Information Retrieval and Classification based Approaches for the Sexual Predator Identification

Darnes Vilariño, Esteban Castillo, David Pinto, Iván Olmos, and Saul León

Benemérita Universidad Autónoma de Puebla
Faculty of Computer Science, Mexico

{darnes, dpinto, iolmos}@cs.buap.mx
ecjbuap@gmail.com, saul.ls@live.com

Abstract In this paper we present the evaluation of two different approaches with the aim of tackling the task of Sexual Predator Identification of PAN 2012. The first approach uses a dictionary of sexual terms in order to identify those documents associated in some manner with a sexual predator behavior. In order to do so, we use the sexual terms of the dictionary as a query in an information retrieval system, thus, retrieving the documents that best match with the query introduced. The second approach uses the multinomial Naïve Bayes classifier in order to detect sexual predators. The first approach performed better than the second one with low percentages of precision and high values of recall.

Keywords: Sexual predator, Chat messages, Sexual terms, Information retrieval, Supervised classifiers

1 Introduction

Nowadays there has been a growing number on the use of messaging systems such as chats and instant messaging which provide sexual predators a good platform for sexual purposes. Thus, it becomes very important to tackle the problem of sexual predator identification in order to ameliorate cases of sexual harassment. In this context, the 6th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN'12) has proposed a task named Sexual Predator Identification with aims to generate a framework in which different teams around the world may compare their approaches solving this particular problem. The goal is to provide an automatic method that permits to detect chat conversations in which one person attempts any erotic or suggestive remarks. The proposed task is subdivided into two sub-tasks:

- Identification of sexual predators. To detect the users considered to be sexual predators by classifying those conversations containing sexual predator behavior.
- Identification of sexual predator chat lines. To detect the specific lines in which the sexual remark is done.

The above mentioned problem has been tackled in this research work by means of two different approaches. The first one uses techniques of information retrieval, whereas the second uses classical methods of supervised machine learning. The aim of this work is to determine which one obtains the best performance.

The remaining of this paper is structured as follows. In Section 2 and 3, the two different approaches are explained. Section 4 shows the results obtained for each approach. Finally in Section 5 the conclusions of this work are given.

2 Information Retrieval Based Approach

As mentioned before, this approach considers a number of sexual terms as query for an information retrieval system. Each query is constructed with one original sexual term with its corresponding synonyms. For the indexing process, the chat messages that belong to the same conversation are considered to be a document d. Thus, we constructed a posting list by using as document the target conversations.

With the purpose of detecting those conversations using terms associated with a sexual orientation, we use the cosine similarity metric in order to determine the matching degree between each conversation and each query. The matching procedures is shown in the Algorithm 1, with $tf_{t,d}$ equal to the term frequency of t in document d. For the implementation, we considered the normalized version of tf as $f_{i,j} = \frac{tf_{i,j}}{max(f_j)}$, where $max(f_j)$ is the maximum term frequency in the document. The number of queries was equal to 919 which matches the original sexual terms considered, and the weight of each query term is calculated as $w_{t,q} = (0.5 + (0.5 \times f_{i,j})) \times log_{10} \frac{N}{df_i}$, with N equal to the number of conversations.

```
Algorithm 1: CosineScore(q)
  Input: Posting List
  Input: $K$ : number of documents to return
  Input: length[$N$]: length of each document of the collection
  float $Scores[N] = 0$;
  Initialize $length[N]$;
  foreach term $t$ in $q$ do
      calculate $w_{t,q}$ and fetch postings list for $t$;
      foreach pair(d,$tf_{t,d}$) in postings list do
          $Scores[d] + = wf_{t,d} \times w_{t,q}$;
  Read the array $Length[d]$;
  foreach $d$ do
      $Scores[d] = Scores[d]/Length[d]$;
  return Top $K$ components of $Scores[]$
```

The top 10 documents for each query are returned as conversations associated with a sexual predator. Since, 10 documents at most are obtained for each entry, we should be returning 9190 conversations in total, but in our case we returned 9071 documents.

3 Multinomial Naïve Bayes Approach

We have used a probabilistic supervised learning method named multinomial Naïve Bayes in order to determine sexual predators (as described in [1]). The probability of a document (message) d being written by sexual predator a is computed as shown in Eq.(1).

$$P(a|d) \approx P(a) \prod_{1 \leq k \leq n_d} P(t_k|a) \tag{1}$$

where $P(t_k|a)$ is the conditional probability of the k-th term (t_k) occurring in a message written by sexual predator a. Actually, $P(t_k|a)$ measures the contribution of term t_k so that the message d belongs to class a. n_d is the number of terms in message d. $P(a)$ is the prior probability of a message written by sexual predator a. Since we are really interested in finding the best class (sexual predator) for the document, we may calculate the maximum a posteriori (MAP) as shown in Eq.(2).

$$a_{map} = \arg\max_{a \in A} P^*(a|d) = \arg\max_{a \in A} P^*(a) \prod_{1 \leq k \leq n_d} P^*(t_k|a) \tag{2}$$

$P^*(t_k|a)$ is estimated by using Laplace smoothing, which simply adds one to each count (See Eq. (3)).

$$P^*(t_k|a) = \frac{T_{at_k} + 1}{\sum_{t' \in V}(T_{at'} + 1)} \tag{3}$$

where T_{at_k} is the number of occurrences of t_k in training documents from class a, including multiple occurrences of a term in a document and V is the corpus vocabulary.

4 Obtained results

In Table 1, the obtained results for the two approaches are shown. As can be seen, the number of conversations retrieved produced a result with high recall and low precision.

Table 1. Sexual Predator Identification (Goal: Identify the predators)

Task	**Retrieved**	**Relevant**	**precision**	**recall**	**F1**	**F($\beta = 0.5$)**
dictionary of terms	9071	232	0.0256	0.9280	0.0498	0.0378
multinomial Naïve Bayes	5225	97	0.0186	0.3880	0.0354	0.0272

In Table 2 the number of correct lines obtained are evaluated. Since the organizers of the competition evaluated this part manually, we do not have results for the information retrieval based approach. The number of lines retrieved is so high in comparison with those lines that really show a sexual predator behavior.

Table 2. Sexual Predator Identification (Goal: Identify predators line)

Task	Retrieved	Relevant	precision	recall	F($\beta = 1$)	F($\beta = 3$)
multinomial Naïve Bayes	6787	47	0.0069	0.0073	0.0071	0.0072

5 Conclusions

We have attempted two basic approaches for the sexual predator identification task. One approach based on information retrieval techniques, and the second one that uses a supervised classifier based on Naïve Bayes. The first approach performed better than the second one with low percentages of precision and high values of recall.

References

1. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)