
UNIVERSIDADE DA CORUÑA


Javier Parapar, David E. Losada and Álvaro Barreiro

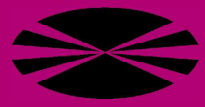
A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs

Rome, September 20th 2012

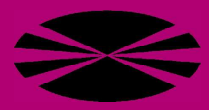
PAN: Uncovering Plagiarism, Authorship and Social Software Misuse

CLEF 2012

Outline

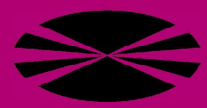


Our task: Sexual Predator Identification



- Author Identification
  - Sexual Predator Identification
    1. **Identify the predators**
    2. *Identify the part (the lines) of the predator conversations which are the most distinctive of the predator bad behavior*
- We tested different Machine Learning strategies and thoroughly tuned the classifiers
- Main contribution is the proposal of an innovative set of features to drive the classification of chat participants
  - Standard term-based features (tf/idf).
  - Psycholinguistic features (deception language).
  - User-level features based on activity.
- We have been centred on sub-task 1

Representing the chat users for learning



- Great **challenges** open in representing subjects and temporally modelling the **predation process** (gaining access → deceptive trust development→grooming→ isolation→ and approach)
- We opted for approaching this year's task in **a simple way**.
- For every individual, we **concatenated** together **all the lines** written by him/her in any conversation in which he/she participated (**user's document)**.
- It is a recognizably **simplistic** strategy but we expect that it **still contains the basic clues to identify** predation.
- These document-based representations were used as an input to extract the **content-based features**
- In addition **we devised other non-text based features**.

Extracting the features I

- We studied different strategies to extract a feature-based representation for the chat participants:
  - *tf/idf* features. Baseline representation consisting of a standard uni-gram representation of the texts.
    - No stemming, terms with $df < 10$ removed, terms longer than 20 chars were also removed.
    - Each term was weighted with a standard tf/idf weighting scheme (Sparke-Jones Journal of Documentation 1972)

$$tf/idf_{t,d} = (1 + \log(tf_{t,d})) \times \log(\frac{N}{df_t}) \quad (1)$$

    - **2-grams and 3-grams also considered** and tested in combination with uni-grams (unigrams only, bigrams only, trigrams only, unigrams+bigrams, unigrams+trigrams, bigrams+trigrams, and all n-grams)

Extracting the features II

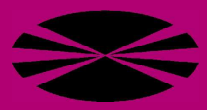
- *LWIC* (psycholinguistic) features.
  - Linguistic Inquiry and Word Count (LIWC) (Pennerbaker et al ARP 2003), which is a text analysis software program that calculates the degree to which people use **different categories of words**.
  - The **ways that individuals talk** and write provide **windows into their emotional and cognitive worlds** and can be used to analyze aspects such as **deception**, honesty, etc.
  - Overall there are **80 different LWIC dimensions** spanning from psychological constructs (e.g. affect, cognition) to personal concerns (e.g. work, home, leisure)

Extracting the features III

- *chat-based* features. Finally, we defined **11 additional features** that capture some global aspects related to the **activity** of the individuals in the **chatrooms**.

Feature Name	Feature Description
avgLineLengthChars	Average size (in characters) of the user's message lines in the collection.
avgTimeOfDayOfMessages	Average time of day when every message line was sent by the user. Time of day is measured in minutes from/to midnight (the smallest amount applies).
noOfMessageLines	Number of message lines written by the user in the collection
noOfCharacters	Character count of all the message lines written by the user in the collection
noOfDifferentUsers-Approached	Number of different users approached by the user in the collection
percentOfConversations-Started	Percentage of the conversations started by the user in the collection
avgNoOfUsersInvolved-InParticipedConversations	Average number of users participating in the conversations with the user
percentOfCharacters-InConversations	Percentage of the characters written by the user (computed across the conversations in which he/she participates)
percentOfLines-InConversations	Percentage of lines written by the user (computed across the conversations in which he/she participates)
avgTimeBetween-MessageLines	Average time, in minutes, between two consecutive message lines of the user
avgConversation-TimeLength	Average conversation length, in minutes, for the user (computed across the conversations in which he/she participates)

Learning and training I



- We used **LibLinear**, a highly effective library for large-scale linear classification, (Fan et al. JMLR 2008) for learning the classifiers.
- We tested against the training collection all the classifiers supported.
- We **chose SVMs**, precisely, a **L2-regularized L2-loss SVM primal solver**, as our classifier for all our submitted runs.
- As we are dealing with a highly unbalanced two-class classification problem we have to avoid resulting in a trained trivial classifier which ignores the minority class

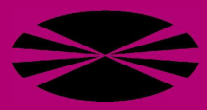
Learning and training II

- In order to do so we decided to **carefully adjust the misclassification costs** to penalize the error of classifying a positive example as negative (i.e. a sexual predator classified as a non-predator).
- Training with **4-fold cross-validation** and focused on optimizing F1 computed with respect to the positive class (being a predator):
- We **fine tuned** the weights that adjust the **relative cost of misclassifying positive examples** (w_1)
- Given **different combinations of the feature sets** described before, we did not apply any feature selection strategy.

Learning and training III

- Performance with the different types of **features in isolation** is **not very good** (best F1 0.56) but **combining** them we achieved up to **0.84** F1.
- **Best results** in training and test were obtained combining **unigram tf/idf features and chat-level features**.
- For the **sub-task 2** we did **not** have **training data** so we simply processed the lines of the supposed SP with a tf/idf classifier trained for the sub-task 1 and submitted the ten lines with higher SP estimation, our **expectations** for this task were **rather low**.

Conclusions and Future Work



- We believe that we have successfully shown that a **learning-based approach** is a **feasible** way to approach this problem.
- We have proposed **innovative sets of features** to drive the classification of chat participants
- Our experiments demonstrated that the set of **features** utilized and the r**elative weighting of the misclassification** costs in the SVMs are the two main factors that should be taken into account to optimize performance.
- We want to carefully **analyse** the relative **importance of the individual features** in each feature set.
- We want to move to **more evolved representations** of the subjects taking into account the **sequential process of predation**


UNIVERSIDADE DA CORUÑA


Javier Parapar, David E. Losada and Álvaro Barreiro

A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs

Rome, September 20th 2012

PAN: Uncovering Plagiarism, Authorship and Social Software Misuse

CLEF 2012