Conversation Level Constraints on Pedophile Detection in Chat Rooms

PAN 2012 — Sexual Predator Identification

Claudia Peersman, Frederik Vaassen, Vincent Van Asch and Walter Daelemans
Task 1: Sexual Predator Identification
Preprocessing of the Data

- Data: PAN 2012 competition training set
  - info on the conversation, user and post level
  - predator ID list
- Two splits: training and validation set
- No user was simultaneously present in training and validation

→ prevent overfitting of user-specific features
Experimental Setup

- Features: token unigrams
- LiBSVM
- Probability output
- Parameter optimization
- Experiments on 3 levels
- Data resampling
Level 1: the Post Classifier

• Resample the number of posts
  → Equal distribution of posts per class
• About 40,000 posts per class in training
• No resampling in the validation sets
Level 1: the Post Classifier (2)

- Only output on the post level
- Aggregate the post level predictions to the user level:
  - LiBSVM’s probability outputs
  - Predators = average of the 10 highest predator class probabilities $\geq 0.85$
# Results for the Predator Class

<table>
<thead>
<tr>
<th>Scores</th>
<th>Post Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.93</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
</tr>
<tr>
<td>F-score</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Level 2: the User Classifier

- Resampling on the user level
  → exclude users with no suspicious posts
- Filter: dictionary of grooming vocabulary
  → see Task 2
- Why?
  - reduce the amount of data
  - “hard” classification → higher precision?
## Update Results (1)

Data reduction: up to 48.4%

<table>
<thead>
<tr>
<th>Scores</th>
<th>Post Classifier</th>
<th>User Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.93</td>
<td>0.82</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
<td>0.88</td>
</tr>
<tr>
<td>F-score</td>
<td>0.52</td>
<td>0.84</td>
</tr>
</tbody>
</table>

→ Combine systems?
Combining the systems

- Weighted voting using LiBSVM’s probability outputs
- 70% of the weight on the high precision User Classifier
## Update Results (2)

<table>
<thead>
<tr>
<th>Scores</th>
<th>Post Classifier</th>
<th>User Classifier</th>
<th>Combined Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.93</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>F-score</td>
<td>0.52</td>
<td>0.84</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Level 3: Conversation Level Constraints

- Both users in a conversation labeled as predators
- Our approach:
  - go back to predator probability output
  - use the high precision user classifier
  - Predator probability $\geq 0.75$
System Overview

Post Classifier

User Classifier

Combined Prediction

Apply Conversation Level Constraints

Final Predator ID List
## Update Results (3)

<table>
<thead>
<tr>
<th>Scores</th>
<th>Post Classifier</th>
<th>User Classifier</th>
<th>Combined Results</th>
<th>Combined + Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.93</td>
<td>0.82</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
<td>0.88</td>
<td>0.84</td>
<td>0.94</td>
</tr>
<tr>
<td>F-score</td>
<td>0.52</td>
<td>0.84</td>
<td>0.84</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Results on the PAN 2012 Test Set

<table>
<thead>
<tr>
<th>Scores</th>
<th>Combined + Constraints</th>
<th>PAN Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.85</td>
<td>0.60</td>
</tr>
<tr>
<td>Precision</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>F-score ($\beta = 1$)</td>
<td>0.89</td>
<td>0.72</td>
</tr>
</tbody>
</table>

- Future research:
  - more splits
  - investigate ensembles
Task 2: Identifying Grooming Posts
Identifying Grooming Posts

- From the final predator ID list \(\rightarrow\) detect posts expressing typical grooming behavior
- No gold standard labels \(\rightarrow\) What is grooming?
- Predator conversations have predictable stages (e.g. Lanning, 2010; McGhee et al., 2011)
Identifying Grooming Posts (2)

- Dictionary containing references to 6 stages:
  - sexual topic
  - reframing
  - approach
  - data requests
  - isolation from adult supervision
  - age (difference)
Identifying Grooming Posts (3)

- Resources:
  - McGhee et al. (2011)
  - English Synonyms [http://www.synonym.net/](http://www.synonym.net/)
  - cf. user classifier filter
Results on the PAN 2012 Test Set

- Precision = 0.36
- Recall = 0.26
- F-score ($\beta = 1$) = 0.30
Discussion

• Use of $\beta$-factors to calculate the F-score:
  • Task 1: focus on precision ($\beta = 0.5$)
  • Task 2: focus on recall ($\beta = 3.0$)

• However, in practice:
  • find all predators (recall in Task 1)
  • find the most striking posts (precision in Task 2)
Questions?

Contact: claudia.peersman@ua.ac.be