Sexual predator identification

Luka Križan, Marko Lukić, Josip Užarević

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{luka.krizan, marko.lukic, josip.uzarevic}@fer.hr

Abstract

This paper describes the system for sexual predator identification in internet chat developed as a project for Text Analysis and Retrieval course at Faculty of Electrical Engineering and Computing. Tasks for the project were to identify sexual predators within all users and to identify lines of conversations which are the most distinctive of the predator bad behavior. We developed a simple system using machine learning with TF-IDF weighted vector space model and compared our results with results from PAN 2012 competition.

1. Introduction

According to Wikipedia, a sexual predator is a person seen as obtaining or trying to obtain sexual contact with another person in a metaphorically "predatory" or abusive manner. During past few decades, Internet has become the most dominant social media giving the opportunity to users from all around the world to communicate, e.g. through instant messaging services, social networks, forums and blogs. Such services give users option to hide their personal information, which while giving opportunity for users to establish new connections, can also represent a threat from sexual predators, especially to children and underage users. According to NBC news program *Dateline*, USA law enforcement officials in 2006 estimated that around 50000 sexual predators are online at any given moment. Although the origins of that figure have been questioned by many sources, the figure still shows that threat from sexual predators to underage users in online activities is present and real.

In 2012, *PAN*[1] ran a competition in sexual predator identification with a task to identify sexual predators from given chat logs which involved two or more people. The task was divided into two parts:

- identifying the predators within all the users
- identifying the lines of the predator conversations which are the most distinctive of the predator bad behavior

Training corpus for the competition included 66928 conversations between 97690 different users from which 148 were tagged as sexual predators. For the second part of the task, no direct training data was given.

2. Related work

As this task was the part of PAN 2012 competition, we focused on participants' approaches (Inches and Crestani, 2012).

2.1. Identifying predators

Collection given as a training set was intentionally unbalanced (having less than 1% of positive class examples), so most of the participants used two stage classifier to filter out negative examples, which got highest score in competition (Villatoro-Tello, et. al). In first stage, classifier distinguished between conversations involving a predator and conversations without a predator. Second stage was used for predator-vs-victim classification. Other approach in filtering out negative examples was to remove conversations that included only one participant, those that had less than six messages per user, etc. Most of the participants used unigram or bigram model with TF-IDF weighting to extract features. In general, features have been used without stemming and stopword removal to preserve author's style, including misspelling and grammatical errors. Some of the participants included "behavioural" features, e.g. the number of times a user starts a dialogue, the number of questions asked, LIWC features, etc. Linguistic Inquiry and Word Count (LIWC) is a text analysis software that computes the degree in which people use different categories of words. This can be used to determine individual's psychological aspects given in natural language of their messages. For classification, participants used various methods, focused mostly on machine learning algorithms. The most used method was support vector machine (SVM).

2.2. Identifying predators' lines

The simplest solutions used for this problem was to return as relevant all the conversations lines of all identified predators from the first problem. The most used method was filtering of all predator conversations through a dictionary of "perverted" terms or with a particular score (e.g. TF-IDF weighting).

3. Proposed method

To implement a system for sexual predator identification, we used simpler approach than most of the participants and included some of the methods mentioned above, but with a focus on only lexical features of messages sent between users. System was implemented in Python using text analysis tools from NLTK [2] module and machine learning tools from scikit-learn [3] module.

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

[2] http://www.nltk.org/

[3] http://scikit-learn.org/

Table 1: Number of users removed from training data related to the value of threshold (minimal number of messages sent)

Threshold	Users removed	Positive users removed
1	20804	1
2	57962	4
3	72720	4
4	75644	4
5	76578	4
6	77236	4
7	77956	4
10	81013	5

3.1. Identifying predators

3.1.1. Preprocessing

After taking into account the nature of chat messages, we decided only to perform tokenization as a part of preprocessing. Chat messages are usually poorly structured with many grammatical errors and misspellings. They also include emoticons, internet slang and intentional deviations from rules of grammar. Performing stopword removal, stemming or lemmatization in this case wouldn't have significant effect in classification and it would greatly increase computation time. Also, intentionally misspelled words and emoticons can contain significant information for context of messages which would get lost after performing additional preprocessing. In the end, we used lemmatization, but it had almost none effect in reducing total number of words appearing in messages.

Due to lack of efficient preprocessing, total number of words appearing in messages was too large to efficiently do any kind of further computation on input data. To reduce the total number of words, we decided to remove words which in total occurred less than 5 times. After word filtering, total number of words was reduced to approximately half of its original value.

3.1.2. User filtering

As was mentioned in introduction, training data included conversations between 97690 users, from which only 148 were tagged as sexual predators. To reduce the number of users tagged as negative, we decided to put a threshold - minimal number of messages sent per user. If number of messages sent by particular user was less or equal than value of threshold, he/she was removed from training data. Number of users (tagged both positive or negative) removed from data related to the value of threshold is shown in Table 1. After analysis, we decided to set value of threshold to 5 because no significant improvement was made with higher threshold value. Total number of users was reduced almost to a quarter of its original value, while only 4 users tagged as positive were removed.

Also, to reduce time spent in computing user representation, we removed one user who was tagged as negative, because his TF-IDF computation time was longer than computation time of all other users combined.

3.1.3. User representation

For user representation, we used a very simple approach. Every chat user is represented with a document - a collection of all messages that he/she sent in any conversation that he/she participated. Every document was transformed into a unigram representation, weighted with a standard TF-IDF weighting scheme. As a result, every user was simply represented as a floating point vector. Additionally, we also included a non-text feature - average time of the day when user was chatting. We divided time of the day into three categories:

- morning - between 7:00 and 14:59
- afternoon - between 15:00 and 22:59
- night - between 23:00 and 6:59

Because of an equal distance between these categories, we added them to feature vectors using one-hot encoding. In the end, total representation of a user was a vector with a size of 59378.

The simplicity of this approach also has its negative aspects. It is not possible to extract most of "behavioural" features that could also be important for classification, e.g number of conversations started by user, number of users that were included in conversations with this particular user, frequency of turn-taking, etc. However, we expect that this representations still contains enough amount of information to identify bad behaviour distinctive for sexual predators.

3.1.4. Training

Due to highly unbalanced dataset, straight-forward approach in classifier training (performing training with exact data which was left after filtering) would result in trivial classification in which would all users be classified as negative. To tackle that problem, we considered two approaches - oversampling the positive class and undersampling the negative, majority class. As we already encountered problems with too high computational costs, we decided for undersampling the negative class.

In the first step of training, we split training collection into training and test parts (with 70-30 ratio, while keeping the number of positive class examples balanced). In the second step, we performed undersampling of the training part, using different amount of negative class examples ($\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\} * n$, where n is a number of negative class examples in training part). After undersampling, we performed 5-fold cross-validation on resized training part, which was used to find classifier model and its hyperparameters that maximize F0.5 score (measure which was used in PAN 2012 for evaluation of predator identification). For experiments, we used SVM (with linear and RBF kernel), logistic regression and k-nearest neighbours classifier. Classifier with best score in cross-validation was tested on test part and also evaluated with F0.5 measure. To find optimal amount of undersampling, whole cross-validation process was then performed again, with lower rate of undersampling. Due to non-deterministic way in which negative users were chosen for undersampling, we repeated each test several times and took the best result.

Table 2: Results for sexual predator identification with L2-regularized logistic regression on test collection

Precision	Recall	F1 score	F0.5 score
0.9175	0.7008	0.7946	0.8641

Logistic regression and SVM with linear kernel showed best results, while KNN classifier showed very poor results. The highest score on test part scored L2-regularized logistic regression, while using only a quarter of negative class examples from training part. Also, SVM with linear kernel while using the same number of negative class examples scored almost the same score as logistic regression. Due to very high computation costs of training and classification with SVM, we decided to focus only on logistic regression. Logistic regression classifier was then retrained only on data used in cross-validation, to keep the same negative class examples from undersampling which showed best results. The downside was that the classifier was trained using only about 70% of positive class examples from the training collection.

3.2. Identifying predators' lines

To identify predators' lines, we considered two approaches. First approach was to return all messages from all users who were classified as predators in the previous step. Even though it is an oversimplified approach, participants who used it in PAN 2012 competition received good results. For second approach, we decided to modify previous approach. Instead of returning every message sent by users classified as predators, we filtered messages by classifying them individually using the classifier from the previous step. Our assumption was that the second approach would outperform the first approach because it should have a better precision score while keeping a similar recall score.

4. Results

The test collection contained conversations between 218702 users, from which 254 were marked as sexual predators. Both training and test collection were equally unbalanced, containing around 0.1% of positive class examples. For the second part of the problem, 6478 messages from the test collection were tagged as distinctive of the predator bad behaviour. F0.5 score was used for evaluating the system in predator identification and F3 for identifying predators' lines.

4.1. Identifying predators

The results for predator identification are shown in Table 2. Classifier had a high score in precision and a lower score in recall, but it was trained to maximize F0.5 score (which emphasizes precision) so total result was not unexpected. F0.5 score of 0.8641 would place our system at 5th place (out of 16 participants) in predator identification in PAN 2012 competition, with less than 0.01 lower score than systems that were ranked as 3rd and 4th.

Table 3: Results for identifying predators' lines problem for both approaches used (returning all messages sent by users marked as predators in first step and filtering messages by their individual classification using the classifier from first step)

Method	Precision	Recall	F3 score
Returning all messages	0.0967	0.8902	0.4889
Message filtering	0.0721	0.0012	0.0013

4.2. Identifying predators' lines

Our results for the second part of the problem are shown in Table 3. Results were completely opposite from our initial expectation, since second approach, which used message filtering, scored a really low score, both in precision and recall. First approach had a low score in precision because it returned all messages sent by predators, but it showed good result in total because F3 score was used (which emphasizes recall). Our system, while using very simple approach of returning all messages sent by users classified as positive in previous step, outperformed all systems in identifying predators' lines and with F3 score of 0.4889 it would be ranked as 1st in PAN 2012 competition.

5. Conclusion

Our system, even though it was much simpler than most systems used by participants in PAN 2012 competition, ended with good results in both parts of the sexual predator identification task. As always, there is a lot room for improvement and with inclusion of more behavioural features in user representation, system would probably end up with better results in predator identification (and consequentially with even better result in identifying predators' lines).

Taking into account that this was only a course project and that it was our first project including text analysis, we are satisfied with results, especially with the fact that our system would be ranked as 1st in identifying predators' lines.

References

Giacomo Inches and Fabio Crestani, Overview of the International Sexual Predator Identification Competition at PAN-2012

Javier Parpar, David E. Losada and Alvaro Barriero, A learning-based approach for the identification of sexual predators in chat logs - notebook for PAN at CLEF 2012

Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda, A Two-step Approach for Effective Detection of Misbehaving Users in Chats - notebook for PAN at CLEF 2012