Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 189 (2021) 116–127

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

Sexual-predator Detection System based on Social Behavior Biometric (SSB) Features

Mudasir Ahmad Wani*, Nancy Agarwal, Patrick Bours

Department of Information Security and Communication Technology (IIK)
Norwegian University of Science and Technology (NTNU)
Teknologivegen 22, 2815 Gjøvik, Norway

Abstract

This study designs an online sexual predator detection system using Social Behavior Biometric (SSB) features. Social biometric focuses on extracting the pattern a user exhibits while interacting and communicating through social networks. The paper addresses the online sexual predator problem by mining the vocabulary and emotional behavior, which could assist in identifying if the user is a benign or predator. The feature-set consists of vocabulary terms that appear differently in predator and victim content. In order to strengthen the detection model, the paper also focuses on distinguishing the two classes of users based on emotions reflected in their conversation. The experiments are performed on the PAN 2012 corpus. Two datasets are created with respect to vocabulary-based and emotion-based features. The results obtained on the test set have proved that by integrating the vocabulary and emotion-based attributes, the performance of the system is significantly enhanced. While comparing, the proposed approach has outperformed top existing methods by obtaining F_1, F_2, and $F_{0.5}$values of 0.95, 0.94, and 0.96 respectively. Furthermore, we also recorded the best accuracy compared to state-of-the-art studies for our proposed SBB-based approach with 99.86%, 99.51%, and 99.88% for Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) respectively.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on AI in Computational Linguistics.

Keywords: Online Sexual Predators; Emotion mining; Lexical analysis; Machine Learning;

1. Introduction

Social networking applications play an essential role in our daily lives by providing a platform to connect, communicate and socialize with other people easily. We may or may not know our online contacts also in the real world. According to an online abuse report (2019) [2], around 90% of the population in the age group 11–16 years possess a social media account. Alongside amazing opportunities, these sites also open doors for various safety risks to their users. For example, the anonymity characteristic of social networks allows a person to behave in whatever manner

[*] Corresponding author. Tel.: +47 46593757 ; fax: +0-000-000-0000.
E-mail address: mudasir.a.wani@ntnu.no

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on AI in Computational Linguistics.
10.1016/j.procs.2021.05.075

wanted. This feature can put the users, especially children, and youngsters, in extreme danger since sexual predators [6] can easily deceive and lure them by adopting fake identities. It has also been reported that the number of cases of online sexual offenses are continuously growing with time.

The online child sexual abuse problem is being studied from different dimensions. Law enforcement is working towards the prevention of child and sexual abuse, while mental health experts and psychologists are studying and investigating the phenomenon behind the abnormal behavior. The present time demands a system that automatically detect the behavior of a sexual predator in these online networks and alerts the relevant authorities. The system will be helpful for a parent, a local authority such as police, or the minor who is involved in an online chat conversation.

Our work primarily attempts to capture the Social Behavioral Biometrics (SBB) [18] traits of the users that could distinguish online sexual offenders from victims and assist in keeping the young generation safe on these platforms. SBB is a new trend that, unlikely biological biometrics, focuses on analyzing the social interaction and activities of the users. Researchers have primarily studied the vocabulary and the way a predator interacts with the target in the conversations to combat the predator issue. However, emotion analysis has gained little attention in this direction. Since predators are considered emotionally unstable by psychologists, emotion mining from chat conversations as new SBB features would add significant benefit to the detection approach.

In this paper, we extract both vocabulary and emotion-based features to design a sexual predator detection model. The experiments are conducted on the PAN 2012 dataset [13], which is the largest dataset used for detecting sexual predators in online conversations so far. The more detailed information about the dataset is provided in the data collection and pre-processing description in Section 3. In the vocabulary set, instead of using all the words in the conversations, we filter those words which are differently used by predators and victims. Bag of Words (BoW) [23] approach has been employed to design the language model of both classes of users. The emotion behavioral-based features have been extracted using the MoodBook lexicon [20]. The lexicon provides lists of emotion terms, divided into eight categories of emotions. These categories are *fear*, *anger*, *sad*, *joy*, *surprise*, *disgust*, *trust*, and *anticipation* and represent Robert Plutchik's emotion wheel [1]. The authors have employed this emotion lexicon for several studies, such as fake profile detection [21], gender prediction [22], etc. One of the main goals of this study is to identify the potential of emotion-based features over the dictionary ones in identifying predators. The results show that only emotion-based features are not enough to design a sexual predator detection system. Therefore, an efficient sexual offender detection has been designed by combining the proposed vocabulary and emotion-based features.

The main contributions of this study are:

- Identification of a set of vocabulary terms used by predators and victims differently;
- Mining of emotional behavior of users to observe their mental state to aid in the detection of sexual predators;
- Designing an integrated feature-based sexual predator detection system;
- The two datasets will be made available for the researchers of this domain.

The rest of the paper is as follows: Section 2 provides the literature around sexual predator detection systems and the features employed by other researchers to train their systems. In Section 3, we have discussed the datasets on which we conducted our experiments, along with some pre-processing requirements. Section 4 presents an exploratory data analysis to summarize the main characteristics of victims and predators. In Section 5, we have discussed emotion and vocabulary-based features used in this study. Section 6 presents the experiments and results of this study and finally, section 7 concludes the overall work of designing a sexual predator detection system.

2. Background Study

Many studies have been conducted with the aim to understand how sex offenders are leveraging cyberspace to commit different crimes involving children, and what are their characteristics and demographics. For example, the study [12] observed that online offenders exhibit different behavior from offline sexual offenders. In [17], the authors conducted a detailed analysis to study the background and behavior of online sexual offenders. They sampled the data from the online child sexual exploitation-related cases that were under the Innocent Images National Initiative (IINI) investigation. They observed that offenders were showing diverse characteristics in terms of age, education, occupation, and family dynamics except for gender, which was greatly male. Also, the majority of the offenders did not show any criminal history.

There are various studies [5, 19], which were keen to develop a tool for detecting the presence of an offender in a chat, by processing and analyzing a conversation. The authors in [15] performed experiments on the PAN 2012 dataset to address the issue of detecting online sexual offenders in chatting applications. The authors employed an n-gram model and used the LIWC (Linguistic Inquiry and Word Count) [7] software to draw the distinguishing features for training an SVM classifier. On the same corpus, another study [19] used a two-stage classifier to identify the sexual predators in online chatting. The first layer classification was designed for filtering the suspicious conversations containing potentially sexual offensive content from the normal chatting, while the second layer was used for the task of identifying the actual predator in the suspicious conversations. Their work was mainly grounded on two theories; first, words used in the chat related to child exploitation are significantly different than in a normal chat, and second, predators usually adopt the same behavior to approach a target victim. They developed a language model using BoW technique to train the classifiers. In [11] a similar approach was used where the authors tested various combinations of features and classifiers, including fusion of various classifiers. The authors in [5] also applied the two-step classification scheme for the predator identification task. However, they also captured the behavioral patterns of the sexual predators for the solution. Furthermore, the authors of [16] devised a three-stage identifier by combining post-level and user-level classifiers in order to enhance the performance of the detection system. Bours and Kulsrud [4] used a two-step approach in an attempt to detect the sexual predators as early as possible during a chat.

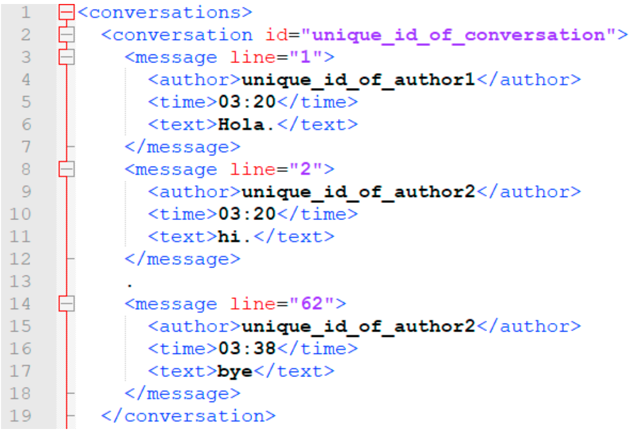
Earlier studies have mainly focused on the vocabulary or behavior of predators to address the problem. Since sexual predators are considered emotionally unstable by psychologists, there is a study [3] that investigates the potential of sentiment-based features in the predator detection process. The following six emotions categories were considered: *fear*, *joy*, *anger*, *sad*, *disgust* and *surprise*. The feature-set is also integrated with other content-based attributes such as the use of personal and reflexive pronouns, imperative sentences, relationships words, etc. In our study, we investigated eight dimensions of emotions along with positive and negative sentiments, and the number of emotion categories revealed by a user during chatting. Furthermore, we also constructed a BoW model based on the vocabulary used by predators and victims separately.

3. Data Collection and Pre-processing

The sexual predator identification problem can be clearly seen as a supervised machine learning task. In order to train the prediction models, we need labeled data. In this study, we used the data provided in PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) 2012 competition for the Sexual Predator Identification task [13]. This data contains almost 67k conversations for training and over 155k conversations for testing, extracted from several chat repositories and Internet Relay Chat (IRC) channels. Furthermore, a list with the ids of sexual predators is also provided. The dataset is in XML format containing conversation_id, author_id, message time, and message text in the format as shown in Figure 1.

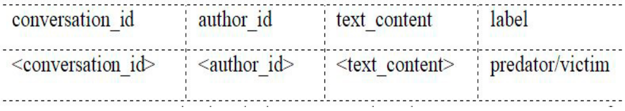
In order to make the dataset ready to train the different machine learning algorithms, various pre-processing functions were applied on the datasets. First, those conversations in the dataset which contains messages from one user only or more than two users are treated as noisy as they are not providing knowledge of the distinguishing behaviour between normal and predatory chatting. For experiments the conversations in which exactly two users are involved are taken into consideration. Furthermore, those conversations that had less than 6 messages were also removed. Afterwards, the text messages of each user in a conversation are accumulated in a file with their respective conversation_id, author_id, text_content and label, as shown in Figure 2.

Once we obtained the accumulated message content of every user from several conversations, we applied basic pre-processing techniques to the whole corpus, including tokenization, same casing, stop word removal, removal of special characters, etc. We aggregated the content of every predator and every victim from the conversations to see the topmost vocabulary terms used by both user groups. Furthermore, we also found the vocabulary terms which have been predominantly used by predators but not by victims and vice-versa. Also, We discarded the instances from the processed dataset, where we could not find enough vocabulary terms to profile a user. The statistics of the resulting training dataset are given in Tables 1 and 2 at conversation and user level respectively.



```
<conversations>
  <conversation id="unique_id_of_conversation">
    <message line="1">
      <author>unique_id_of_author1</author>
      <time>03:20</time>
      <text>Hola.</text>
    </message>
    <message line="2">
      <author>unique_id_of_author2</author>
      <time>03:20</time>
      <text>hi.</text>
    </message>
    .
    <message line="62">
      <author>unique_id_of_author2</author>
      <time>03:38</time>
      <text>bye</text>
    </message>
  </conversation>
```

Fig. 1. Raw structure of experimental Data



conversation_id	author_id	text_content	label
<conversation_id>	<author_id>	<text_content>	predator/victim

Fig. 2. Dataset initial schema

Table 1. Dataset (Conversation level).

Conversations	Size
Predatory	972
Normal	8783
Total	9755

Table 2. Dataset (user level).

User	Size
Predators	137
Victims/non-predators	15683
Total	15820

This data has been used to train and validate the model. The final testing has been done separately on the testing dataset provided by the PAN 2012 for the sexual predator detection task.

4. Predator-Victim Exploratory Data Analysis

In this section, we analyzed our datasets to summarize their main characteristics. After aggregating the message content of all the predators and victims from different predatory conversations, we were keen to observe the topmost words used by the users. We found words including "*like*", "*want*", "*want*", "*know*", "*call*", "*home*", etc. are being used by both user groups while chatting with each other. We noticed that these words belong to different word categories including approach words (e.g. "*meet*", "*together*", "*car*" "*room*", "*hotel*"), family words (e.g. "*mom*", "*dad*", "*sister*", "*brother*"), and relationship words (e.g. "*boyfriend*", "*partner*", "*date*"), as suggested in [16]. Furthermore, we also observed other words such as "*pretty*", "*beautiful*", "*cute*", "*sweetie*", "*princess*", "*like*", were used most of the time by both the user groups. We call these words positive terms as they are usually used to praise someone. Similarly, we found words such as "*lips*", "*legs*", "*mouth*", "*eyes*", "*tongue*", "*hair*" and categorize them as body-parts words. Table 3 shows examples of some of the high-frequency words used by both victims and predators while chatting with each other. The table also lists each word with the example sentence from the dataset.

Table 3. Terms Commonly used by Victims and Predators.

Category	***Word(s)***	***Example (from dataset) P: predator, V: victim***
Approach words	hotel, car, etc.	..so do we have to get a hotel or can we stay at your place? (P) did u get a car yet? (V)
Connection words	talk	i got online to talk (P) yes! talk to you later night:-* (V)
Family words	mom, family, etc.	u having other family over or just u and your mom? (P) i wont bug u while ur with ur family! (V)
Body-parts Words	lips, eyes, etc.	its nice to feel lips on mine. . . (P) u really do have green eyes they r really pretty (V)

In order to see the distinguishing characteristics of predators and victims, we found some of the words which were being used by predators but not victims, out of all the topmost common words. Similarly, we recorded words used by victims but not by predators. For example, words like "*sweetie*", "*feeling*", "*please*", "*body*", "*touch*", "*18*" have been found mostly in predator messages. While as topmost words mostly found in the victim message but not in predator's content include words such as "*kno*" (from know), "*2nite*" (from tonight), "*broke*", "*hurts*", "*idk*" (meaning: I don't know), "*14*". Table 4 shows some of the words used by either of the groups, with some sentence examples from the collected dataset.

Apart from the words shown in Table 4, there are plenty of other terms that can distinguish predators from victims (or normal) users. Please note that the words described in Table 4 are based on the statistics of topmost words from the two groups. Therefore, it is quite likely that a word occurring in both predator and victim content fails to secure a position in the topmost list of one of the groups. For example, the word "*horny*" appears in predator and victim messages 82 and 2 times respectively, so it is assigned to the topmost list of predator class only. The other examples include words such as "*friends*" (p:82, v:2) or "*swear*" (p:10, v:25).

Table 4. Terms Commonly used by Victims or Predators.

Category	***Word(s)***	***Example (from dataset) P: predator, V: victim***
Approach words	apartment	Come to my apartment. Here i am alone. Will u come? (P)
Connection words	chatting	... a little bore but better now that im chatting with you (V)
Feeling words	awww	awww thats rly nice, awww im sorry, awww ur so sweet, etc. (V)
Sexual words	horny	... I am getting horny thinking about it. (P)

5. Feature Engineering

Once the data is pre-processed, the next step in machine learning is to derive the optimal feature vector from the content which provides the most discriminative information and has the best potential to distinguish between the two classes. In this work, we extract two categories of the features from the chat messages, namely, vocabulary-based and emotion-based features. In the above section, we observed that predators and victims use somewhat different vocabulary while chatting. Based on this observation, we evaluate two sets of words as follows.

$$Set(P - V) = \{\text{Top } n \text{ Pred words}\} - \{\text{Top } n \text{ Vic words}\}$$
$$Set(V - P) = \{\text{Top } n \text{ Vic words}\} - \{\text{Top } n \text{ Pred words}\}$$

After conducting the exploratory data analysis we first created the set of n most used words by predators (P) and similarly by victims (V), where we used $n = 10000$ as we observed most of the words were covered when $n = 10000$. Therefore, we did not realised the need of increasing the value of n. Afterwards, we created the two sets, $Set(P - V)$ and $Set(V - P)$ by taking the difference of the P and V sets. Also, in the $Set(P - V)$ and $Set(V - P)$, we removed the words with a frequency of 25 or less. We also used words with occurrence less 25 (for example ¡23,¡ 18,etc.) but did not obtained promising results. Therefore we stick to the words with frequency less or equal to 25. Finally, the set $Set(P - V)$ contains all the words of the set P that are not in the set V and based on our dataset, we noticed that this set contained 299 different words. Similarly, the set $Set(V - P)$ contains 304 words. In the end, we combined both sets into one set and named it "*PreVicVocab*", containing 603 words. Using the words appearing in PreVicVocab, the vocabulary-based features of the users are determined based on the BoW language model. The number of features is equal to the size of PreVicVocab, i.e., 603, where each feature is linked to a word holding the count of the number of times the word appears in the user content. The vocabulary feature values of i^{th} user are calculated as given in Equation (1) where $PreVicVocab_j$ returns the word to be counted in the aggregated content of the i^{th} user.

$$V_{i_j} = frequency(PreVicVocab_j), \forall j = 0..603 \quad (1)$$

The next set of feature categories comprises the evaluation of the emotions of the users. The MoodBook lexicon [20] has been utilized to obtain these attributes. This lexicon provides a list of emotion-terms for the eight classes of emotions, namely, fear, anger, sad, joy, surprise, disgust, trust, and anticipation. For example, the joy category includes words like "*happy*" and "*awesome*" and the sad category includes terms like "*cry*" and "*hopeless*". Furthermore, the lexicon also lists the emotion-terms for positive and negative sentiments. Based on the MoodBook lexicon, a total of 11 emotion-based attributes are constructed. The first ten attributes of the i^{th} user $(E_{i_1}, E_{i_2}, \ldots, E_{i_{10}})$ correspond to eight emotions and two sentiments (positive and negative) classes of Moodbook, respectively. Let $moodbook_{j,k}$ represents the k^{th} emotion term in the j^{th} class. Then, values from E_{i_1} to $E_{i_{10}}$ can be determined by Equation (2), where $moodbook_{j,k}$ returns the emotion word to be counted in the i^{th} user text.

$$E_{i_j} = \Sigma_k frequency(moodbook_{j,k}), \forall_{j=1..10} \quad (2)$$

The 11^{th} emotion attribute (E_{11}) captures the number of emotion categories found in the user content. Let $tokenize(text_i)$ return the list of word tokens of the messages of i^{th} user. Equation (3) provides the formula for deriving the E_{11} value. The expression $n([tokenize(text_i) \cap ([moodbook_j])]$ calculates the number of emotion terms appears in the user content of the j^{th} emotion class of Moodbook.

$$E_{i_{11}} = \Sigma_j \max(0, \min(n(tokenize(text_i) \cap moodbook_j, 1))), \forall_{j=1..8} \quad (3)$$

We integrate both vocabulary and emotion-based feature sets for our experiments. Table 5 provides a clear overview of all the features with their possible values used in our experiments.

Table 5. Features used in the Study

Feature(s)	***Feature type***	***Range***
$f_1 - f_{603}$	Vocabulary-based (cBoW)	0 to # specific vocabulary words in the user content
$f_{604} - f_{611}$	Specific Emotion-based (MoodBook)	0 to # emotion words in user content
$f_{612} - f_{613}$	Positive/negative Emotion-based (MoodBook)	0 to # positive or negative emotion words in user content
f_{614}	# Emotion categories (MoodBook)	0-8
f_{615}	Class Label	Predator/Victim

In this section, we also plotted some of the features to clearly visualize their distinguishing characteristics. Figure 3 shows the box plots of 8 emotion-based attributes for two classes of users, which depicts the difference in the predators and victims. For example, in the case of fear emotion category, the predators' range (0-4) is slightly higher than the victims' range (0-2). The same trend is observed for the trust category with (0-9) and (0-7) as predators and victims' values, respectively. However, the difference in range values for sad and joy emotion gets bigger. The predators' score is (0-30) and (0-100) for sad and joy, respectively, whereas, for the same emotion categories, the victims' score (0-25) and (0-80), respectively. Furthermore, the anticipation category records a much higher difference in the range values for the predator (0-30) and victim conversations (0-18). Since one of the main objectives of the sexual offenders is to gain the confidence of the victim, predators highly make use of words like "*believe*", "*trust*", "*faith*", and "*hope*" while chatting.

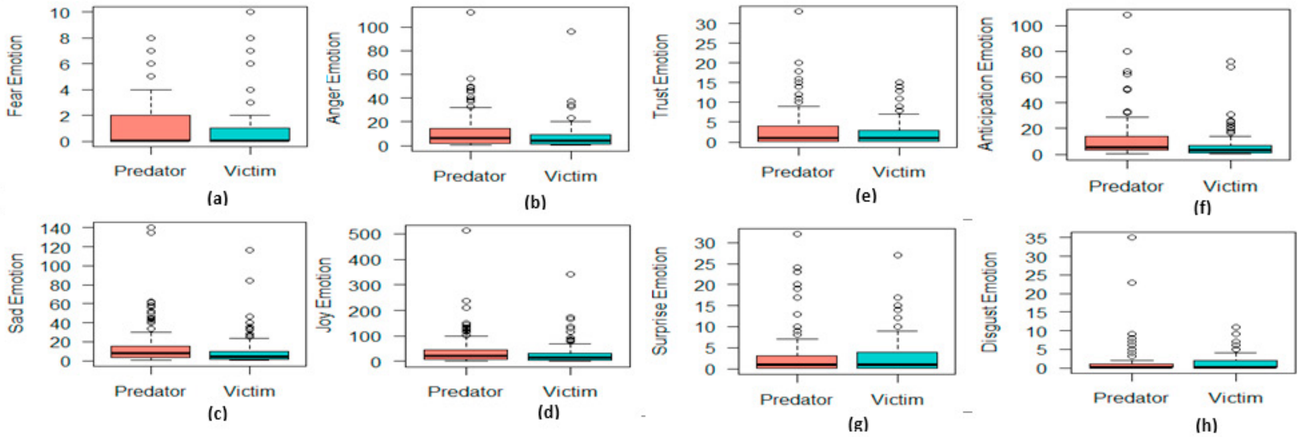


Fig. 3. Boxplots of Emotion features used in the study.

It can be observed that all the six emotion categories discussed above obtain higher values for the predator classes. The other two emotion classes, disgust and surprise, however, notice different patterns as victims receive higher values in these classes. The victim users score (0-10) and (0-4) for surprise and disgust respectively, whereas predators score (0-7) and (0-2) for the two categories, respectively.

Figure 4 shows the scatter plots of some of the topmost words from predator and victim class. The words "*understand*", "*beautiful*" and "*horny*" belong to $Set(P - V)$, whereas, the words "*idk*", "*swear*" and "*aww*" belong to $Set(V - P)$. The scatter plots of these words clearly explain the difference in the frequency of the usage of these terms in the content of two user groups.

While analyzing the vocabulary features, one more interesting behavior has been observed. The word "*chatting*" appears more in predator content, whereas "*chattin*" (incorrect spelling of chatting) appears in the victim content. Similar examples include the word pairs "*hanging*"-"*hangin*", "*making*"-"*makin*", and "*shopping*"-"*shoppin*", where the former is always preferred by predators and the latter is preferred by victims. It indicates that victims who are

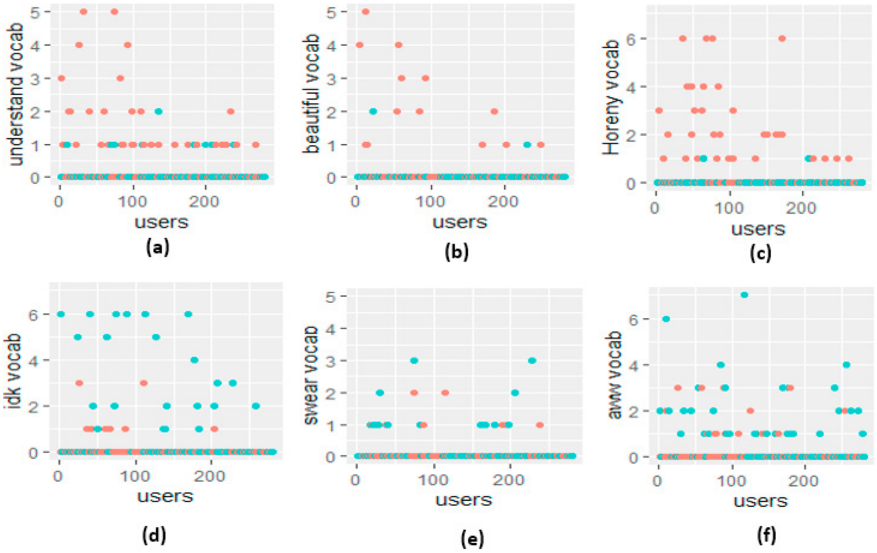


Fig. 4. Scatterplots of some Vocabulary features used in the study.

mainly youngsters or children tend to use informal vocabulary in their conversation. On the contrary, since predators are mature adults, they are not seen following such behavior. The scatter plots of few such words are shown in Figure 5.

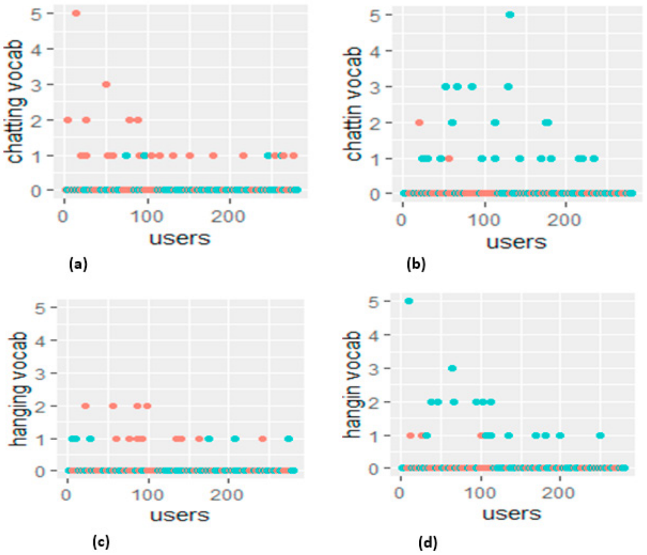


Fig. 5. Scatterplots of few unique vocabulary terms used in the sexual predators and victims in chatting.

6. Experiments and Results

The prime objective of this paper is to detect a sexual predator based on his/her behavior while chatting with an unknown person. For our experiments, we used the PAN 2012 dataset, which contains an XML file containing different conversations, including predatory conversations. Out of the given conversations, we designed two datasets to analyze the distinguishing power of different features. A brief description of these datasets is as under.

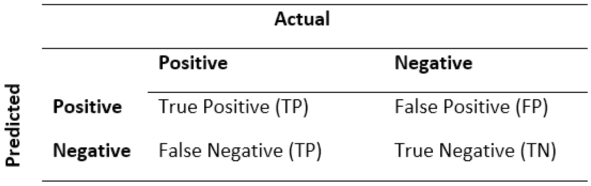
- Dataset 1 (D1): Emotion-based features (calculated using MoodBook)
- Dataset 2 (D2): Vocabulary- based features (Calculated using CBOW (Count of Bag Of Words) + Emotion features (D1))

In the two subsections below we will first describe how training and testing is done in Section 6.1, while in Section 6.2 we will compare our results with other existing approaches to the same problem.

6.1. Training and Testing

To observe the potential of our selected features in designing the predator detection model, we performed experiments using three widely used classification techniques, i.e. Decision Tree (DT), Support Vector Machine (SVM) and, Random Forest (RF). The training data has been divided into 80% for training and 20% for validating the model. Final testing has been done separately on the testing set. We trained and validated every algorithm using the two designed datasets D1 and D2, thus conducted 6 experiments in total. The commonly used performance evaluation measures for classification models were employed, such as accuracy, precision, recall and F-measure

The confusion matrix for this setting is given in Figure 6. Precision (P) calculates ratio of correct decision for the case that the classifier returns a positive verdict, whereas, recall (R) computes the ratio of correct predictions for the positive class. Maximizing precision can be achieved by minimizing the false-positive errors, whereas maximizing the recall means minimizing the false-negative errors. The F_1-score is harmonic mean of precision and recall and gives both the same weight. In the F$_\beta$-score, precision and recall are weighted differently. The formulas for calculating each of the metrics are given in Equations 4, 5, and 6.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (TP)	True Negative (TN)

Fig. 6. Confusion Matrix

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F_\beta - score = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \tag{6}$$

The F$_\beta$-score is a generalization of the F-score that is used to put more weight on either precision or recall. Here we will use, besides the F_1-score also the $F_{0.5}$-score (to emphasis the importance of fewer false positives) and the F_2-score (to emphasis the importance of fewer false negatives).

Table 6 shows the results obtained after applying the selected machine learning techniques to the two datasets. The results clearly show that all three classification techniques achieved good results on both the datasets. However, models trained on the dataset D2, that contains both emotion and vocabulary-based features, performed better than the models built using the dataset D1, that comprises only emotion features. In both the datasets, the SVM algorithm shows the weakest performance with F_1-scores of 57% and 86% on D1 and D2 respectively. In the case of dataset D1, DT yields the highest values for accuracy (99.76%) and F_1-score (89%), while for dataset D2, RF just outperforms RF with 99.88% and 95% for accuracy and F_1-measure respectively.

Table 6. Performance of Several Supervised Learning Techniques using Vocabulary- and Emotion-based Features.

Dataset	**Decision Tree (DT)**		**Support Vector Machne (SVM)**		**Random Forest (RF)**	
	Accuracy	F_1-Score	Accuracy	F_1-Score	Accuracy	F_1-Score
D1	99.76	0.89	95.34	0.57	99.35	0.81
D2	99.86	0.94	99.69	0.86	99.88	0.95

It is to be noticed that the difference between the best scores of accuracy for D1 and D2 is only 0.12% (99.76% compared to 99.88%), whereas, in terms of F_1-measure, the difference between the highest scores is much higher, i.e., 6% (89% compared to 95%). This is due to the fact that the two datasets are highly imbalanced with approximately 1% predators in the testing sets. Since accuracy is insensitive to the distribution of the proportion of two classes, victims and predators, the potential of the model for classifying predator as predator or misclassifying predator as victim becomes invisible in the evaluation metric. In contrast, the F_1-score being the harmonic of precision and recall is capable of evaluating the strengths and weakness of the models even for imbalanced datasets, and therefore, 6% gain in the performance can be seen by utilizing emotion as well as vocabulary-based features. Please note that here we did not report the results of two feature sets namely emotion and vocabulary separately as we observed almost same figures (for accuracy and F_1-score) for both the feature sets. Also, the aim is to present the potential of emotion features in detecting the sexual predators.

6.2. Comparison with Existing Approaches

Finally, in this section, we compare our work with the previous approaches based on the evaluation metrics precision, recall, and different variants of the F-score ($F_{0.5}$, F_1, and F_2). For the predator detection problem, most of the authors focus on precision more than recall to avoid falsely accusing a person for being a sexual predator. We do believe that it is also important to detect as many sexual predators as possible, hence focusing on a higher recall to avoid that some sexual predators go undetected. In Table 7, we aimed to compare our results with the previous works based on all the above metrics.

Table 7. An example of a table.

Approach	**Prec.**	**Rec.**	$\boldsymbol{F_1}$	$\boldsymbol{F_{0.5}}$	$\boldsymbol{F_2}$
Bours & Kulsrud (2019) [4]	0.89	0.87	0.88	0.89	0.87
Ebrahimi et al. (2016) [8]	0.92	0.72	0.81	0.87	0.75
Ebrahimi et al. (2016) [9]	0.78	0.50	0.61	0.70	0.54
Eriksson & Karlgren (2012) [10]	0.86	0.89	0.87	0.86	0.88
Fauzi & Bours (2020) [11]	0.96	0.86	0.90	0.93	0.88
Morris (2013) [14]	0.74	0.95	0.83	0.77	0.90
Parapar et al. (2012) [15]	0.94	0.67	0.78	0.87	0.71
Peersman et al. (2012) [16]	0.89	0.60	0.71	0.81	0.64
Villatoro-Tello et al. (2012) [19]	0.98	0.77	0.87	0.93	0.80
Proposed approach (RF-based)	***0.97***	***0.94***	***0.95***	***0.96***	***0.95***

It can be clearly observed in Table 7 that the proposed approach has obtained the highest values for F_1-score (0.95), $F_{0.5}$-score(0.96), and F_2-score (0.95) in comparison to the previous approaches. The ranking metric for the PAN 2012 competition was the $F_{0.5}$-score [13], where the argument was to not falsely accuse non-sexual predators, and thereby, not overloading law enforcement officers with a task of investigating many false cases. In case of the $F_{0.5}$-score, our approach again is ranked highest, even higher than the solution in [19] which was the best submission during the PAN 2012 competition. When considering the F_2-score (related to detection of as many sexual predators as possible), our approach gave also the best performance, even better than the top performer [14] in the given list. Furthermore, based on the values of precision (0.97) and recall (0.94), the proposed model performs better than most of the given approaches.In contrast to the existing studies the proposed approach made use of emotion-based features for distinguishing sexual-predators and normal user in chat conversations.

7. Conclusions and Future Work

This paper primarily focuses on extracting the communicating and emotional patterns of users as social biometric traits to identify sexual predators in online chatting systems. A vocabulary has been designed (PreVicVocab) that stores the vocabulary-terms preferred by victims and predators differently. It is observed that approaching words such as *call* and sexual words like *horny* appear more often in the sexual offense content. The CBoW approach has been used to represent the vocabulary-based features where we determine the frequency of the presence of the respective word in the user-content. *MoodBook* assists in extracting the 11 emotion-based features of the users. The box plots of the emotions clearly revealed the difference in the emotion behavior of sexual predators and victims.

The potential of the proposed approach is tested on the test set of PAN 2012 corpus, using machine learning techniques, including SVM, DT, and RF. The experiments are conducted on two datasets comprising of emotion, and a vocabulary (+ emotion) feature-set, respectively. The training dataset has been divided into two sets in the ratio of 80:20 for training and validation, respectively. Finally, trained models have been tested separately on the test set provided by PAN 2012. On the test set, all three algorithms also achieve the best results on the combined dataset. Here, we recorded accuracy of 99.86%, 99.69%, and 99.88% by DT, SVM, and RF, respectively which outperform the existing approaches. Furthermore our model obtained values of 0.95, 0.96, and 0.95 for F_1, $F_{0.5}$, and F_2-score, respectively. While comparing our results with existing studies, the proposed approach has obtained the highest values for F_1, $F_{0.5}$, and F_2-metrics and nearly the highest values for Precision and Recall. This study demonstrated that emotional attributes together with vocabulary features significantly aid in the sexual predator identification task.

Since sexual predator detection in online chats is more like a classification task and *Deep Learning* has been seen best suited and promising in this domain, therefore, in the future, we will investigate the application of deep learning algorithms for solving this task. Furthermore, instead of considering the full-length conversation for experiments, we will focus on recognizing a predator at his/her early stage in order to protect a user from being victimized.

Acknowledgements

This work was carried out during the tenure of an European Research Consortium for Informatics and Mathematics (ERCIM) Alain Bensoussan Fellowship Program.

References

[1] Ben-Zeev, A., 1987. The nature of emotions. Philosophical Studies 52, 393–409.

[2] Bentley, H., Burrows, A., Clarke, L., Gilligan, A., Glen, J., Hafizi, M., Kumari, P., Mussen, N., O'Hagan, O., Peppiate, J., 2019. How safe are our children? 2019. an overview of data on child abuse online.

[3] Bogdanova, D., Rosso, P., Solorio, T., 2012. On the impact of sentiment and emotion based features in detecting online sexual predators, in: CLEF (Online Working Notes/Labs/Workshop), pp. 110–118.

[4] Bours, P., Kulsrud, H., 2019. Detection of cyber grooming in online conversation, in: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE. pp. 1–6.

[5] Cardei, C., Rebedea, T., 2017. Detecting sexual predators in chats using behavioral features and imbalanced learning. Natural Language Engineering 23, 589–616.

[6] Chang, F.C., Chiu, C.H., Miao, N.F., Chen, P.H., Lee, C.M., Chiang, J.T., 2016. Predictors of unwanted exposure to online pornography and online sexual solicitation of youth. Journal of health psychology 21, 1107–1118.

[7] Chung, C., Pennebaker, J., 2012. Linguistic inquiry and word count (liwc): pronounced "luke,"... and other useful facts, in: Applied natural language processing: Identification, investigation and resolution. IGI Global, pp. 206–229.

[8] Ebrahimi, M., Suen, C.Y., Ormandjieva, O., 2016a. Detecting predatory conversations in social media by deep convolutional neural networks. Digital Investigation 18, 33–49.

[9] Ebrahimi, M., Suen, C.Y., Ormandjieva, O., Krzyzak, A., 2016b. Recognizing predatory chat documents using semi-supervised anomaly detection. Electronic Imaging 2016, 1–9.

[10] Eriksson, G., Karlgren, J., 2012. Features for modelling characteristics of conversations, in: CLEF (Online Working Notes/Labs/Workshop), pp. 1–8.

[11] Fauzi, M.A., Bours, P., 2020. Ensemble method for sexual predators identification in online chats, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), IEEE. pp. 1–6.

[12] Gottfried, E.D., Shier, E.K., Mulay, A.L., 2020. Child pornography and online sexual solicitation. Current psychiatry reports 22, 1–8.

[13] Inches, G., Crestani, F., 2012. Overview of the international sexual predator identification competition at pan-2012., in: CLEF (Online working notes/labs/workshop), pp. 1–12.

[14] Morris, C., 2013. Identifying online sexual predators by svm classification with lexical and behavioral features. Master's thesis. University Of Toronto, Canada.

[15] Parapar, J., Losada, D., Barreiro, A., 2012. A learning-based approach for the identification of sexual predators in chat log, in: CLEF (Online working notes/labs/workshop), pp. 1–12.

[16] Peersman, C., Vaassen, F., Van Asch, V., Daelemans, W., 2012. Conversation level constraints on pedophile detection in chat rooms, in: CLEF (Online working notes/labs/workshop), pp. 1–13.

[17] Shelton, J., Eakin, J., Hoffer, T., Muirhead, Y., Owens, J., 2016. Online child sexual exploitation: An investigative analysis of offender characteristics and offending behavior. Aggression and violent behavior 30, 15–23.

[18] Sultana, M., Paul, P., Gavrilova, M., 2015. Social behavioral biometrics: An emerging trend. International Journal of Pattern Recognition and Artificial Intelligence 29, 1–20.

[19] Villatoro-Tello, E., Juárez-González, A., Escalante, H.J., Montes-y Gómez, M., Pineda, L.V., 2012. A two-step approach for effective detection of misbehaving users in chats., in: CLEF (Online Working Notes/Labs/Workshop), pp. 1–12.

[20] Wani, M.A., Agarwal, N., Jabin, S., Hussain, S.Z., 2018. User emotion analysis in conflicting versus non-conflicting regions using online social networks. Telematics and Informatics 35, 2326–2336.

[21] Wani, M.A., Agarwal, N., Jabin, S., Hussain, S.Z., 2019. Analyzing real and fake users in facebook network based on emotions, in: 11th International Conference on Communication Systems & Networks (COMSNETS), IEEE. pp. 110–117.

[22] Wani, M.A., Bours, P., Agarwal, N., Jabin, S., 2020. Emotion-based mining for gender prediction in online social networks, in: International Conference on Machine Learning and Data Science (ICMLDS), ACM. pp. 100–109.

[23] Zhang, Y., Jin, R., Zhou, Z.H., 2010. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1, 43–52.