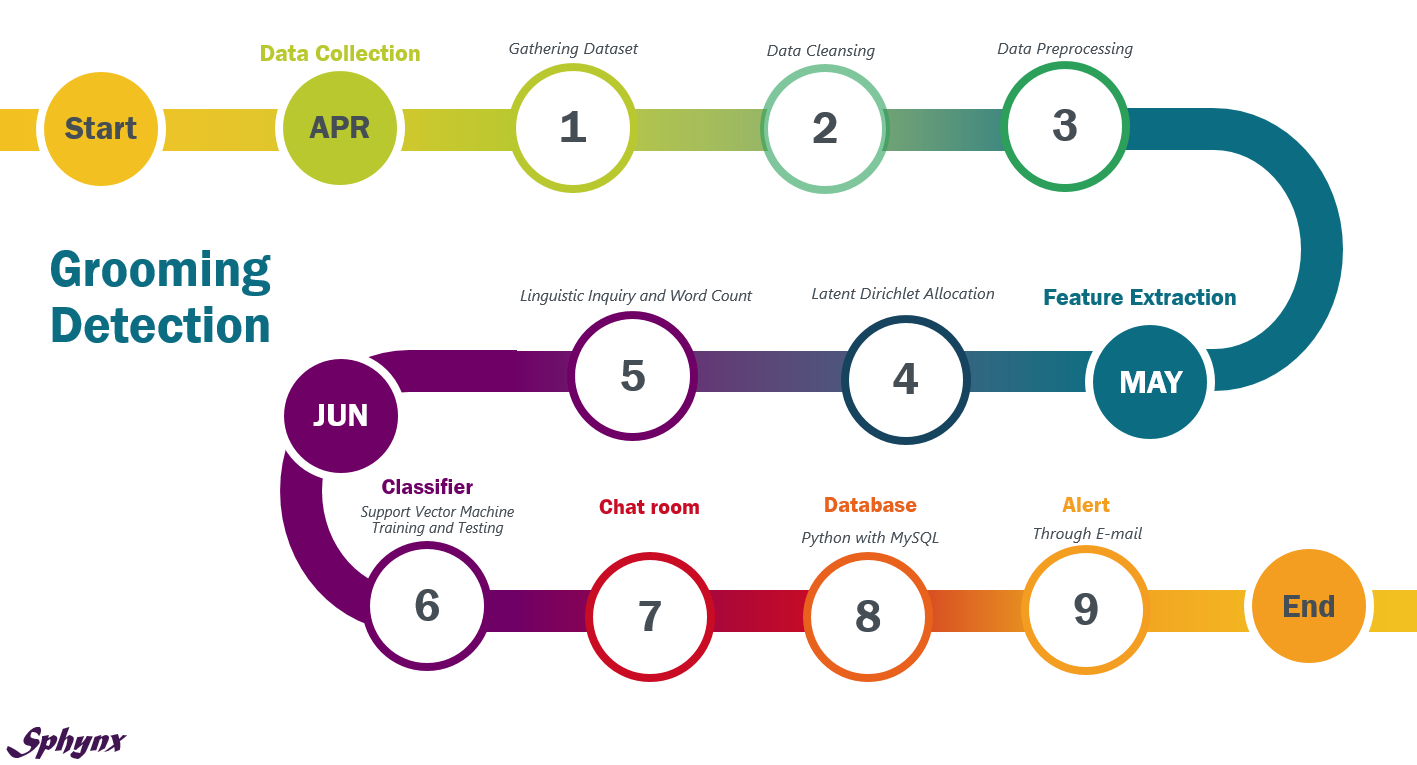
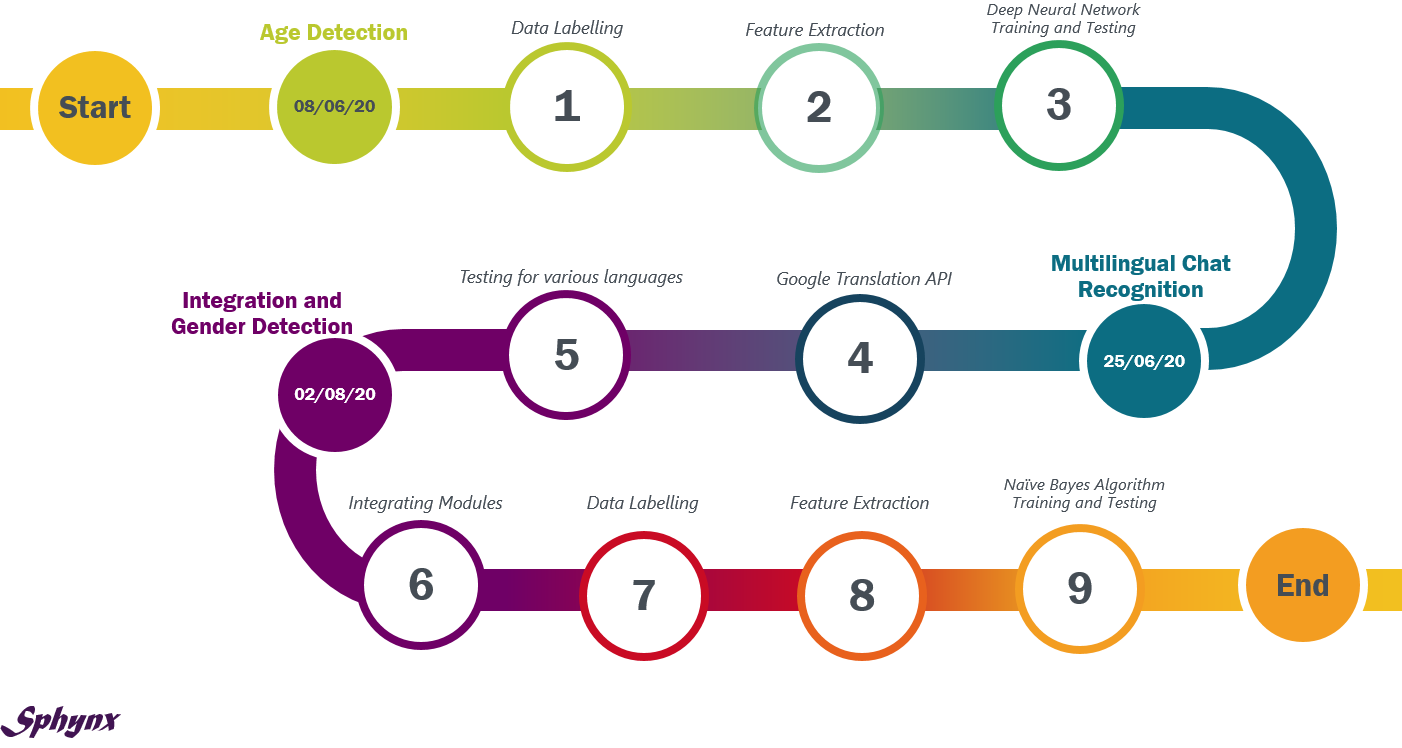
PREVENTION OF EMOTIONAL ENTRAPMENT OF GIRLS ON SOCIAL MEDIA PLATFORMS

**Roadmap**



****

**Literature Survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research Paper** | **Methodology** | **Conclusions/Findings** | **Limitations** | **Publication Year** |
| Topic Detection and Extraction in Chat,  Paige H. Adams and Craig H. Martell | TF- IDF based vector space model approach with temporal relationship information. | Time- distance penalisation shows promise. | Time- distance penalisation has the effect of pulling apart strongly linked messages. Collocation detection is missing. This method becomes expensive as the number of posts increases. | 2008 |
| Our Little Secret: Pinpointing Potential Predators, Anna Vartapetiance and Lee Gillam | Compared Naïve Bayes, J48 decision tree and SVM supervised machine learning approaches to detect paedophilia. | The Naïve Bayes algorithm performed better with an F1 score of 0.76. | Variation in age leads to inconsistency. | 2014 |
| Detecting Child Grooming Behaviour Patterns  on Social Media,  Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani | Feature selection done using information gain. Cosine similarity with LIWC was used by the SVM classifier. | Combined features in classifiers boosts the performance. High recall values ensure good coverage of stage detected in the sentences. LIWC2007 covers over 60 dimensions of language. | Precision value should be high to minimize the number of false negatives. Classification performance drops while using sentiment polarity and content features independently. | 2014 |
| Detecting sexual predators in chats using  behavioral features and imbalanced learning, Claudia Cardei and Traian Rebeda | Used a two stage classifier- Suspicious Conversation Identifier (SCI) using SVM and identifying predators using Random Forest classifier. | It handles dataset imbalance. Higher precision achieved due to behavioural features. The presence of SCI reduces the second classifier dataset. | The exact messages showing misbehaviour are not identified. Chats must be in English and must not contain emoticons. | 2017 |
| Grooming Detection using Fuzzy-Rough Feature  Selection and Text Classification,  Zheming Zuo, Jie Li, Philip Anderson, Longzhi Yang  Nitin Naik | This approach applies the conventional text feature extraction along with fuzzy rough feature selection. GNB, Random Forest, Adaboost and Logistic Regression classifiers were used. | The logistic regression classifier combined with the fuzzy rough feature extraction gave highest accuracy of 73.10%. | It requires further investigation to test the proposed approach while dealing with online text streaming in real time. | 2018 |
| An Intelligent Online Grooming Detection System  Using AI Technologies,  Philip Anderson, Zheming Zuo, Longzhi Yang, Yanpeng Qu | This approach used BoW model and fuzzy rough feature selection. Fuzzy twin SVM was used to classify conversations. | Better performance due to fuzzy rough feature selection and combination of BoW model and MM. | This method doesn’t use advanced AI techniques for parameter settings. It also does not work for live streaming chats. | 2019 |
| Technical Mapping of the Grooming Anatomy  Using Machine Learning Paradigms: An Information Security Approach  Patricio Zambrano, Jenny Torres, Luis Tello-Oquendo, Rubén Jácome,  Marco E. Benalcázar, Roberto Andrade, And Walter Fuertes | Latent Dirichlet Allocation (LDA) topic modeling was applied to determine the stages of the attack. Once the number of stages was determined, linguistic context was given and a linear model was trained. | Identifying these stages supports investigations related to identifying patterns of malicious behavior online | The classifier does not indicate whether a conversation is grooming or not. | 2019 |

**Dataset Gathering**

A collection of chats, both grooming, non- grooming as well as conversations with abusive content were gathered to train the classifier. Grooming conversations were collected from the perverted justice website [1]. Non- grooming conversations were collected from IRC logs [2]. Finally, the false- positive conversations were gathered from fugly website [3].

**Data Cleansing**

The chats collected as a result of the above mentioned process were cleansed by removing timestamps, commentary, converting group conversations to those between two people. This was done to aid the conversion of the said conversation to csv format.

**Data Labelling**

The conversations collected were labelled as follows for each of the three classifiers-

* Grooming or non-grooming
* Teenager or adult
* Female or male

**Data Pre-processing**

This step included removal of emoticons, conversion to lower case (transformation), tokenisation, removal of stop-words and punctuation, lemmatisation. This was done in order to facilitate the feature extraction process. The conversation was pre-processed and stored line by line.

**GROOMING DETECTION**

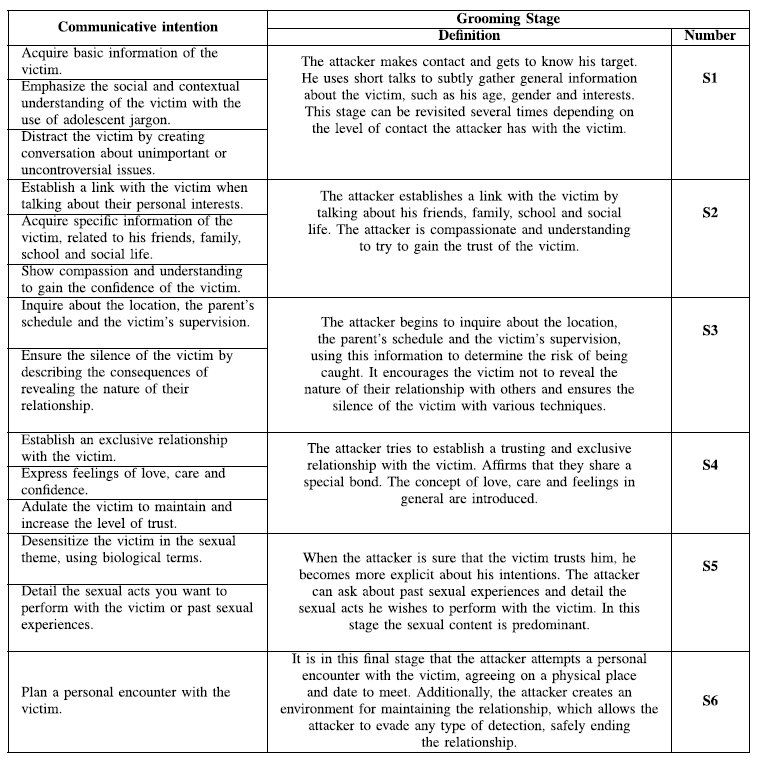
**FEATURE EXTRACTION**

**Latent Dirichlet Allocation (LDA)**

The grooming process was scattered across six stages. Subsequently, the words in the conversation were grouped into six categories with the help of LDA. However, the basis or sentiment on which these words were grouped could not be identified using LDA. Hence, feature extraction was performed with the help of LIWC.

**Linguistic Inquiry and Word Count (LIWC)**

A dictionary consisting of the words from the LIWC 2007 dictionary, common slang words used on social media platforms and sexual and abusive words were collected and split across the six grooming stages. The words in each line of the conversation were assigned a stage and the highest stage was recorded for the corresponding lines. The count of these six stages (S1- S6) are given as input to the classifier.



**Classifier**

A Support Vector Machine (SVM) was used to classify the conversation as grooming or non- grooming based on the features determined by the LIWC process. Sixty percent of the conversations were used for training, twenty percent for validation (in order to fix the kernel type) and twenty percent for testing.

**AGE DETECTION**

**Feature Extraction**

Nine features were chosen to indicate the age of users. In order to accurately identify the age and at the same time detect all grooming activities, the users will be monitored for two weeks and hence chats of varying lengths were taken for training and testing. The features used were average number of words per sentence, average number of characters per word, number of punctuations used, occurrence of duplicate characters in words, usage of slang words, abbreviations and acronyms, number of emoticons, number of posts, number of followers and number of following.

**Classifier**

A deep neural network was used to classify the age of the users as teenager or adult based on the six features extracted from the chat. Sixty percent of the dataset was used for training, twenty percent for validation and the remaining twenty percent for testing.

**Multilingual Chat Recognition**

As an additional feature, we used the Google translator API to extend the scope of the project to various regional as well as foreign languages and not just English. This API supports a total of 109 languages.

**Chat room**

To understand and demonstrate the application of this software in real time, which is to be continuous monitoring, a localised chat room server and client environment was created. As each line was typed, it was pre-processed and the features were correspondingly extracted and updated in a database. After every pair of exchange, the updated features were passed to the classifier and the result was updated in the database.

**Database**

Details regarding the users, chats, as well as users to be monitored were stored in tables in a chat room database. The users flagged by the age classification module were appended in the monitoring table. The chats of these users were continuously monitored and the corresponding features as well as output of the classifier were continuously updated. Once a user exceeded the monitoring age limit (not a teenager), her record was removed from the monitoring table.

**Alert**

Once a chat was identified to contain grooming characteristics, an e-mail was sent to the moderator containing both user IDs and with the subject line ‘Possible Grooming Alert.’

**Integrating Modules**

Both the grooming as well as the age modules were connected to the MySQL database so as to show the real time working or application of this project on a small scale. After monitoring the user for two weeks, the age is determined and based on this result their corresponding chats are continuously monitored for grooming activities.

**Test Case**

A number of different inputs were taken and the expected outputs were also noted for each component as well as the entire integrated module in order to test and detect any anomalies if present.

**Test Scenario**

The inputs mentioned in the test case were executed to check for anomalies. In the cases were such anomalies did occur, the code was corrected. The cases covered a wide range of inputs in order to increase accuracy and decrease anomalies.

**GENDER DETECTION**

**Feature Extraction**

Six features were used to detect the gender of the user from his or her chat. Similar to the age component, the user was monitored for the same two weeks after which the chat was analysed to detect the gender. The features chosen were average message length, average number of characters per word, number of punctuations, occurrence of duplicate characters in words, frequency of stop words and average number of distinct words.

**Classifier**

A naïve Bayes classifier was used to detect the gender of the user based on the features extracted from his or her chat. Sixty percent of the dataset are used for training, twenty percent for validation and the remaining twenty percent for testing.

**Technology Stack**

This software is coded using Python 3.

* NLTK (Pre- processing)
* GENSIM (LDA)
* LIWC
* SCI-KIT LEARN ( SVM and Naïve Bayes Classifier)
* TENSORFLOW WITH KERAS (DNN Classifier)
* GOOGLETRANS API (Multilingual)
* SMTPLIB (Mail alert)
* SOCKET (Chat room)
* MINIDOM (Xml parser)
* SVM algorithm
* Deep Neural Network
* Naïve Bayes Algorithm
* MySQL Database

**References**

[1]- <http://www.perverted-justice.com/>

[2]- <https://krijnhoetmer.nl/irc-logs/>

[3]- <https://www.fugly.com/victims/>

**Referred paper:**

**Grooming Detection**

Patricio Zambrano, Jenny Torres, etc. “Technical Mapping of the Grooming Anatomy Using Machine Learning Paradigms: An Information Security Approach”, in Proc. IEEE Int. Conf. Artificial Intelligence in Cyber Security, October 2019.

**Age Detection**

Rita G. Guimaraes, Renata L. Rosa, etc. “Age Groups Classification in Social Network using Deep Learning”, in IEEE Access, Access-2017-02532.

**Gender Detection**

Tayfun Kucukyilmaz, B. Barla Cambazoglu, etc., “Chat Mining for Gender Prediction”, Bilkent University, Department of Computer Engineering, Turkey.