achieves a higher precision score than the other classifiers. The careful approach makes the NB implementation more useful in an early detection approach because the results are more reliable than the results of the other classifiers. The difference that makes NB achieve a higher precision score than the other classifier is because it classifies both victim and predator as victims when it does not have enough information to tell which author is the predator. The other classifiers try to guess since the classifiers trained on predatory conversations where half of the authors are predators. The downside with the NB classifier is that it needs more messages to detect the same amount of predators as the other classifiers. Furthermore, it is not able to detect as high total number of predators as the other classifiers. The precision was higher than 0.8 for all classifiers after 21 messages, and the precision score was not lower than 0.91 for the NB classifier.

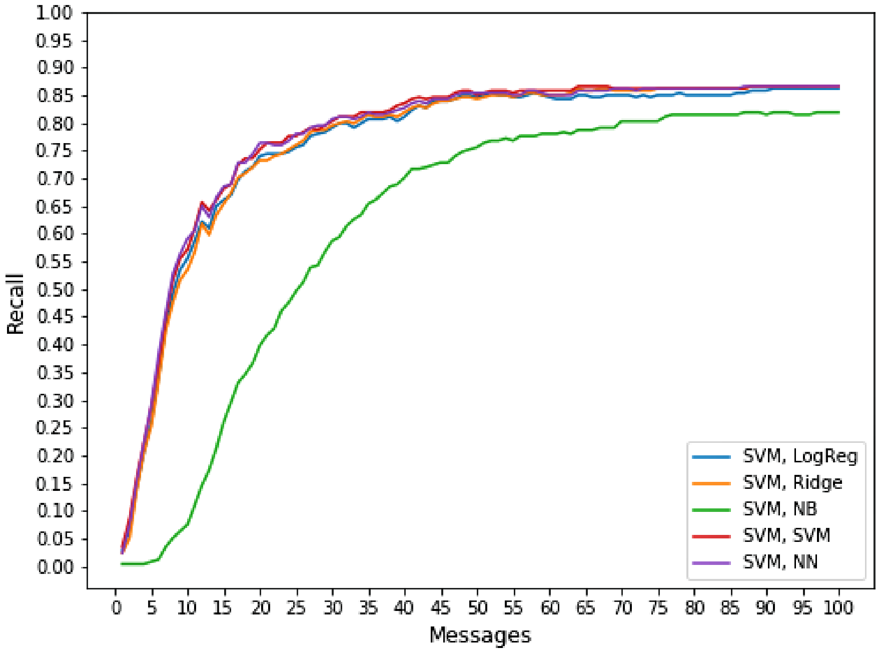


Figure 5.8: The recall score for classification of unique predators based on the number of messages provided during early detection.

Figure 5.8 presents the recall score for the classification of predatory authors. The recall score is a scaled version of the true positives, where the only difference between Figure 5.5 and Figure 5.8 is that the Y-axis is measured from zero to one instead of numbers of predators. The recall is the number of true positives divided by the sum of true positives and false negatives. The sum of true positives and false negatives is constant and equal to the number of predators in the testing set. The recall score reflects how many percentages of the predators that are detected. The logistic regression, ridge, SVM and NN classifiers achieved a recall score above 0.8 within 36 messages, while the NB classifier needed 70 messages.

Figure 5.9 presents the F_1-score for the classification of unique predators. The F_1-score equally weighs precision and recall. The Ridge implementation of the victim from predator classifier achieved the highest F_1-score of 0.88. The Logistic regression, Ridge, SVM and NN classifiers reached a score of 0.8 between message 24 and 25, while the NB classifier needed 40 messages to achieve the same score.

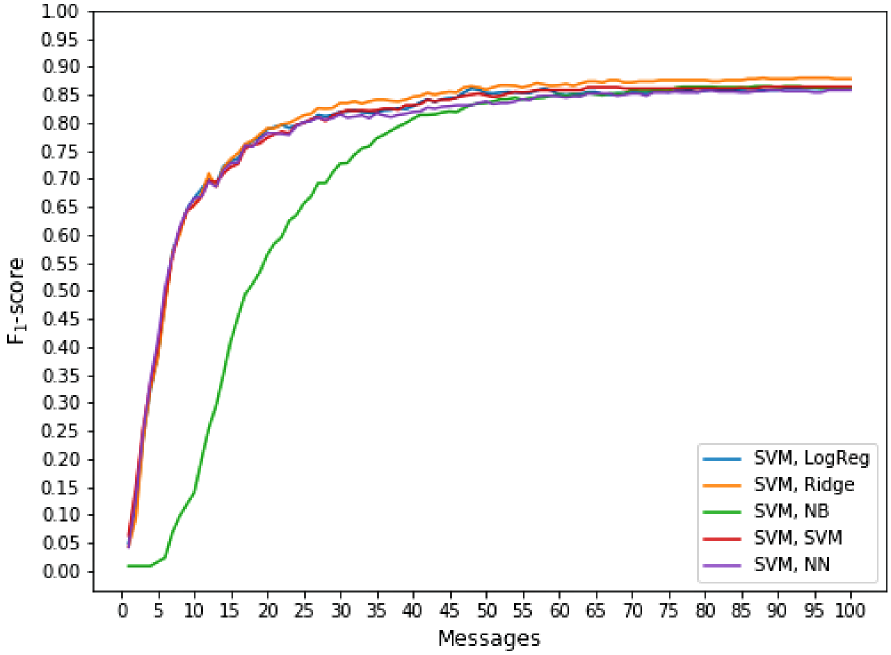


Figure 5.9: F_1-score for classification of unique predators based on the number of messages provided during early detection.

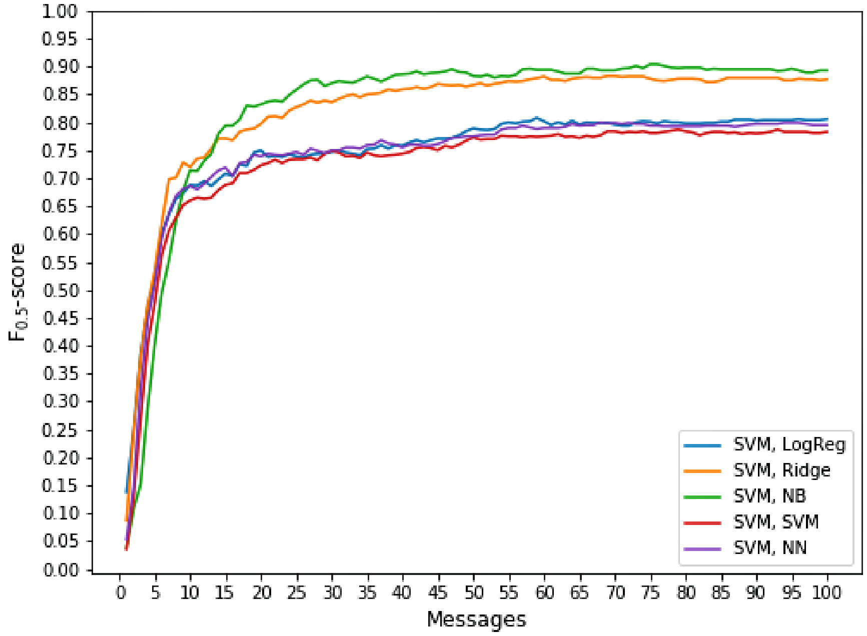


Figure 5.10: $\mathrm{F}_{0.5}$-score for classification of unique predators based on the number of messages provided during early detection.

Figure 5.10 presents the $F_{0.5}$-score for the classification of unique predators. The $F_{0.5}$-score emphasizes precision, which is the score that the PAN 2012 competition used to rank the submissions. The NB classifier achieved the highest $F_{0.5}$-score, which was 0.898. All of the classifiers reached a score higher than 0.8 after 24 messages.

The early detection implemented with the CBD approach fulfilled all of the requirements from Section 3.3. The early detection was automated, returned predictions before the end of the conversations and used precision and recall to measure the results.

5.4.1 Full-length Predatory Conversations

The early detection with the CBD approach was further tested on ten full-length conversations from the PJ website. The tests were similar to the work in [Pen07] and [EEL10], where full-length conversations from the PJ website were used to differentiate between the victim and the predator in predatory conversations. The work on full-length predatory conversations in this project differs from the previous work by including early detection. The CBD approach was tested on full-length conversations to see if there were any differences from the shorter conversation segments and also to see at what message the conversations turned into suspicious conversations and at what message the victim and predator were correctly identified. The tests were conducted with the NB implemented CBD approach from Section 5.2.

Conversation	1	2	3	4	5	6	7	8	9	10
Suspicious	7	58	49	89	35	39	16	26	32	48
Correct	37	58	49	89	46	128	58	26	32	-
Stable	37	114	125	89	46	128	63	26	161	-
Length	4937	2465	2573	1230	6023	3629	3743	766	764	4348

Table 5.20: Suspicious conversation identification and victim from predator detection of full-length predatory conversations.

Table 5.20 presents the message lines where the ten full-length predatory conversations were detected as suspicious, where the victim and the predator were correctly classified and where it stabilized. The row called "stable" in the table indicates the point where both victim and predator were correctly classified, and no later messages changed that classification. Conversations with different values for "correct" and "stable" were correctly classified, but later messages changed the classification until it again was correct and the classification no longer changed.

The tenth conversation did not correctly classify the victim, not even at message line 4347, which was the last message in the conversation. Both victim and predator were labeled as predators. An interesting observation from that conversation was that the victim was leading the conversation, which was not the case in the other conversations. The predator was passive and mostly answered the victim's question. Furthermore, the predator stated his concern for undercover police several times throughout the conversation. Even so, the victim stated her age of 13 and wrote several messages supporting that statement. A human review of the conversation would have correctly classified the victim.

The ten analyzed conversations all contained age, sex and location of both the victim and the predator. All of the predators frequently encouraged the victim with compliments. The victims repeatedly gave indications of their young age, and often mentioned their parents. Most of the predators screened their victim before making a predatory approach. However, it was not the case for all of them. In two of the conversations, the predator made the predatory approach from the first message. The victims wrote less formal than the predators, and they used much more emoticons. The length of the conversations varied between 764 and 4937 messages.

The two-stage CBD approach marked all of the full-length conversations as suspicious within 90 messages. The victim and predator were correctly classified within 130 messages and stabilized within 170 messages for nine out of ten conversations. Stable classification of full-length conversations needed more messages than for the conversation segments in PAN 2012.

Chapter 6 Discussion

This chapter discusses the results presented in Chapter 5. Unexpected results were investigated to find out why they occurred. The discussion includes the unexpected results.

6.1 Message Based Detection

Even though the MBD approach proved not to be a good way to solve the predator identification problem, the techniques used in the MBD worked the way they were supposed to. This section presents a discussion of the results from applying those techniques on single messages. The discussion of pre-filtering and pre-processing from the MBD approach is also relevant for the other approaches.

6.1.1 Pre-filtering

The pre-filtering performed in the MBD approach utilized different filters to reduce the dataset. The first round of the pre-filtering removed conversations with only one author and conversations with few message exchanges. Conversations with only one author were assumed to be unsuccessful attempts to start a conversation. The pre-filtering removed short conversation because of the difficulty of determining a predator from a tiny amount of information. The second round of the pre-filtering removed all conversations with more than two authors since this project does not focus on group conversations. The third round of pre-filtering filtered out messages containing long sequences of special characters and empty messages as they did not provide value to the machine learning algorithms. The last round of pre-filtering removed 3126 messages, which was a small amount compared to the total number of messages in the dataset. The last round of pre-filtering was not necessary, but the removal of empty messages made it easier to work with the data.

The organizers of PAN 2012 divided the conversations from the Perverted Justice (PJ) website into several smaller conversation segments, as described in Section

4.1. They divided conversations containing consecutive messages sent more than 25 minutes apart. Several of the predatory conversation segments lost all of its predatory content, and the pre-filtering did not recognize them as predatory. However, it does not explain the five removed predators. Conversation segments without predatory content would likely pass undetected, yet the predator should still appear in some of the predatory conversation segments.

C_id	Author	M_id	Text
1	A	1	so wat now ?
1	A	2	????
2	B	1	im sorry but im not comming! Good bye!
3	C	1	hey are you going to be on today.
4	C	1	HEY DON'T WANNA TALK NO MORE
4	C	2	WHATS GOING ON
5	C	1	HEY HOW ARE YOU DOING
6	C	1	hey how are you doing is everthing fine, reply back.
7	D	1	hi
7	D	2	???
8	C	1	are you ok, what did you do. I was there driving around and waited but i had to drive back to get to work. So is he going to be there this weekend, what are you going to do. can you you use the computer now.
9	C	1	hey you there
10	C	1	hey where were you u didn't call me
11	E	1	pm me sometime i miss u
12	E	1	sorry i havent been around lately i hadda go outta town for awhile
13	C	1	yeah I'm here
13	C	2	how have you been doing. is everything ok.
14	C	1	hey liz are u ok

Table 6.1: All of the removed predatory conversations, predators and messages written by predators from the training set.

Table 6.1 displays all the lost messages from the training set that were written by predators. "C_id" and "M_id" stand for conversation id and message id, respectively. Conversation ids and author ids were changed from unique strings to single numbers and letters due to their length. Even though the conversations were anonymized, it was possible to retrace individual conversations by searching for distinctive messages. To find out why some of the predators were lost, the entire conversation of a lost predator from the PJ website was compared to the content of the dataset. It turned out that the remaining messages, that are not displayed in Figure 6.1, sent by the author "C" were not included in any of the datasets. The reason behind the exclusion is uncertain. However, the full conversation contained more than 150 messages without any 25 minutes breaks. The organizers of PAN 2012 claimed that they did not include such conversations as previously explained in Chapter 4, yet there are several examples of both predatory and non-predatory conversations longer than 150 messages. Short segments of a predatory conversation should not be included without the remaining segments of the conversation. Without the remaining conversation segments, the classifiers may end up training only on a predator's non-predatory behavior. The non-predatory conversation segments from predators impacted the classifiers and the final scores negatively. The content of the messages displayed in Figure 6.1 does not explicitly indicate grooming. Those messages are something regular users could write, which is the reason why it is important to include all of the conversation segments of a conversation.

Author "B" from Table 6.1 produced only one of the messages from the entire dataset, which was *"im sorry but im not comming! Good bye!"*. The authors "B", "D" and "E" only produced two messages each and author "C" produced 11 messages in total. All of these authors were labeled as predators. Those predators should not be used as predators in the dataset with so few messages.

6.1.2 Pre-processing

The pre-processing technique removed text that turned out to be distinctive features, such as emoticons, which decreased the performance of the classifiers. The decrease in performance indicates that the semantics of the writing for a predator is a part of their distinctive behavior and that any pre-processing of the text is likely to harm the results. The pre-processing also revealed the formatting problems addressed in Section 5.1.2.

6.1.3 Features

The encoding of apostrophes in some of the non-predatory conversations, as described in Section 5.1.2, impacted the classifiers. The encoding made several words very distinctive for non-predatory behavior, and most or all of the conversations that contained the encoding were likely labeled as non-predatory. The encoding made the classifiers focus on the encoded word features making it harder to correctly classify the other non-predatory conversations that did not contain the encoding. The failed encoding was unfortunate and should have been fixed by the PAN 2012 organizers.

Encouraging and complimentary words such as "cutie", "sweetie", "sweety" and "hun" were among the most positive features towards labeling an author as a predator. Manual analysis revealed that such words were widespread, but not exclusive among predators. The word "mom" was used within two bigrams, where the first bigram was one of the most negative features, and the second bigram was one of the most positive features. The negative one was "my mom", while the positive was "your mom", which catches the different focus of the victim and the predator. Moms are central and vital in a minors life, but the victim's mom is also a concern for the predator. The word "omegle" was also a negative feature, which is not surprising considering that none of the predatory conversations were gathered from Omegle.

6.1.4 Classification

The classification results were poor because they only focused on single messages, and not all messages written by predators are predatory. However, the MBD approach was still interesting because it provided information about the most distinctive predatory messages which was not as relevant in the CBD and ABD approaches. Furthermore, an implementation of the MBD approach where a certain percentage of an author's messages need to be detected as predatory could work. Majority voting or a threshold could evaluate whether an author was predatory or non-predatory. However, these suggestions were not implemented in this thesis, as more promising approaches were investigated instead.

6.2 Conversation Based Detection

The CBD approach performed far better than the MBD approach and ended up with decent results towards identifying unique predators in a collection of conversations. This section discusses the results from the CBD approach.

6.2.1 Pre-filtering

Pre-filtering was performed on the testing part of the dataset because of its large size. Besides, this project focuses primarily on conversations with exactly two authors, which was not the case for all of the conversations in the dataset. Furthermore, the best submission of the PAN 2012 competition [VTJGE$^+$12] also performed pre-filtering on the testing part of the dataset.

The pre-filtering technique used for the CBD approach was the same as the technique used for the MBD approach. The majority of the conversations that were filtered out from pre-filtering the testing part of the dataset had only one author. The filtered conversations that contained two authors consisted of very few messages. The longest predatory conversation that was filtered out contained ten messages, where three of the messages were empty. Thus, it was the same case as described for the training part of the dataset where there were not enough messages sent by the filtered predators for them to be detected.

6.2.2 Pre-processing

The pre-processing technique used for the CBD approach reused the technique from the MBD approach, and it turned out that both approaches performed better without the use of any pre-processing. The pre-processing removed most of the special characters which were frequently used in predatory conversations to make emoticons. Emoticons were among the top positive features towards labeling a conversation as predatory.

During pre-processing, some unnatural words were discovered among the top ten most used words. The words "obama" and "faggot" were present among the top ten most used words in the testing set, as shown in Table 5.6. Those two words were among the top ten most used words because they were used a large number of times in successive messages from the same authors. One author wrote "OBAMA" 69536 times within a single conversation while another wrote "FAGGOT" 71362 times within a single conversation. Special cases concerning only a few conversations did not impact the classification process.

6.2.3 Features

The CBD approach used different features in the two classification stages. Some of the most distinctive features for the suspicious conversation classifiers were emoticons. Some emoticons weighed positively towards labeling a conversation as suspicious, while others did the opposite. The most positive emoticon towards labeling a conversation as suspicious was ":-*". Emoticons were frequently used in predatory conversations, more frequent than for the non-predatory conversations. Some of the most positive weighed words for predatory conversations were "kewl", "dad", nite", "kiss" and "sux". The words "kewl", "nite" and "sux" were mostly used by the victims in the predatory conversations. Among the negative weighed features were words with the failed encoding.

The victim from predator classifier had many emoticons among the most negative features towards labeling an author as a predator. This observation indicates that it is the victim that is using emoticons most frequently. The only exception for the emoticons was ";)", which was a positive predatory feature. The features "kewl", "lol", "k", "u 2", "wat", "my mum", "my dad" and "dunno" were also among the most negative weighed features. The positive weighed features included "call me", "are u", "love you", "talk to me", "sweetie", "sweety", "baby", "sexy", "love", "work", "horny" and "doing". The pseudo-victims wrote more informally than then predators and often tried to use slang that is associated with a young age.

6.2.4 Classification

The testing set contained 254 predators, after the pre-filtering phase 228 predators remained. Six more predators went undetected by the suspicious conversation classifier. The NB implemented victim from predator classifier correctly identified 209 predators and misclassified 13 of the 222 remaining predators. The pre-filtering process was the phase where most of the predators were lost. However, the messages of those predators had not enough information to detect them. It is more important to look at the predators that were missed during the two-stage classification and also the non-predatory authors misclassified as predators.

For the suspicious conversation stage, all of the six lost predators participated in one conversation each. Two of the six predators produced messages that should have labeled the conversation as suspicious and them as predators. Those message exchanges clearly stated the age of both victim and predator and the predators asked inappropriate and sexual oriented questions. For the four other predators, one produced only a few and general messages, one was very careful and asked a lot about the police, one did not indicate the age of the victim and one conversation only included daily life conversation topics.

The conversations of the two predators where the conversation should have been labeled as suspicious were both very direct, and the victim actively engaged in sex-oriented talk. The two predators were not cautious at all and made their sexual intentions very clear. Those two conversations may have been mistaken for the non-predatory, but sexual conversations from Omegle, which were similar in structure and content. The difference that the classifiers may not have been able to detect was the written age of the victim. Typical behavior for the pseudo-victims seems to be that they play along with the pace of the predators. The observed behavior might have resulted in a different structure for the two undetected predatory conversations. Based on those two conversations, the second hypothesis from Chapter 1 is clearly false. The second hypothesis suggested that predators must build relations with the victims before they attempt to groom them. While this seems to be the case for the majority of the predatory conversations, it is not valid in all cases.

For the victim from predator classification, the 13 undetected predators participated in 1 to 20 conversations. A manual review of the conversations revealed that five of the predators should have been predicted as predators. The eight others could not manually be predicted as predators with satisfactory certainty. Some of the predators only exchanged messages without any predatory-like content. Those predators should not have been labeled as predators because of the missing conversation segments excluded by the organizers of the PAN 2012 competition. From the 20 wrongfully labeled non-predatory authors, 2 were not in a suspicious conversation. One of those two participated only in one conversation, which was through Omegle. That person was 19 years old while the conversation partner was 12 years old. However, the conversation was general and did not indicate grooming. The second person participated in many conversations and produced 20323 messages. Those messages were not manually reviewed, but they seemed to be of technical orientation. The other 18 misclassified authors were all pseudo-victims in one or more predatory conversations. The pseudo-victims seems to have experience from participating in a lot of predatory conversations, which they have used to develop a specific way to speak. Their language mimics how the predators speak to some extent, but seems mostly to reflect how the pseudo-victims believe that a child behaves.

From the manual analysis of misclassified authors, it was observed that the pseudo-victims are usually easy to recognize as "underage" based on how they write and that they often purposely state their age to use the conversations as evidence. However, it is not always clear what age the other person is. Furthermore, pseudo-victims used emoticons, netspeak, abbreviations and terms of endearment frequently. This behavior is similar to the behavior the predators often try to replicate from children to connect with the victim. The behavior might be part of the reason why the victim from predator classifier sometimes switches the roles of the victim and the predator. The pseudo-victims are also often baiting the predators, which is similar to how predators encourage the victims.

The suspicious conversation classifier performed very well, and in the way it was intended. The low number of false positives from identifying predatory conversations was essential to achieve a high $F_{0.5}$-score. There were two main reasons for the importance of the low number of false positives. Firstly, false negatives of predatory conversations did not necessarily lead to a false negative for unique predators while false positives were more likely to add a false positive to the number of unique predators. This was a result of the divided predatory conversations. Secondly, because of the emphasized precision, false positives were penalized harder than false negatives.

The victim from predator classifier missed more predators than the suspicious conversation detection classifier. The victim from predator classifier had less data to train on than the suspicious conversation classifier. Furthermore, a significant portion of the predatory conversations was incomplete. The predators' and victims' messages in the predatory conversations were also more similar than the messages in predatory and non-predatory conversations. Considering the scores of the two classifiers in the CBD approach, the victim from predator classifier should be prioritized to improve the detection system.

6.3 Author Based Detection

The ABD approach was a more simplistic idea than the CBD approach. It used one classification stage and performed almost equally well to the CBD approach. The results from the ABD approach are discussed in this section.

6.3.1 Pre-processing

The ABD approach was tested both with and without the pre-processing technique described in Section 3.3.4. The results were generally better with the use of pre-processing, and classification on pre-processed text achieved the highest $F_{0.5}$-score. The ABD approach made a dictionary of all the words and their occurrences used by each author. The author-based dictionary made the ABD approach focus more on the content of each author and neglected words that occurred frequently. After the pre-processing, two of the most weighed features were "want come" and "want bring". Stopwords such as "to" were removed during the pre-processing, which made those two features more likely to occur. There were more bigrams among the features for the ABD approach than for the CBD approach, which was a result of the pre-processing.

6.3.2 Features

The BoW features of the ABD approach were made of the top 20000 used words in the training set, which was 5000 more features than for the CBD approach. The assumption to use more features was that less common words could be more important for an author than for a conversation. There were also more words and messages for each author than for each conversation. The results for the BoW implemented ABD approach were better with the increased feature size. The number of TF-IDF features, on the other hand, were reduced because of the removed stopwords and special characters, and also because there were fewer documents while using the same maximum document frequency.

The highest weighted features of the ABD approach contained more words and phrases that provided meaning by themselves compared to the CBD approach. Examples of such words and phrases were "want come", "want bring", "go jail", "pick u", "ill come", "talk phone", "check see", "ur house", "u want call", "call", "meet" and "truck". The comparison between the most distinctive features for the two approaches indicates that the ABD approach captured more of the conversation content than the CBD approach. This might be the reason why the ABD approach had less false positives. The observation supports the first hypothesis from Chapter 1. The first hypothesis stated that terms used in cyber grooming are categorically different from terms used in general conversation. However, it is the combination of the terms in a predatory conversation that is different from the combination of the terms in a general conversation.

6.3.3 Classification

The highest $F_{0.5}$-score for the ABD approach was achieved with a NN classifier. The NN implemented ABD approach correctly categorized 165 predators and 218445 non-predatory authors and misclassified 89 predators and 3 non-predatory authors. The organizers of the PAN 2012 competition emphasized precision in the predator identification task. Without the emphasized precision, the highest score for the ABD approach would have been an F_1-score of 0.809, which was 0.082 lower than the $F_{0.5}$-score. Compared to the highest F_1-score of the CBD approach, it was 0.071 lower.

The three misclassified non-predatory authors were three somewhat different cases. One author was one of the pseudo-victims that was misclassified in the CBD approach. The second author was participating in a sexual conversation on Omegle and claimed to be a 20 years old male while the other participant claimed to be a 19 years old female. The third author participated in several conversations and produced more than 600 messages. The conversations were not from PJ nor Omegle. The third author used sexually oriented words in a few messages, many emoticons, and was the most active person in the majority of their conversations. The third author also talked often about places where either the author or the author's conversation partners lived. Sexual language and location information about where the authors lived were involved in messages of all the misclassified non-predatory authors. Furthermore, two of the three authors used emoticons frequently. The three authors had similar traits to the predators, which is likely why they were misclassified.

The misclassified predatory authors consisted of 26 predators that were removed by pre-filtering the testing set and 63 predators that were not detected by the NN implemented ABD approach. The pre-filtered authors produced between one and 16 messages each, except for one author that produced 29 messages. The majority of the pre-filtered authors produced less than seven messages and received no response to any of their messages. The author that wrote 29 messages got five messages in response which the author did not answer. The 63 predators that were not detected by the ABD approach produced between 10 and 1354 messages each. Those predators were not manually analyzed.

6.4 Early Detection

The main goal in the PAN 2012 competition was to identify predators from texts. In this thesis, methods have also been implemented to identify predators from texts. However, the goal of the thesis is to investigate the feasibility of early detection of predators from texts. The CBD approach was chosen for early detection because it achieved the highest score and the highest number of predators detected out of the three tested approaches. The CBD approach is also the most realistic approach. It is more realistic than the ABD approach because it did not combine messages from different conversations that predators took part in. It would not be feasible in a real-life scenario to combine messages from different conversations across different communication platforms nor in situations where authors changed their id.

The results presented in Section 5.4 shows that the performance of the early predator detection increases for each message at the beginning of a conversation. This observation applies to almost all 100 of the first messages sent. However, it is still possible to obtain good results with fewer messages. Figure 5.10 shows that the $F_{0.5}$-score was above 0.8 after 24 messages for all of the classifiers. Figure 5.7 shows that the NB implemented classifier's precision was higher than 0.9 for all of the messages. Due to the NB classifier's high precision, it is possible to use the CBD approach in a detection system from the first message sent. These results indicate that the fourth hypothesis from Chapter 1 is false. The fourth hypothesis stated that grooming cannot be detected during the initial phase of an online conversation. The results contradict the hypothesis. Furthermore, the observations presented in Section 6.2.4 showed examples where predators attempted a straightforward approach to groom the victim from the first message in a conversation. In those cases, early detection is necessary during the initial phase of the conversation in order to prevent the grooming.

The emphasized precision, of the PAN 2012 competition, made the classifier with the least amount of detected predators to achieve the highest score because of its lower number of false positives. The organizers thought it was better to let a predator walk than to falsely accuse somebody. However, it can be argued that this is a bad approach as an automated program will often play an advisory role. Meaning the program will select a number of cases that a human investigator needs to analyze further. If the predator detection were to be used in such a scenario, it might have been better to emphasize recall over precision in order to detect as many predators as possible.

The early detection approach was not tested on live chatting, only on old messages. An early detection system would have to be dynamic and efficient in order to keep up with the number of messages sent in an online conversation. The purpose of this thesis was not to implement live predator identification in a system, but rather to investigate its feasibility. The results in this thesis indicate that it is possible to implement a live predator identification system to detect child grooming during online conversations. The results also indicate that predator identification is possible at the beginning of a conversation, depending on the approach of the predator.

6.4.1 Full-length Predatory Conversations

The results from early detection of full-length predatory conversations show that conversations starting from scratch can be detected early in the conversation. Furthermore, there was not a significant difference in what number of messages needed for full-length conversation and shorter conversations segments to correctly classify the victim and the predator. The number of correctly classified victims from predators in full-length conversations are likely to improve by training the classifiers in the CBD approach on full-length conversations. However, training the classifiers on full-length conversations may result in slower detection. The experiments for the detection of predators in full-length conversations did not account for non-predatory conversations. The number of conversations tested was neither sufficient to conclude, but rather indicate the differences between full-length conversations and shorter conversation segments.

Chapter 7 Conclusion

This thesis investigated the feasibility of detecting child grooming during an online conversation. Three different approaches to detect predators in conversations were implemented and tested. The three approaches were Message-Based Detection (MBD), Conversation-Based Detection (CBD) and Author-Based Detection (ABD). The three approaches classified single messages, all messages in a conversation and all messages produced by an author as either predatory or non-predatory. The CBD approach consisted of two classification stages, the first stage differentiated between predatory and non-predatory conversations, while the second stage differentiated between predator and victim. The CBD approach achieved the highest performance and detected the largest amount of predators. Thus, the CBD approach was tested on an increasing number of messages to investigate the possibilities of early detection within a conversation.

The results from the two-stage CBD approach implemented with an SVM classifier for the first classification stage and a NB classifier for the second classification stage were good. The classification system detected 209 out of 254 unique predators and misclassified 20 non-predatory authors in a dataset with 218702 authors. The classification resulted in an $F_{0.5}$-score of 0.893. When used for early detection, the system detected 101 predators within 20 messages, 191 within 50 messages and 207 within 80 messages. The corresponding $F_{0.5}$-scores were 0.754, 0.891 and 0.897. The length of the conversation segments ranged between 1 and 150 messages. The CBD approach was also tested on a limited amount of full-length predatory conversations. The results of full-length predatory conversations and shorter conversation segments were similar, but most of the full-length predatory conversations needed more messages to detect the predator. The results of the CBD approach indicate that it is possible to detect online grooming during an online conversation and that the detection is possible already at the beginning of a conversation.

Intermediate results and manual analysis showed that the combination of terms used in the process of cyber grooming is different from the combination of terms used in general conversations. Predatory terms were used by the different classifiers as features to detect the predators. Not all of the analyzed predators built relations to their victims before they attempted to groom them. Some of the predators even made their intention to be sexually involved with a minor clear from their very first message. Most of the analyzed predators applied the same course of conduct to approach a child. However, the pace of the predators varied. There were significant differences in the precaution taken by the predators. The most careful predators asked the victims several times whether they were law enforcement and whether they were aware of law enforcement that tried to trap predators. The conduct applied by the predators when they were confident that the victim was not law enforcement was similar to the conduct of those predators that had a more direct approach.

Predator detection during online conversations can help to mitigate the societal problem of online grooming. Predator detection is a well-researched area, but it has not been tested during conversations before. This thesis puts light on the importance of early detection in order to detect predators before any physical or psychological harm is caused to the victims.

Future Work

For future work, the use of text classification within a Recurrent Neural Network (RNN) could be exciting for real-time predator identification. Recurrent networks take as their input not just the current input example, but also what they have perceived previously. An RNN can build up a large memory of predatory and non-predatory messages and use them to evaluate new messages. Future work should also investigate predator identification and early detection of predators in conversations with actual underage victims. However, it is a task for law enforcement as the data is susceptible. Further work should also investigate how to integrate a predator detection system with different communication platforms and what functionality is needed for users and platform owners.

References

[AKM15] Ashcroft, M. ; Kaati, L. ; Meyer, M.: A Step Towards Detecting Online Grooming – Identifying Adults Pretending to be Children. In: *2015 European Intelligence and Security Informatics Conference* (2015), S. 98–104

[Bis06] Bishop, Christopher M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg : Springer-Verlag, 2006. – ISBN 0387310738

[ECP18] ECPAT international: Trends in online child sexual abuse material. (2018), April. https://www.ecpat.org/wp-content/uploads/2018/07/ECPAT-International-Report-Trends-in-Online-Child-Sexual-Abuse-Material-2018.pdf

[EEL10] Edwards, April ; Edwards, Lynne ; Leatherman, Amanda: Text Mining and Cybercrime. In: *Text mining: Applications and theory* (2010), 03, S. 149 – 164. http://dx.doi.org/10.1002/9780470689646.ch8. – DOI 10.1002/9780470689646.ch8. ISBN 9780470689646

[EHS11] Egan, Vincent ; Hoskinson, James ; Shewan, David: Perverted Justice: A Content Analysis of the Language Used by Offenders Detected Attempting to Solicit Children for Sex. In: *Antisocial Behavior: Causes, Correlations and Treatments* (2011), Februar, S. 119–134

[ESO16] Ebrahimi, Mohammadreza ; Suen, Ching Y. ; Ormandjieva, Olga: Detecting predatory conversations in social media by deep Convolutional Neural Networks. In: *Digital Investigation* 18 (2016), Sept, S. 33–49

[EURa] EUR-Lex: *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1552295163478&uri=CELEX:32011L0093. – Last accessed 2019-03-11

[Eurb] European Union: *Regulations, Directives and other acts*. https://europa.eu/european-union/eu-law/legal-acts_en. – Last accessed 2019-03-11

[FS96] Freund, Yoav ; Schapire, R. E.: Experiments with a new boosting algorithm. In: *In Proceedings of the thirteenth International Conference on Machine Learning* (1996), S. 148–156

[Gar] Garbade, Michael J.: *A Simple Introduction to NLP*. https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32. – Last accessed 2019-05-31

[GB17] Gerven, Marcel van ; Bohte, Sander: Editorial: Artificial Neural Networks as Models of Neural Information Processing. In: *Frontiers in Computational Neuroscience* 11 (2017), 12. http://dx.doi.org/10.3389/fncom.2017.00114. – DOI 10.3389/fncom.2017.00114

[GKS12] Gupta, Aditi ; Kumaraguru, Ponnurangam ; Sureka, Ashish: Characterizing Pedophile Conversations on the Internet using Online Grooming. In: *CoRR* abs/1208.4324 (2012). http://arxiv.org/abs/1208.4324

[IC12] Inches, Giacomo ; Crestani, Fabio: Overview of the International Sexual Predator Identification Competition at PAN-2012. (2012)

[JWHT14] James, Gareth ; Witten, Daniela ; Hastie, Trevor ; Tibshirani, Robert: *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. – ISBN 1461471370, 9781461471370

[Kit04] Kitchenham, Barbara: Procedures for Performing Systematic Reviews. In: *Keele University Technical Report TR/SE-0401* (2004). http://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf

[Leg] Legal Information Institute: *18 U.S. Code Chapter 117 - TRANSPORTATION FOR ILLEGAL SEXUAL ACTIVITY AND RELATED CRIMES*. https://www.law.cornell.edu/uscode/text/18/part-I/chapter-117. – Last accessed 2019-02-26

[MBK$^+$11] McGhee, India ; Bayzick, Jennifer ; Kontostathis, April ; Edwards, Lynne ; McBride, Alexandra ; Jakubowski, Emma: Perverted Justice: Learning to Identify Internet Sexual Predation. In: *International Journal of Electronic Commerce* 15 (2011), Nr. 3, S. 103–122. – ISSN 1557–9301

[Mey15] Meyer, Maxime: *Machine learning to detect online grooming*, Uppsala university, Master thesis, July 2015. http://uu.diva-portal.org/smash/get/diva2:846981/FULLTEXT01.pdf

[Min] Ministry of Justice and Public Security: *The Penal Code: Chapter 26. Sexual offences*. https://lovdata.no/dokument/NLE/lov/2005-05-20-28/KAPITTEL_2-11#KAPITTEL_2-11. – Last accessed 2019-02-26

[MS99] Manning, Christopher D. ; Schütze, Hinrich: *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA : MIT Press, 1999. – ISBN 0–262–13360–1

[MVA$^+$18] Madigan, Sheri ; Villani, Vanessa ; Azzopardi, Corry ; Laut, Danae ; Smith, Tanya ; Temple, Jeff R. ; Browne, Dillon ; Dimitropoulos, Gina: The Prevalence of Unwanted Online Sexual Exposure and Solicitation Among Youth: A Meta-Analysis. In: *Journal of Adolescent Health* 63 (2018), Nr. 2, 133 -

141. http://dx.doi.org/https://doi.org/10.1016/j.jadohealth.2018.03.012. – DOI https://doi.org/10.1016/j.jadohealth.2018.03.012. – ISSN 1054–139X

[Nat] National Center for Missing and Exploited Children: *Online Enticement.* http://www.missingkids.com/theissues/onlineenticement. – Last accessed 2019-02-27

[Nav] Navlani, Avinash: *Understanding Logistic Regression in Python.* https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python. – Last accessed 2019-04-26

[NMEL18] Ngejane, C.H ; Mabuza, Gugulethu ; Eloff, J.H.P ; Lefophane, Samuel: Mitigating Online Sexual Grooming Cybercrime on Social Media Using Machine Learning: A Desktop Survey. (2018), 08, S. 1–6. http://dx.doi.org/10.1109/ICABCD.2018.8465413. – DOI 10.1109/ICABCD.2018.8465413

[Oxf] Oxford English Dictionary: *Overfitting.* https://en.oxforddictionaries.com/definition/overfitting. – Last accessed 2019-04-26

[PAN] PAN: *Author Identification.* https://pan.webis.de/clef12/pan12-web/author-identification.html. – Last accessed 2019-02-21

[Pat] Patel, Savan: *Chapter 2: SVM.* https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effZ72. – Last accessed 2019-04-29

[Pee12] Peersman, C: Conversation Level Constraints on Pedophile Detection in Chat Rooms. In: *CLEF 2012 Conference and Labs of the Evaluation Forum* (2012), 01, S. 1–13

[Pen] Pennebaker Conglomerates: *LIWC.* http://liwc.wpengine.com/. – Last accessed 2018-10-03

[Pen07] Pendar, Nick: Toward Spotting the Pedophile Telling victim from predator in text chats. (2007). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4338354

[Per] Perverted Justice Foundation: *Perverted Justice.* http://www.perverted-justice.com/. – Last accessed 2018-10-05

[PKM12] Pandey, Suraj J. ; Klapaftis, Ioannis ; Manandhar, Suresh: Detecting Predatory Behaviour from Online Textual Chats. In: *Multimedia Communications, Services and Security* (2012), S. 270–281. ISBN 978–3–642–30721–8

[QSR] QSR: *What is NVivo?* https://www.qsrinternational.com/nvivo/what-is-nvivo. – Last accessed 2018-09-26

[Sas07] Sasaki, Yutaka: The truth of the F-measure. In: *Teach Tutor Mater* (2007), January

[Scia] Scikit-learn: *1.1. Generalized Linear Models.* https://scikit-learn.org/stable/modules/linear_model.html. – Last accessed 2019-04-26

[Scib] Scikit-learn: *1.17. Neural network models (supervised).* https://scikit-learn.org/stable/modules/neural_networks_supervised.html. – Last accessed 2019-04-29

[Scic] Scikit-learn: *1.9. Naive Bayes.* https://scikit-learn.org/stable/modules/naive_bayes.html. – Last accessed 2019-04-29

[Scid] Scikit-learn: *3.1. Cross-validation: evaluating estimator performance.* https://scikit-learn.org/stable/modules/cross_validation.html. – Last accessed 2019-04-29

[VTJGE$^+$12] Villatoro-Tello, Esaú ; Juárez-González, Antonio ; Escalante, Hugo J. ; Gómez, Manuel M. ; Pineda, Luis V.: A Two-step Approach for Effective Detection of Misbehaving Users in Chats. (2012)

[Wie14] Wieringa, Roelf J.: *Design science methodology for information systems and software engineering.* Springer, 2014. http://dx.doi.org/10.1007/978-3-662-43839-8. http://dx.doi.org/10.1007/978-3-662-43839-8. – ISBN 978–3–662–43838–1. – 10.1007/978-3-662-43839-8

[Wol11] Wollis, Melissa: *A Linguistic Analysis of Online Predator Grooming*, Cornell University, Honors thesis, Jan 2011. https://core.ac.uk/download/pdf/4918216.pdf